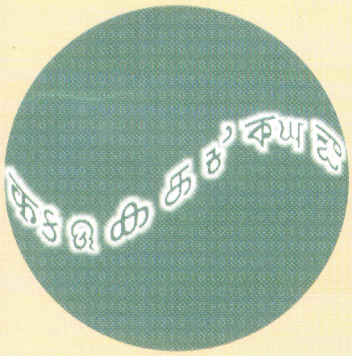


Multilingual Computing and Information Management in Networked Digital Environment

Third International CALIBER 2005

February 2-4, 2005

Jointly Organised by :
Cochin University of Science and Technology, Kochi &
INFLIBNET Centre, Ahmedabad

Proceedings
(Convention on Automation of Libraries in Education and Research Institutions)

Editors
T A V Murthy
S M Salgar
S K Sharma
K Prakash

INFLIBNET Centre
Ahmedabad



Dr. T. A. V. Murthy is the first LIS professional to become the Director of INFLIBNET with the status of a Central University Vice Chancellor and elected President of Society for Information Science. He is also the council member of IASLIC and Secretary of Ahmedabad Library Network. He holds B Sc, M L I Sc, M S L S (USA) and Ph.D from Karnatak University and carries with him a rich experience and expertise of having worked in managerial level at a number of libraries in many prestigious institutions in India including National Library, IGNCA, IARI, University of Hyderabad, ASC, CIEFL etc. apart from Catholic University and Casewestern Reserve University in United States. His highly noticeable contributions include KALANIDHI at IGNCA, Digital Laboratory at CIEFL, UGC Infonet E-Journal Consortium etc. He has been associated with number of universities in the country as visiting professor and has guided number of Ph.Ds and actively associated with the national and international professional associations, expert committees and has published good number of research papers and books. He visited several countries on professional activities and organized several national and international conferences and programmes. He is the recipient of SIS Fellowship, SATKAL Librarian Award and SALIS Harish Chandra Sushil Chandra Best Librarian Award.



Mr. S. M. Salgar is Scientist-G at INFLIBNET Centre, Ahmedabad, India. He is an Electronics & Communication Engineer, having done his M.E. (Hons.). He served for 20 years in Indian Space Research Organisation (ISRO) before joining INFLIBNET in 1991. He has published many scientific papers in national & international journals and successfully completed important projects of national importance. His specialization includes "Satellite Communication and Networking" and visited several countries on official assignments. He has been with INFLIBNET from its inception and has been actively involved and took leading role in all its developmental activities.



Mr. S K Sharma is Scientist-B, leader of Networking, Testing and Quality control group in Information and Library Network Centre, UGC-INFLIBNET Ahmedabad. He has a M.Sc. in Physics and Master of Computer Application (MCA). He has nearly 7 years rich experience in IT and published papers in national conference/journals. Currently he is doing research in Web Services Security. His major research interests are distributed computing, Network security and management.



Mr. K Prakash is working as Scientific/Technical Officer-I with INFLIBNET Centre since 1995. He holds BSc, MLIS from Karnatak University, Dharwad. He has qualified SLET. He has done specialization course in "IT Applications to LIS" from NCSI, IISc Bangalore. Pursuing research in Library Automation. Before joining to INFLIBNET, he has worked in academic and industrial libraries. Presently he is working in Serials Union Database and he is involved in Training and other Activities of the centre. He has contributed 20 papers in seminars, conferences and journals. He is a life member of ILA, IASLIC, SIS & KLA and he is managing digilib_India discussion forum also. His areas of interests are Library Automation, Database Management, Information Retrieval, Organisation of E-resources, Digital Libraries and Training.

Third International CALIBER-2005, Proceedings
ISBN : 81-902079-0-3

Multilingual Computing and Information Management in Networked Digital Environment

Convention on Automation of Libraries in Education and Research Institutions
[Third International CALIBER - 2005]
Kochi, INDIA, February 2-4, 2005

Jointly Organised by
Cochin University of Science and Technology, Kochi
Information and Library Network Centre, Ahmedabad

Proceedings

Editors

T A V Murthy
S M Salgar
S K Sharma
K Prakash

INFLIBNET **Centre**
Ahmedabad

EDITORS

T A V Murthy
INFLIBNET Centre, Ahmedabad, India
E-Mail: tav@inflibnet.ac.in

S M Salgar
INFLIBNET Centre, Ahmedabad, India
E-Mail: salgar@inflibnet.ac.in

S K Sharma
INFLIBNET Centre, Ahmedabad, India
E-Mail: sharma@inflibnet.ac.in

K Prakash
INFLIBNET Centre, Ahmedabad, India
E-Mail: prakash@inflibnet.ac.in

ISBN : 81-902079-0-3
© INFLIBNET Centre (Ahmedabad, INDIA) 2005

Published by

Information and Library Network Centre
An IUC of University Grants Commission
PB 4116, Navrangpura,
Ahmedabad – 380 009, India

Phone: 91-79-26304695 / 5971 / 8528
Fax: 91-79-26300990 / 7816
URL: http://www.inflibnet.ac.in
E-Mail: root@inflibnet.ac.in

Secretarial Assistance	:	Hiren Jani
Typeset & Design	:	Prakash Rathod and Maqsud Shaikh
Printed at	:	R . K. Print-art, Ahmedabad

No part of this publication can be reproduced in any form by any means without the prior written permission of the publisher. All data, views, opinions etc being published are the sole responsibility of the authors. Neither the publishers nor the editors in anyway are responsible for them.

COMMITTEES

Prof. Arun Nigavekar, Chairman,
University Grants Commission, New Delhi

PATRONS

Dr. S Ramani, Chairman,
INFLIBNET-GB, Ahmedabad

Prof. V N Rajshekharan Pillai, Vice-Chairman
University Grants Commission, New Delhi

Dr. P.K. Abdul Azis, Vice Chancellor,
Cochin University of Science & Technology, Kochi

Dr. T A V Murthy, Director,
INFLIBNET Centre / UGC, Ahmedabad

NATIONAL ADVISORY COMMITTEE

Prof. M.P. Satija, DLIS,
Guru Nanakdev University, Amritsar

Dr. R.D. Mehla, Librarian,
Kurukshetra University, Kurukshetra

Dr. S.D. Vyas, Librarian,
Banasthali Vidyapeeth, Banasthali

Prof. G. Devrajan, DLIS, Kerala University,
Thiruvananthapuram

Prof. V.G.Talwar, DLIS,
University of Mysore, Mysore

Dr. (Mrs.) R.S.R. Varalakshmi,DLIS,
Andhra University, Vishakhapattanam

Dr. Madhuri Devi,DLIS,
Manipur University, Imphal

Dr. Manibhai Prajapati, Librarian,
Hemchandracharya North Gujarat University, Patan

Prof.(Mrs.) A. Vaishnav, DLIS,
Babasaheb Ambedkar University, Aurangabad

INTERNATIONAL ADVISORY COMMITTEE

Mr. K.M. Abdul Awwal, Director, Library & Publication, UGC, Bangladesh

Mr. Krishna M Bhandary, Tribhuvan University, Nepal

Mr. Yuan Goujing, Associate Librarian, Shanghai Jiaotong University Library , China

Mrs. Habeeba Hussain Habeeb, National Library of Maldives, The Maldives

Dr. Muhammad Ramzan, Chief Librarian, Lahore University of Management Sciences, Pakistan

Prof. Dr. Kay Raseroka, President, IFLA

Mr. Ramchandran Rasu, Secretary General, IFLA
University of Colombo, Sri Lanka

Mr. Mynak R. Tulaku, National Library, Bhutan

Dr. N.U. Yapa, President, Srilanka Library Association, Srilanka

EDITORIAL COMMITTEE

Editor-in-Chief:

Dr. T A V Murthy, Director,
INFLIBNET Centre / UGC, Ahmedabad

Members

Mr. S.M. Salgar, INFLIBNET Centre, Ahmedabad

Mr. S.K. Sharma,INFLIBNET Centre, Ahmedabad

Mr. K. Prakash, INFLIBNET Centre, Ahmedabad

PROGRAMME ORGANISING COMMITTEE

Chairman :

Mr. S M Salgar, Scientist-G
INFLIBNET Centre / UGC, Ahmedabad

Organizing Secretary :

Dr. (Mrs.) M.D. Baby, Librarian, CUSAT, Kochi

Convener :

Mr. S.K. Sharma, Scientist-B, INFLIBNET Centre

Joint Convener :

Mr. Rajesh Chandrakar, Scientific & Technical Officer,
INFLIBNET Centre

Members :

Dr. R.Vengan, Librarian, University of Madras, Chennai.

Mr. P.K. Rajendrakurup, Dy. Librarian In-charge, M. G. University, Kottayam

Dr. N. Ushakumari, Librarian, Avinashilingam Institute for Home Science & Higher Education for Women, Tamil Nadu

Mr. V. C. Abdul Khader, Deputy Librarian, Incharge, Kerala University Kerala

Mr. P. Sanjeev, Kerala Agriculture University, Trichur

Technical Coordinators :

Mr. Umesh Gohel, STA, INFLIBNET Centre

PREFACE

Dr T A V Murthy
BSc, MLSc, MSLS (USA), Ph D, CBIS, CAA
Chairman, Editorial Committee for CALIBER 2005

Director
INFLIBNET **Centre**
An IUC of UGC
Gujarat University Campus
PB 4116, Navrangpura
Ahmedabad – 380 009 (Gujarat India)
E-mail: tav@inflibnet.ac.in
Telephone 91-79-26305702 [Dir]
26304695 26305971 26308528 [EPBAX]



President & Fellow, Society for Information Science
Secretary, Ahmedabad Library Network
Council Member, IASLIC

It is believed that all peoples would benefit from the use of computers and access to the Internet, but the computers, and now the internet, are dominated by English language and other major languages of the developed world. If computers and Internet are to be widely used, clearly it should take place using language of that society, just as all other activities do. In India with one billion people speaking many languages, localization has become an important requirement for the information technology industry and library and informational professionals. Local language computing in India started way back in 1980s, and the advent of Internet forced encoding standardization. Most of the existing software now is either based on Unicode or provide export and import facility. Due to the increased usage of computers and internet, localized operating systems are also under development.

The use of digital information in the modern world is increasing at a phenomenal rate. At the same time an increasing proportion of new information is only being conceived, produced, and distributed in electronic form and librarians and other information professionals are facing a new world of primary electronic objects, requiring professional management. Digital libraries have emerged as a crucial component of global information infrastructure, adopting the latest ICT to promote an organizational structure that encourages communication between scholars across nations, and helps disciplinary boundaries. The growth of Internet, the increased sophistication of web based tools, the intranet and campus networks within the organizations have changed the role of libraries. Huge volumes of data are available in this networked environment for academic sharing. The concern is how to integrate and handle this information and how quickly and seemingly access can be provided. Content Management increases document management efficiencies to capture, manage, store, preserve and deliver content.

The 3rd International CALIBER in the series, will be hosted by Cochin University of Science and Technology, Kochi during February 2-4, 2005. The main theme chosen for CALIBER – 2005 is **Multilingual Computing and Information Management in Networked Digital Environment**. The convention is intended to bring together researchers, information professionals, and developers working on multilingual computing, digital libraries, and related areas for in-depth analysis and discussion of new models, theories, frameworks, and solutions to interoperability in digital libraries. There were considerable response from research scholars, senior and eminent professionals in the

field who have contributed papers to this convention. There are around 119 papers received by us covering range of topics from natural language processing, metadata and content management and finally 86 papers are selected for full text publication and 25 papers are included in the 'Abstracts' section in this proceedings. On behalf of the Organising, Programme and Editorial Committees of CALIBER-2005, I thank all authors for their submissions and camera-ready copies of papers, and their cooperation at every stage. We are also delighted to see a sizable number of academics attending this programme from USA, Germany, Sri Lanka, Pakistan, Ghana, Bangladesh and other countries.

I sincerely acknowledge with thanks the help and encouragement of all those who were involved in bringing out this proceeding. My sincere appreciation goes to the entire team of INFLIBNET especially Shri D P Negi, Shri H G Hosamani, Shri Rajesh Chandrakar, Shri J K Vijayakumar, Shri Umesh Gohel and Shri S R Shah besides the members of editorial board, Shri S M Salgar, Shri S K Sharma and Shri K Prakash for their constant support throughout the entire process of this publication and for organizing this event. I am thankful to Shri K Haridas, Shri Hiren Jani, Shri Vinod Dantani and Shri Bakul Parmar of INFLIBNET for secretarial assistance and Shri Raghavendra Patil, Shri Maqsud Shaikh, Shri Prakash Rathod of INFLIBNET for Type Set and Design.

I am also thankful to the Bibliothek International, Germany for their collaborative support. The encouraging support from sponsors M/s Elsevier India (Science & Technology), Global Information System Pvt. Ltd. (GIST), Cambridge University Press and Informatics (India) Limited and the printer M/s R K Print-art, Ahmedabad for their timely help is greatly acknowledged.

At the end, I wish to appreciate the candid support received from academics, professionals and officials of the UGC in our endeavors. I would like to thank Chairman and Vice Chairman of UGC, Chairman of Governing Board and Dr. (Mrs.) M.D. Baby, Organizing Secretary of CALIBER-2005 (also member of Governing Board of INFLIBNET) for their wholehearted support and encouragement. I have special thanks to Dr. P K Abdul Azis, the Vice Chancellor, Cochin University of Science & Technology and his academic & administrative staff for their enthusiasms and support.

I hope that CALIBER-2005 will be a great convention and everyone will enjoy attending it !



1 February, 2005
Ahmedabad

Dr. T A V Murthy

CONTENTS

Committees iii
Preface iv

Theme 1 : Multilingual Computing and Natural Language Processing

1. **Multilingual Computing For Indian Languages - An Overview *(Theme Paper)*** 1
   Shivashankar B. Nair
2. **Intelligent Agent-Based Multilingual Information Retrieval System** 8
   Suman Mary Idicula and David Peter S.
3. **A New Architecture For Braille Transcription From Optically Recognises Indian Languages** 22
   Omar Khan Durrani and K. C. Shet
4. **A Document Reconstruction System For Transferring Bengali Paper Documents into Rich Text Format** 32
   Anirban Ray Chaudhari, Debnath Singh, Mita Nasipuri and Dipak Kumar Basu
5. **Natural Language Requirements to Executable Models of Software Components** 43
   V. R. Rathod, S. M. Shah and Nileshkumar K. Modi
6. **Analysis and Synthesis for Pyramid Based Textures** 50
   V. Karthikeyani, K. Duraiswamy and P. Kamalakkannan
7. **Critical Challenges in Natural Language Processing** 62
   Veena A. Prakashe
8. **UNL Nepali Deconverter** 70
   Birendra Keshari and Sanat Kumar Bista
9. **Preprocessing Alogorithms for the Recognition of Tamil Handwritten Characters** 77
   N. Shanthi and K. Duraiswamy
10. **Performance of Memoized - Most- Likelihood Parsing in Disambiguation Process** 83
    Maya Ingle and M. Chandwani
11. **A New Contour Based Invariant Feature Extraction Approach for the Recognition of Multi-Lingual Documents** 94
    Manjunath Aradhya V N, Hemantha Kumar G, Shivakumara P. and Noushath S.
12. **Current Status & Process in ihe Development of Applications through NLP** 109
    V. R. Rathod, S. M. Shah and Nileshkumar K. Modi
13. **Two-Tier Performance Based Classification Model for Low Level NLP Tasks** 117
    S. Sameen Fatima and R. Krishnan
14. **Globalization of Software Applications Using UNICODE Based Multilingual Approach** 128
    Sonia Dube, Yatrik Patel and T A V Murthy
15. **Enabling Indic Support in Library Information Systems: An Opensource Localizer's Perspective** 132
    Indranil Das Gupta and Najmun Nessa
16. **Multilingual Computing in Malayalam: Embedding the Original Script of Malayalam in Linux and Development of KDE Applications** 146
    Rajeev J. S, Chitrajakumar R, Hussain K. H and Gangadharan N.
17. **Digital Mapping of Area Studies: A Dynamic Tool for Cultural Exchange** 158
    Chitra Rekha Kuffalikar and D. Rajyalakshmi

THEME 2 : Content and Information Management

18. **Technology Enablers for Building Content Management Systems*(Theme Paper)*** 167
Vasudeva Varma

19. **Searching Patent and Patent Related Information on Internet** 178
Sumati Sharma and Mohinder Singh

20. **XFML, Standard for Distributed Information Architecture** 186
Aparajita Suman

21 **Content and Information Management with Special Reference to India** 192
J C Sharma

22. **DLIST: Distributed Digital Management of the Scholarly Publication** 197
Kamalendu Majumdar and U N Singh

23. **Content Management in Digital Libraries** 209
Mohd. Nazi and Faizul Nisha

24. **Knowledge Management in Bangladeshi Libraries: A Long Way to Go** 214
Kazi Mostak Gausul Hoq and M Nasiruddin Munshi

25. **A Brief Evaluation of Search Facililties and Search Results of few Resources Accessible through INDEST Consortium** 221
Kshyanaprava Sahoo and V K J Jeevan

26. **The Needs for Content Management with Special Reference to Manuscripts of Manipur** 230
Th. Satyabati Devi and T A V Murthy

27. **Streaming Communication for Web Based Training** 236
E. Jayabalan, R. Pugazendi and A. Krishnan

28. **Streaming Media to Enhance Teaching and Improve Learning** 244
E Jayabalan, R Pugazendi and A Krishnan

29. **Information Life Cycle Management for LIS Professionalsi in the Digital Era** 249
Ramesh R Naik

30. **Challenges of Multimedia Watermarking Techniques** 253
E. Jayabalan, R. Pugazendi and A. Krishnan

31. **Automatic Ontology Generation for Semantic Search System Using Data Mining Techniques** 259
K. R. Reshmy, S. K. Srivatsa and Sandhya Prasad

Theme 3 : Digital Information Processing and Interoperability

32. **Digital Libraries in Knowledge Based Society: Prospects and Issues *(Theme Paper)*** 271
Om Vikas

33. **Mining of Confidence-Closed Correlated Patterns Efficiently** 290
R Hemalatha, A Krishnan, C Senthamarai and R Hemamalini

34. **Mining Frequent Item Sets More Efficiently Using ITL Mining** 300
R. Hemalatha, A. Krishnan and R. Hemamathi

35. **Temporal Association Rule Using without Candidate Generation** **309**
Keshri Verma and O P Vyas

36. **Platform Independent Terminology Interchange Using MARTIF & OLIF** **318**
M. Ramshirish

37. **Digitization : Basic Concepts** **325**
B. Mini Devi

38. **Features in the Web Search Interface: How Effective are They?** **331**
Deepak P and Sandeep Parameswaran

39. **Mutual Authentication Protocol for Safer Communication** **342**
S. P. Shantharajah And K. Duraiswamy

40. **Effectiveness of Name Searching In Web OPAC: From Authority Control to Access Control** **348**
Veerankutty Chelatayakkot and V. Jalaja

41. **Digital Preservation of Art, Architectural and Sculptural Heritage of Malwa (Madhya Pradesh)** **358**
S. Kumar, Mukesh Kumar Shah and Leena Shah

42. **Digital Preservation of Indian Manuscripts - An Overview** **370**
Y. V. Ramana

43 **A Novel Approach for Document Image Mosaicing Using Wavelet Decomposition** **377**
P. Shivakumara, G. Hemantha Kumar, D. S. Guru and P. Nagabhushan

44. **Enhanced Information Retrieval** **392**
R. Bhaskaran

45. **Preservation and Maintenance of the Digital Library - A New Challenge** **396**
K. R. Mulla, A. S. Shivakumara and M. Chandrashekara

46. **Meta Search in Distributed Electronic Resources: A Study** **404**
M. Krishnamurthy

47. **Web Based Library Services: A Study on Research Libraries in Karnataka** **409**
Vijayakumar M, B. U. Kannappanavar and Madhu K. N

48. **Digital Library of Theses and Dissertations** **414**
G. Rathinasabapathy

49. **Preservation of Digital Cultural Heritage Materials** **420**
P. Lalitha and T A V Murthy

50. **Preservation and Digitisation of Rare Collection of Dr. Panjabrao Deshmukh Smruti Sangrahalaya, Amravati** **428**
Vaishali G. Choukhande and Jitendra Dange

51. **Web Services and Interoperability : Security Challenges** **432**
S. K. Sharma, G. K. Sharma and P. N. Shrivastava

52. **Legal Text Retrieval and Information Services In Digital Era** **441**
Raj Kumar Bhardwaj

Theme 4 : Digital Libraries and Services

53. **Building the German Digital Library Vascoda : Status Quo and Future Prospects** **448**
(Theme Paper) *Tamara Pianos*

54. **Digital Knowledge Resources for Agribusiness Development** **457**
J. P. S. Ahuja and M. R. Rawtani

55. **Networking and Security Aspects in Libraries** **470**
P. Balasubramanian, K. Paulraj and S. Kanthimathi

56. **Role of the Library Homepage as a New Platform for Library Services: A Case Study** **475**
Hemant Kumar Sahu

57. **Wireless Network Connections Policies & Standards** **484**
Atul M. Gonsai, N. N. Jani and Nilesh B. Soni

58. **Library Consortia Model for Country Wide Access of Electronic Journals and Databases** **497**
A. T. Francis

59. **Subject Gateways: An Overview** **505**
R. T. Yadav

60. **Students Attitudes towards Digital Resources and Services in B.I.E.T, Davanagere: A Survey** **517**
Manjunath S. Lohar and Mallinatha Kumbar

61. **Digital Libraries and Services** **526**
K. Paulraj, P. Balasubramanian and S. Kanthimathi

62. **Consortia Developments in Library and Information Centers : Some Issues** **531**
Jayaprakash H and Bachalapur M M.

63. **Enhancing Network Applications in a University Library : A Case Study** **539**
Suresh Jange, R. B. Gaddagimath, Amruth Sherikar and S. B. Policegoudar

64. **Importance of Digital Library for E-Learning In India** **549**
H. S. Chopra

65. **Transformation of Library Services: with Special Emphasis on Digital Reference Service** **553**
Padmini Kaza

66. **Role of Digital Libraries in E-Learning** **559**
Prachi Singh

67. **Subject Gateways - A Case Study of the Science Campus Library, University of Madras** **565**
R. Samyuktha

68. **Role of Information Technology in Ayurveda in the Digital Age** **575**
G. Hemachandran Nair

69. **Institutional E-Print Repositories for Scholarly Communication : Issues and Implications** **580**
B. Maharana, D. K. Pradhan, B. K. Choudhury and S. K. Pathy

70. **Digital Library Management in German University Libraries: The Bochum Perspective** **589**
Erda Lapp

71. **Digital Libraries and Open Source Software** 594
Umesha Naik and D. Shivalingaiah

72. **Building Up Digtial Resources for Effective E-Learning Programmes** 606
T. Rama Devi

73. **Library Portal - A Knowledge Management Tool** 612
Daulat Jotwani

74. **An Improved Hybrid Routing Protocol for Mobile Ad Hoc Networks** 621
P. Kamalakkannan, A. Krishnan and V. Karthikeyani

75. **Student's Perceptions toward the Use of the Digital Library for Higher Learning** 630
A. Manoharan, M. Anuvasanthi and T. Deepa

76. **Is the Big Deal Mode of E-Journal subscription a Right Approach for Indian Consortia? A Case Study of Elsevier's ScienceDirect Use at Indian Institute of Technology Roorkee** 635
Yogendra Singh and T A V Murthy

77. **Familiarity and Use by the Students' of Digital Resources Available in the Academic Libraries of Medical Science University of Isfahan (MUI), Iran** 648
Asefeh Asemi

78. **UGC-Infonet: E-Journals Consortium an Indian Model - Bridging the Gap Between Scholarly Information and End User** 658
T A V Murthy, V. S. Cholin, Suresh K. Chauhan and Raghavendra Patil

79. **Potential Role of Subject Gateways, Portals and OPAC's in Electronic Journals Access** 668
K. Prakash, V S Cholin and T A V Murthy

80. **Security for Libraries in the Digital Networked Environment** 679
Manoj Kumar K and Haneefa K. M.

81. **Transition in Information Services : A Digital Experience** 688
Kalyani Accanoor

82. **Indian Academia on Copyright and IPR Issues of Electronic Theses and Dissertations** 697
J. K. Vijayakumar, T A V Murthy and M T M Khan

83. **Use of Information Source in Digital Environment - A Case Study** 705
D. Rajeswari

84. **Role of Telecommunication and Networking Technology in the Development of Digital Libraries** 712
Mamata P. K, G. Gopal Reddy and P. K. Kumbargoudar

85. **Digital Libraries; A Boon for Information Seekers** 720
Mridulata Shrivastava and Chitra Ingle

86. **Towards the Design and Development of E-Books: An Experience** 727
P. Rajendran, B. Ramesh Babu and S. Gopalakrishnan

Abstracts 735

Author Index 748

Keyword Index 750

Multilingual Computing for Indian Languages - An Overview

Shivashankar B Nair

Abstract

While English has predominantly maintained its lead both as the lingua franca of the Internet and also for basic man-machine interactions, the need of the day is a system that can cater to the native by facilitating such communication in the language he is most comfortable. This calls for the realization of Multilingual Systems that can present the same information in a variety of languages. The term Multilingual Computing in the present context refers to systems that are capable of running programs that accept, process and present data in more than one natural language. The concerned language for interacting with the computer may be selected at the time of invocation or use of the program. Such multilingual systems do not fully overcome the man machine barrier. The user still has to comprehend and react to the cryptic messages presented in the language of his choice. The problem can be overcome only by the use of natural language processing systems that can translate user commands and queries in his native language to machine level commands, facilitate execution and present the results in a manner that is very akin to the responses of a human being. This paper discusses the basic issues in the formulation of such multilingual systems.

Keywords : Indian Language, Multilingual Computing, Natural Language Processing.

0. Introduction

The advent of the Internet and the concept of ubiquitous computing have lead to a plethora of applications of computing systems. Information is now available virtually at every point but in the form of web pages or some such related form via the Internet. Web sites that host information in a wide variety of areas ranging from commercial, scientific, medical to educational and literary aspects are currently available. A person literate in computers does not really feel the pinch of accessing and weaning out the information he desires. On the contrary, a person not well versed in the same faces numerous diverse problems. Since most of these sites host information in English, those comfortable in this language find accessing information a lot easier. The net effect is that most people who wish to use computers have to essentially possess good knowledge of English. Though many countries have now opted for using operating systems in their national languages, accessing information across the World Wide Web still requires knowledge of other languages. The problem is compounded in countries where people speak more than one language. India is possibly one of the best-known examples. With about twenty two official languages, and many more unofficial ones being spoken prominently by large populations, this country faces an uphill task of disseminating information not just in English or its national language (Hindi) but also in a diverse set of languages that are grossly different in their phonetic and linguistic aspects. Keeping this problem in mind the Ministry of Information and Communications Technology of the Government of India [1] took up an initiative in the year 2000 to set up several Resource Centres for Indian Language Technology development across the country, to act as language clinics. Assigned with one or more Indian languages, each Centre aims at developing tools for multilingual computing. These include editors, morphological analyzers, spell checkers, language corpora and the like, all essential aids for a proper multilingual computing and language processing environments.

The recent information technology boom and the widespread use of computers have still not percolated to the common man. One of the prime issues that contributes to this problem is the absence of a user-friendly human-computer interface. The advent of graphical user-interfaces has, to a great extent, made the

use of computers an easy task. But these interfaces have their own disadvantages. They provide near pictographic information of what can be done but provide no more. A user may want to do more than just click on icons. This calls for a higher degree of sophistication - systems that can accept commands in speech and also talk back. Realizing such systems falls under the domain of speech and natural language processing, both fairly simple tasks for human beings but highly complex ones for computer systems. This paper describes the meaning of multilingual computing, the techniques used to realize it and winds up with an overview of language processing and the tools required in such a scenario. The latter being a vast and only partially solved problem domain is only described in brief.

1. Multilingual Computing

As the very term suggests, multilingual computing refers to using the computer to communicate with humans in different languages [2]. The term also encompasses browsing and searching the World Wide Web in different languages, even typing from right to left as in some languages like Arabic and printing on different paper sizes. Issues like transliteration, spellchecking [3] and reading in one language and typing and printing in another also fall in this domain. Multilingual computing also deals with mechanisms for accepting input in the desired language, rendering and storing information. It is not only a software and a hardware problem but also one that deals with conforming to world-wide standards [4]. This is because each language may otherwise use a different mechanism for input, rendering and storage that can lead to problems in compatibility.

2. Representation of Information

Information is represented in the form of codes, the most common being the ASCII (American Standard Code for Information Interchange). ASCII is a 7-bit code that encodes 32 control characters, plus 96 alphanumeric characters (A to Z, 0 to 9 plus symbols). The Extended ASCII is an 8-bit code and facilitates more information representation.

2.1 ISCII

Many Indian languages use the ISCII (Indian Script Code for Information Interchange) code [5] that allows usage of 7 or 8 bit characters. In the 8 bit mode, the lower 7 bits (128 combinations) comprise the ASCII character set while the upper 128 characters cater to the Indian scripts. In the 7-bit mode the control character SI is used for activating the ISCII code set. A script is a set of symbols required to represent a writing system. It can be used to represent many languages. For instance the Latin script is used to represent languages like English, French and German. Most Indian languages have evolved from the Brahmi script. They have a common phonetic structure making a common character set viable.

ISCII has been christened lately as ACII (Alphabetic Code for Information Interchange) and now caters to scripts of SAARC countries. ACII is an 8-bit code that has the ASCII character set on the bottom half while the upper half is occupied by ACII characters. In the Indian context, ACII accommodates around 10 Indian scripts including Assamese, Devnagiri, Malayalam and Punjabi. The basic characters have been positioned to enable direct sorting.

It is thus obvious that 8-bit character codes are adequate for languages which have a small alphabet set and when written text in these languages comprise of individual alphabet and punctuations. Things however become difficult for languages, which do not satisfy these conditions. This is true for Indian languages, which have conjunct characters that express combinations of sounds. Data entry thus becomes a problem in such cases. Before displaying the final version of the character (or conjunct), the terminating vowel has to be determined so as to generate the shape to be rendered. Thus an algorithm to tackle variable number

of bytes, comprehend and translate them to a shape formed using one or more glyphs is required. In case of Roman letters such a problem does not arise as each byte corresponds to a unique glyph.

To make things more vivid we take a look at a simple example.

Input String System Alphabet followed by the respective Codes

trupti ASCII 116(t) 114(r) 117(u) 112 (p) 116(t) 105 (i)

ISCII 264(cT) 258(0)184(^)227(0) 176(*T) 258(O) 264(cl) 223(6t)

It can be seen that the characters for the ISCII code by themselves carry no meaning. A program has to find the sequence of codes entered, comprehend the order and finally aggregate the associated glyphs and send them to the rendering device. The string however may be stored as a chain of associated ISCII codes.

2.2 Unicode

Of late the problem of multilingual computing has become an international issue. More number of scripts and languages has compounded the problem. New attempts were therefore made for standardization of multilingual documents. A new code to support all languages was looked into. A new code termed the Unicode was evolved in the year 1991. Unicode is an extension to the ASCII code but also accommodates International languages and scripts. Most international languages require only 7 or 8-bit character codes, as they need only a small set of symbols to represent the letters. Unicode permits the use of this 8-bit representation but augments the code with an extra 8-bit language identifier. These extra 8 bits that form the most significant byte of the 16-bit Unicode can cater to 256 different languages and about 128 characters per language. Thus Unicode compatible software can identify the language of each character and also use the appropriate rendering system. For Indian languages Unicode has greatly conformed to ISCII. Minor variations however do exist for some languages. The embedding of ISCII has resulted in the fact that Unicode inherently bears with it all drawbacks posed by ISCII. A more comprehensive discussion on the limitations of Unicode and ISCII can be read from [6].

3. The Concept of Fonts

The previous paragraphs discussed how text is represented in a coded form and interpreted for rendering. Fonts provide for displaying the symbols on the rendering device. From the computer's point of view, a font is a file (or files) required to display and print in a particular style. A set of fonts that are similar in looks but have different attributes is termed as a font family. Fonts may be of various types. A bitmapped font, at times referred to as a screen font, contains all information regarding the pixels. This information is used by the rendering software to display the font on the screen. Printer fonts generally referred to as PostScript fonts comprise of more than one file. These files include the bit mapped screen font files and a printer font. The former are used for display and the latter by a PostScript compatible printer. The advantage is that printouts are smooth unlike those made using bitmapped fonts. TrueType fonts allow smooth screen displays and facilitate printing without the use of extra screen font sizes or PostScript. They can print on non-PostScript printers. Another category called Dynamic fonts allows delivery of TrueType fonts to the client side.

Fonts are generally stored in an OS specified directory called the font folder. At the time of booting the system checks for these fonts and activates them. Alternatively a font manager may be used to activate or deactivate fonts. When a key is pressed, the keyboard driver looks into the code generated by the keystroke, interprets the same and passes it onto the rendering software. This in turn refers to the font files, grabs the information about the associated glyph and renders it on the screen. As mentioned earlier rendering

conjuncts requires an extra element of processing to aid the exact rendering on the screen. A point that may be emphasized here is that keyboard layouts may also vary. A Romanized keyboard layout uses phonetic English mappings to compose the text. For example, the keystrokes *aa* or A would give the *matra* for the corresponding language. The Typewriter Layout on the other hand is structured based on the normal Hindi Typewriter layout. This enables typists to easily adapt to the keying system. The Phonetic layout, standardized by the Department of Electronics of the Government of India, has the same layout for all Indian Languages. This also facilitates transliteration and ease of typing across languages.

The commonest way of displaying Indian language text in Web pages is the HTML based approach. The problem here is that some browsers may not support the encoding of the specified font. It may happen that the user has to manually switch to the correct encoding. Another approach comprises of using dynamic fonts wherein the fonts specified in the pages are sent along with it. However there are restrictions again in the type of encoding used. They generally work only with true type fonts and may not be rendered properly in Unix systems.

4. Transliteration

This is the processes of converting each alphabet in one language to the equivalent in another. Transliteration is advantageous in the Indian language context due to the common phonetic structure possessed by these languages. Unlike in English, the *aksharas* (mildly equivalent to the alphabet) in Indian languages, which refer to sounds, are pronounced the same way irrespective of their position in the word. This property facilitates easy conversion from one Indian language to the other. Thus if a proper noun or an address or date were written in an Indian language which the user does not comprehend, he can easily transliterate and read the same in a language he is well versed with. For instance a title of a book written in Marathi can be read in Assamese. This also helps people learn a new language.

5. Natural Language Processing

Natural Language is the language used for communication amongst human beings in the real world. The term *real world,* makes the problem much more difficult. If we were to converse using telegraphic language, building systems that understand such conversations would become a much simpler task. For instance consider understanding the two sentences -

1. We all should use telegraphic language.
2. Use telegraphic language.

It is apparent that the second sentence is more precise and extracting its semantic content is obviously easier and quicker. It is thus obvious that the main aim of natural language understanding systems is to translate natural language sentences (of type 1) into a comparatively simpler form (of type 2). Type 2 sentences are finally used by systems aided by some *a priori* or gained logic to comprehend and execute desired actions.

While the term Natural Language (NL) refers to the language spoken by human beings, Natural Language Processing (NLP) refers to an area of Artificial Intelligence (AI) that deals with systems and programs that can accept, comprehend and communicate in natural language. Systems that are capable of processing and understanding natural language bridge the man-machine communication barriers to a great extent. They facilitate interaction with computers without resorting to the memorizing of complex commands. Computational Linguistics (CL) is a closely related and often talked of field in conjunction with Natural language processing. It uses knowledge of both Linguistics and Computer Science to study the

computational aspects of the human language faculty. There is thus a major overlap in the areas AI, CL and also Cognitive Science.

While speech recognition systems convert the speech input to a textual representation, natural language processing systems perform the job of understanding the meaning of the input. NLP entails several phases. As part of the initial processing stage, a morphological analyzer does the job of finding the root words of each word or token within the sentence. These root words along with the associated morphemes allow us to understand the correctness of the word and also the inflections. This phase also facilitates spell checking and correction. The second phase of processing is carried out by a syntactic analyzer, often referred to as a parser. It verifies the grammatical correctness of the sentence and finds whether or not the various words conform to the grammar of the language. Several types of grammar representations and associated parsers exist [7,8], the more common one being the context free grammar. The grammatical representation thus generated is then used by the semantic analyzer to wean the actual meaning of the sentence. The analyzer uses semantic and pragmatic knowledge in the extraction. In most cases this is a very domain specific job.

6. Multilingual Computing and Natural Language Processing

Though definitions of Natural Language Processing do not really include Multilingual Computing, experience has shown that realizing NLP systems across languages requires more than just understanding the linguistic, morphological, grammatical and semantic aspects of the language. From the Engineering perspective, it requires everything available in the domain of multilingual computing. From the perspective of Multilingual Computing too the understanding of many language-related issues such as phonetics, linguistics, etc. are highly necessary to fix and standardize the manner in which the system is to work and deliver. The point being emphasized here is that the two areas are greatly interlinked and issues in one significantly affect the other. Efficient machine translation systems can be formulated but their viability and scalability depend on the quality of the multilingual computing platform.

Almost all linguistic theories and mathematical and computational models are formulated so that they are in principle applicable to natural languages at large, and not just to one or more specific languages. Computational models that deal with more than one language form the field of multilingual NLP. Machine translation (MT) is a sub-category of this field. Conventional MT systems try to produce one translation for the source sentence, but, of course, in order to achieve this, as is well known, all kinds of other information are usually needed. This includes discourse context (i.e. the previous text), the context of the situation (particularly if it is a dialog), and a variety of domain-knowledge, exactly the kind of information that a natural language understanding and generation system needs. Though a clear-cut solution to real MT is definitely not around the corner, there are systems constrained to specific domains that have been attempted and even put to use. Generation of multilingual texts from a source text that is prepared in a greatly constrained, unambiguous, highly stylized language for instance is one such area. These systems form tools to translate manuals written in one language to several others. There are others like the TAUM METEO system that translates weather reports from French to English. The system has been in operation for quite some years in Canada. In weather reports the sentences are usually short, highly stylized and mostly employ a standard phraseology. SYSTRAN is another effort to translate between several European languages.

In India too there have been serious attempts at machine translation. Anglabharati [9] is one such English to Hindi translation system currently available on the web. The system uses the *Interlingua* technique for translation and is thus scalable. The English sentence is first translated into an *Interlingua* termed as PLIL (Pseudo Lingua for Indian Languages). A text generator then transforms it into its equivalent translation in the target language, Hindi in the present case. Thus if text generators for other languages are written the same PLIL can be converted into the language of choice facilitating one to many language conversion.

Attempts are currently being made by many Resource Centres across India to write such text generators for the assigned languages.

7. Tools for Multilingual Language Processing

Multilingual Language Processing calls for the existence of several tools required at various stages of processing. One or more tools may be instantiated at each phase of processing. Some of the more important ones have been briefly described below.

i. Corpora

A *corpus* is a collection of a large number of pieces of a language stored in a standard form based on some explicit linguistic criteria. They are used to serve as a sample of the language. Several types of corpora exist. A *tagged corpus* contains words with their parts of speech alongside. *A. parallel corpus* contains a collection of texts translated into one or more languages other than the original. In its simplest form it comprises of two languages with one of the corpora containing the exact translation of the other. Such corpora are useful in translation from one language to the other.

ii. Morphological Analyzers

These analyzers are required to find the stem or the root of a word. By doing so a word is split into its basic parts called morphemes. Such analyzers require a high amount of linguistic information to be embedded. Thus construction of such analyzers entails collection and representation of knowledge provided by expert linguists in the concerned language. The process aids in finding whether the word is a valid one in the language.

iii. Spell Checkers

Spell Checkers form a vital tool and perform the job of detecting errors in the text and suggest the relevant corrections. They use a lexicon or a dictionary of words or even the corpus and together with the morphological analyzer perform the jobs of detection and correction.

iv. Parsers

These enable checking the grammar of the sentence in question. Panian grammar used for Indian languages has been described in [10].

v. Multilingual Dictionaries

These dictionaries provide for machine translation by providing the equivalent words, category, etc. in the target language. Since no dictionary representation standards have evolved so far the formats in which data is stored have to be known a priori.

8. Conclusion

The mythological problem caused by the Tower of Babel culminating in the Confusion of Tongues has found its way into the era of computing. Large amounts of information in numerous languages have forced us to think seriously of standardizing methodologies for information representation. The birth of Unicode may pave the way to a better future for greater flexibility in the use of multilingual computing platforms. Natural language processing, a vastly incomplete area of research, may reap the benefits of such systems.

9. Acknowledgement

The author wishes to express his gratitude for the support received from the Resource Centre for Indian Language Technology Solutions at the Indian Institute of Technology (a project funded by the Ministry of Information & Communication Technology of the Government of India), as also Samir Borgohain and Monisha Das, Project Personnel at the Centre, in the making of this paper.

10. References

1. http://tdil.mit.gov.in
2. http://LanguageLab.csumb.edu
3. Kukich, K., Techniques for Automatically Correcting Words in Text, ACM Computing Surveys, Vol.24, No.4, December 1992, 377-439.
4. Goodwin-Jones, R.; "Emerging Technologies Language & Learning", Language Learning and Technology", Vol.6, No.2, May 2003, pp.6-11.
5. Viswabharat@tdil, Language Technology Flash, Jan 2002.
6. http://acharya.iitm.ac.in/multi_sys/uni_iscii.html
7. www.link.cs.cmu.edu/link/
8. Joshi, A.K., "Natural language processing", Science, Vol.253, No. 5025, September 1991, pp. 1242-1249.
9. Sinha, R.M.K.; "Machine translation : an Indian perspective", Proceedings of the Language Engineering Conference. LEG 2002, 2003, IEEE Computer Society, Los Alamitos, CA, USA, pp. 181-182.
10. Bharati, A.; Chaitanya, V.; Sangal, R.; "Natural language Processing - A Paninian Perspective", Prentice hall of India, 1995.

About Authors

Dr. Shivashankar B Nair is Associate Professor in Department of Computer Science and Engineering at Indian Institute of Technology, Guwahati, Assam.
E-mail : sbnair@iitg.ernet.in

Intelligent Agent-based Multilingual Information Retrieval System

Sumam Mary Idicula David Peter S

Abstract

The goal of this work is to develop an Open Agent Architecture for Multilingual information retrieval from Relational Database. The query for information retrieval can be given in plain Hindi or Malayalam; two prominent regional languages of India. The system supports distributed processing of user requests through collaborating agents. Natural language processing techniques are used for meaning extraction from the plain query and information is given back to the user in his/ her native language. The system architecture is designed in a structured way so that it can be adapted to other regional languages of India.

Keywords : Information Retrieval, Natural Language Processing, Multilingual Computing, Indian Languages.

0. Introduction

Information is playing an increasingly important role in our lives. Information has become an instrument that can be used for solving problems. An intelligent information agent is one that acts on behalf of its user for information gathering [1]. It is capable of locating information sources, accessing information, resolving inconsistencies in the retrieved information and adapting to human information needs over time. An information agent is responsible for providing intelligent information services. Information market has now been transformed from supply-driven to demand-driven.

India being a multilingual country, information agents have to work in multilingual environment. To handle multilingual environment, specific requirements and good language knowledge is needed. Every language has a set of language resources like characters, lexicons, grammars and keyboard mappings. Agent should be open to updates in language resources. An information agent should be able to configure to the user preferred language and usage style. An agent should exhibit the capacity to learn new languages to increase the degree of multilingualism.

A multilingual information agent should have properties like

- Language autonomy- capability to control state and behavior according to the language.
- Reactivity : The ability to perceive and respond to multilingual percepts.
- Pro-activity : Ability to perform language oriented goal-directed behavior.
- Social–ability : Ability to exhibit interact in language sensitive environments.
- Learning : Ability to adapt to user's language preferences and new languages.

1. Objectives & Motivation

This work is aimed at developing an intelligent agent based system for information retrieval from database. Database hold huge quantities of data. There are several artificial languages for manipulating the data. But their usage needs knowledge about the database structure, language syntax etc. Several Natural language front-ends have been developed as a result of rigorous works done in this field of artificial intelligence. But majority of them use English as the natural language. India being a multi-lingual country

with only 5% of the population having education up to matriculation level, their use is limited. In this context, information retrieval in regional languages from database has tremendous impact. It is of very great use in application areas like e-governance, agriculture, rural health, education, national resource planning , disaster management , information kiosks etc. Even research organizations can make of use of such systems to bring their findings to common man.

The agent based information system developed can retrieve information in Hindi and Malayalam from the National Resource Information database. The user can give his/ her query to the system as if he/ she delegates a human being for information gathering. There is no rigid syntax for asking the query. The system using NLP techniques tries to extract meaning of the query and retrieve information from data stores and present the information to the user in his/ her own native language [2]. This system provides the following advantages to the user.

- the user can communicate with data stores in the regional languages
- flexibility in communication
- shielding the users from complexities of database model, DDL, DML etc
- handling anaphoric and elliptical queries
- capability of handling user misconceptions

2. System Design

The information retrieval system is implemented using Open Agent Architecture (OAA). The OAA is a framework in which a community of software agents running on distributed machines can work together on tasks assigned by human or non-human participants in the community [3]. The system is configured as a multi-agent system. The agents collaborating in the system are user-interface agent, morph-analyser agent, parser agent, SQL agent and facilitator agent. The communication language used among the agents are Interagent Communication Language (ICL). Fig.1 presents the basic architecture of the system.

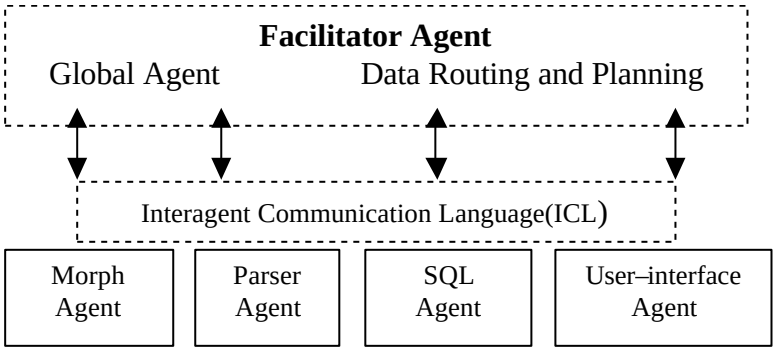


Fig 1 . System architecture

These agents are distributed over a network of machines. There is only one database server. Multiple users can use the system at the same time. For running the system all the machines should have J2DK 1.4.0 and OAA software package. All the client machines should have the User Agent loaded in them.

2.1 User Interface Agent

It accepts queries from the user. Query can be input through the key board or through the window keypad with the help of touch pen. Both Malayalam and Hindi languages have been supported.

Seeing the difficulties of input/output in Indian languages, we decided to develop some convenient approaches to tackle this problem. They include

- Keyboard
- Window Keypad & Touch pen/Mouse

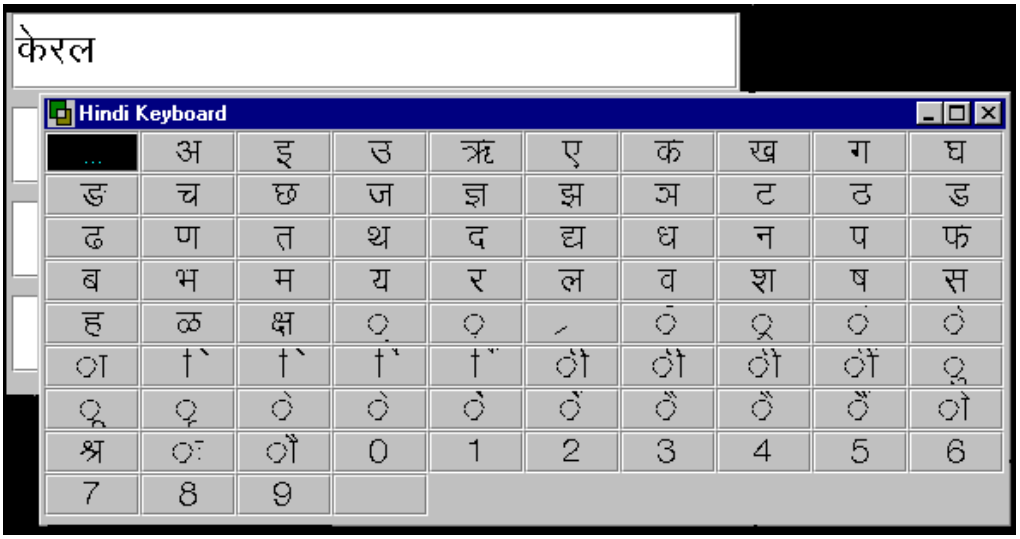
2.1.1 Keyboard Input

The commonly used method for entering text into the computer is with the help of a keyboard. Most of the computers are provided with QWERTY keyboard, which do not have enough keys to support all letters of Indian languages. So better utilization of available keys along with well-designed key assignments play an important role in the usability of keyboard in Indian context. As most of the users are well versed in using English keyboard, we decided to map Indian letters to English keys based on extent of similarity in pronunciation. For example Hindi letter 'Û ' is assigned to key 'k', letter '•Ö' to key 'j' etc. We are inputting joint letters with the help of key combination, for example the typing of the sequence › + Ë + œ will result in ù. The system will check the key sequences and will do automatic transformation. This technique will help to enter all letters in Indian languages within the current keyboard limitations.

Another problem in Indian language software creation is the rendering of characters. The standard used for information exchange is ISCII, which is an 8 bit-character encoding scheme where the Indian language characters are assigned to a unique ASCII value within the range of –127 to –1. A font that satisfies ISCII standard can render letters without any problem. But popular programming languages like Java use two byte Unicode characters. So if we want to render Indian letters in Java GUI controls, we have to map ISCII to Unicode. Now Unicode is emerging as an International standard for all languages. There is a Unicode standard for all Indian language scripts, Devanagari (U0900) ranges from 0900 to 097F and Malayalam (U0D00) ranges from 0D00 to 0D7F. As it's found as difficult to get fonts that satisfies both ISCII & Unicode we are doing a font specific mapping The Indian fonts used for Hindi and Malayalm are DVBW-TTYogeshEN and MLB-TTIndulekha respectively developed by C-DAC.

2.1.2 Entering Text Using Window Keypad & Touch pen / Mouse

The problems in key assignments and lack of sufficient keys to support Indian languages force as to provide a Keypad window with the text controls. The Keypad contains buttons for all letters for a particular language and users can activate the Keypad by right clicking on any text control. The users can enter text by selecting buttons on the Keypad with the help of Mouse or Touch pen. Even though this is not as flexible as typing from Keyboard, this will provide an additional feature mainly for the users who are not well versed with the key assignments. The end-users can either use the keyboard or the Keypad to enter text to a GUI Text component. The sample view of Hindi text control and Keypad is given below.



2.2 Morph Agent

Morphological analysis is the process of finding the root and the grammatical features of a word . This agent performs activities like splitting the query into individual words, spelling correction, domain dependent word grouping and attaching semantic properties to each word group. In this work the Morphological Analysis of words is conducted in a domain (database) specific way rather than a language specific way. For example consider the string “greater than”. In database specific sense it is a relational operator, but in language sense the words “greater” & “than” come under different grammatical categories and later we have to combine these to form a relational operator that requires one more processing step. So the main advantages of domain specific analysis are: -

- It will be easier to tokenize the incoming user query.
- It is the natural way in which human experts are performing.

Here the token represents a “meaningful unit” in database context and the morphological analysis is the process of identification of tokens present in the natural language query with the help of a lexicon. Knowledge content of Tokens are stored as frames. The output of morphological analysis will be a table that contains tokens.

Each token have following slots: -

- Tag
- Id
- Value
- Optional

Tag is a string that implies the database specific category of the token. A list of most significant tags & interpretations are given below in Table 1.

Tag	Interpretation
<CFLD>	Character attribute of a database table
<NFLD>	Numeric attribute of a database table
<CVAL>	Character value
<NVAL>	Numeric value
<FUNC>	An SQL function
<AND>	The logical operator AND
<OR>	The logical operator OR
<COND>	Relational operator

Table 1 Token Table

The tag labels are selected to imply database specific meaning. Id is an integer that along with tag will uniquely identify a token in a token table. The value is the actual content of the token and optional is a location that can be used to store additional information if necessary. The tag sequences that we can extract from the token table can be used for syntactic and semantic analysis. The Morph agent makes use of a Lexicon and Vibakthi Lexicon for its processing.

Lexicon Structures : The objective of the lexicon design is that it should contain sufficient information about words to fill the necessary slots of the token. So the lexicon should also contain almost similar attributes as that of tokens. Each record in a lexicon table contains following set of attributes: -

- ? Word
- ? Tag
- ? Meaning
- ? Optional

The word attribute is a string containing single or multiple words that form a domain specific “meaningful unit”. Tag is having similar purpose as in the case of token and meaning is the domain specific meaning of the word and optional is for additional information about word if required.

A sample filling of lexicon is given in Table 2.

Word	Tag	Meaning	Optional
•Ö®ÖÃÖÓÜµÖÖ	<NFLD>	Demography.totp	
†î¸	<AND>	AND	
ÃÖ²ÖÃÖê †×¬ÖÛú	<FUNC>	MAX	
ÃÖê ÃÖ×¬ÖÛú	<COND>	>	
Ûêú¸»Ö	<CVAL>	Kerala	State.sname

Table 2 Lexicon

2.2.1 Vibhakthi Lexicon

The lexicon structure given above is enough for Hindi morphological analysis, but in Malayalam and other Dravidian languages the Vibhakthi is always attached to the words and will play a significant role in understanding the relation between words in a string. The following example will demonstrate the difference between Hindi & Malayalam: -

For a Hindi string: Ûêú¸»Ö ´ÖêÓ

Malayalam equivalent is: ¶‰¥¨Ì°Þ

Which is a combination of: ¶‰¥¨¹ and ƒÞ

From the above example it is clear that for doing morphological analysis in Malayalam there is a requirement for additional information, which will help to extract word & vibhakthi from a combined word. The conventional way of doing this is either by storing all forms of a word or by storing the type of transformations occuring while adding each vibhakthi. But in our problem the NRI database contains thousands of words that represent proper nouns such as state names, district names, village names etc. We have to address the problem of expansion (for example addition of a new state) and modification (for example the re-naming of an existing place). So following the conventional way will lead to a situation that, once a new word is added to database we have to add that word and vibhakthies manually to lexicon. Another factor of consideration is the size & complexity of the lexicon. To address the above-mentioned problems we have designed an effective strategy by building an additional lexicon for Malayalam morphological analysis. The lexicon structure and sample records are given in Table 3.

Word1-Word2	Word2-Word1	Vibhakthi Tag
¹	µÌ¾¯à	<KAL>
¹	Ì°µú	<INTE>
¹	Ì°Þ	<IL>
¸	°µú	<INTE>

Table 3. Vibakthi Lexicon

The Word1 is the root word that is present in the lexicon and Word2 is the word that is present in the user query. The first attribute is the result of Word1 *minus* Word2 and second attribute is the result of reverse process. The vibhakthi tag is the identifier for a particular vibhakthi. The *minus* operator performs a character level comparison from left to right and if there is a miss match between letters, it will add the letter of the first word to difference. The following examples will give the clear picture: -

Word1	¶	‰	¥	¨	¹		
Word2	¶	‰	¥	¨	Ì	°	Þ
Word1 - Word2	-	-	-	-	¹	-	-
Word1	¶	‰	¥	¨	¹		
Word2	¶	‰	¥	¨	Ì	°	Þ
Word2 - Word1	-	-	-	-	Ì	°	Þ

Only the root word is stored in the lexicon and all valid transformation rules are stored in the vibakthi lexicon in the form as narrated above. So the word in the user query is compared with the words in lexicon and if it satisfies the above equations then we can find the root word (word in lexicon) and the vibhakthi attached to it. The completeness of the vibhakthi lexicon determines the accuracy of the whole process. The noted advantages of this method are: -

- ? As the same vibhakthi transformation rule can be applied to a class of words, the size & complexity of the lexicon is reduced considerably.
- ? In the event of a database record modification, the system can automatically check for new words by querying the database to get values of all attributes that are having string data type.

Automatic Spell Checking & Correction : In an effort to give more flexibility to the end-user, the system will perform a word level automatic spell checking and will correct the spelling if there is a reasonable level of matching. The spell checking is performed with the help of lexicon. The words in the incoming query are compared with the words that are present in lexicon. Note that the lexicon contains string of words in the *Word* field, but the spell checking is performing at individual word level. So there is a necessity to build a linear list of words from lexicon. The process of spell checking will proceed by performing the following steps: -

- ? Check whether the word represents numeric value.
- ? Check for an exact word match
- ? Check for Word + Vibhakthi (For Malayalam only)
- ? Check for a reasonable match
- ? Mark word as *JUNK*

The process will systematically perform each step and if all the initial four steps fail then the word will be considered as a junk word. The incoming word that is having more that 2/3 match to lexicon word(s) is considered as a candidate for spell correction and is replaced with maximum matched lexicon word. A bi-directional checking along with size comparison is required to find the extent of matching between two words.

Word Grouping : This is the final phase of Morphological Analysis. After spell checking & correction the refined words are grouped together to form database specific tokens. Token may contain a single word or group of words. As said earlier tokens are the 'meaningful units' that help in domain specific processing. Word grouping identifies all the tokens that are present in the user query. The algorithm will handle possible ambiguities that might arise during grouping. The algorithm initially check for a valid token by grouping maximum number of words and if no match is found then will proceed the checking by eliminating last word from the group. For example after word grouping the sentence "Ûêú¸»Ö Ûêú ÃÖ²ÖÃÖê †×¬ÖÛú •Ö®ÖÃÖÓÜµÖÖ" will produce following set of tokens: -

Token	Ûêú¸»Ö	Ûêú	ÃÖ²ÖÃÖê †×¬ÖÛú	•Ö®ÖÃÖÓÜµÖÖ
Tag	<CVAL>	<KE>	<FUNC>	<CFLD>

The word grouping will produce a token table that contains all necessary information about tokens, which itself is the final output of the entire Morphological Analysis task. The tag string <CVAL><KE><FUNC><CFLD> is used for doing syntactic and semantic analysis of the user query. The additional information about the tokens are used for SQL generation.

2.3 Parser Agent

The syntactic analysis is the process of checking the validity of a sentence. It checks whether the sentence is grammatically correct . Here pattern driven parsing is used. Each pattern has got specific database mapping meaning. These patterns are then mapped to columns & conditions in a database.

The parsing is with the help of a set of production rules. The production rules contain natural language pattern as antecedent and category as consequence. The incoming query is split to form a tree, which contain patterns found in the natural language query. Since our main aim is to find the conditions and columns for generating SQL, we are having only two category items. They are <CONDITION> and <COLUMN>. If any pattern is not following any of the antecedent of a production rule then it will be treated as a <FILLER>. The Table 4 gives some of the production rules used: -

Pattern (Antecedent)	**Category (Consequence)**
<NFLD><NVAL><COND>	<CONDITION>
<FUNC><NFLD>	<COLUMN>
<CVAL><CFLD>	<CONDITION>
<CFLD>	<COLUMN>
<FUNC><CFLD>	<COLUMN>
<CVAL>	<CONDITION>
<NFLD><NVAL><NVAL><COND>	<CONDITION>
<FUNC><NFLD>	<COLUMN>

Table 4 Antecedents & Consequents of production rules

After parsing the tokens that match with the antecedent of a production rule, they are grouped together to form a node. The node contains information about the pattern, category and participating tokens, which is the final output of the overall understanding process. These nodes are passed to the next module for SQL generation.

The natural language understanding is done by the collaborative work of all the above-mentioned agents. The natural language query will go through each of these agents and undergoes certain type of transformation/processing. The final output is used for SQL generation.

Example

The overall understanding process can be explained with the help of an example as shown below: -

Consider the following natural language query after spell correction: -

×ÛúŸÖ®Öê ÝÖÖÑ¾Ö ´ÖêÓ •Ö®ÖÃÖÓÜµÖÖ 2500 ÃÖê †×¬ÖÛú Æî

The word grouping will result in following set of tokens: -

[×ÛúŸÖ®Öê] [ÝÖÖÑ¾Ö] [´ÖêÓ] [•Ö®ÖÃÖÓÜµÖÖ] [2500] [ÃÖê †×¬ÖÛú] [Æî]

The tagging is done as follows: -

Token	×ÛúŸÖ®Öê	ÝÖÖÑ¾Ö	´ÖêÓ	•Ö®ÖÃÖÓÜµÖÖ	2500	ÃÖê †×¬ÖÛú	Æî
Tag	<FUNC>	<CFLD>	<MEM>	<NFLD>	<NVAL>	<COND>	<HY>

The token table produced as a result of Morphological Analysis is given below. The id is used for identifying tokens of similar category. For example if one more function is available then it will be having an id '2'.

Tag	Id	Meaning	Optional
1	<FUNC>	1	COUNT
2	<CFLD>	1	Village.vname
3	<MEM>	1	mem
4	<NFLD>	1	Demography.totp
5	<NVAL>	1	2500
6	<COND>	1	>
7	<HY>	1	hy

The parsing is done on the tag sequence: -

<FUNC><CFLD><MEM><NFLD><NVAL><COND><HY>

Parsing results in identification of following set of patterns, which is used by the SQL generator: -

Pattern	Category	Tokens
<FUNC><CFLD>	<COLUMN>	Tokens 1 & 2
<MEM>	<FILLER>	Token 3
<NFLD><NVAL><COND>	<CONDITION>	Token 4,5 & 6
<HY>	<FILLER>	Token 7

2.4 SQL Agent

This agent generates SQL equivalent of the user inputted natural language query. After the understanding task, we will get some clear indications of required columns & conditions in the final SQL. The SQL is generated based on the underlying database structure and set of expert rules for query building. So there should be provisions for: -

- Getting information about the underlying database structure, which includes meta-data of individual tables
- Building SQL by interpreting the meaning of particular natural language pattern and with the help of database specific information.

2.4.1 Understanding Database Structure

As many tables are there in the database, the type of relation between tables and the conditions for joining tables have significant effect on the final SQL. Also there should be sufficient data about the individual tables such as fields, data type, constraints etc. The table structures shown in Table 5 and Table 6 are used for SQL generation. The Table 5 is used for storing information about individual tables and Table 6 is used for storing information about the relation between tables in the database.

Database Table Field	Database Table	Data Type	Constrain
State.scode	State	CHAR	PKEY
State.sname	State	CHAR	
Demograhy.totp	Demography	NUMBER	

Table 5 Metadata of database tables

Database Table 1	Database Table 2	Table Join Condition	Relation Type
State	Village	State.scode=Village.scode	Master-Detail
Village	Transport	Village.vcode=Transport.vcode	Detail-Detail
Transport	Landuse	Transport.vcode=Landuse.vcode	Detail-Detail
State	Landuse	State.scode=Landuse.scode	Master-Detail

Table 6 Table Join Conditions

The Tables 5 & 6 will provide database specific information to the SQL generator as and when required.

2.4.2 Building SQL

Interpretation of the natural language patterns that we received as a result of parsing is required for generating SQL. The system will have to interpret the expert rules that are having a format similar to the one that is given below: -

```
IF (pattern = “Some natural language pattern”)
   {
   ____________________________________
   Actions for building SQL
   ____________________________________
   }
```

For example: -

```
IF (pattern = “<FUNC><NFLD>”)
   {
   ADD ‘<FUNC>(<NFLD>)’ TO COLUMNS;
   ADD TABLE(<NFLD>) TO TABLES;
   }
```

There are two methods for storing and interpreting the expert rules. These are: -

- Hard-code the rules in a programming language
- Create rule-interpreter for interpreting rules stored in a rule-base.

The first method is having the advantage that it is easy to implement. But if we want to add new rules then the only way we could do is by making changes to the program itself. If we are using the second method, we can add/modify rules in the rule-base without making any modification to the program. The rule-base can be created as an XML document, so that the processing can be done easily with the help of already available XML parsers.

The rule-base will look something similar to the one given below: -

```
<Rule-Base>
   <Rule>
      <Pattern> FUNC NFLD </Pattern>
      <Action>
         <Exec> ADD 'FUNC(NFLD)' TO COLUMNS</Exec>
         <Exec> ADD TABLE(NFLD) TO TABLES</Exec>
      </Action>
   </Rule>
</Rule-Base>
```

The rule-base given above is the simplified form, the original one will embed programming language codes such as function calls inside the <Exec> </Exec> tags. The rule interpreter will parse this XML document and will execute the statements in the <Exec> </Exec> tags. The great aspect of this approach is that in future we will be in a position to formulate some meta-rules for producing rule-base for a new database. The meta-rules contain instructions to build rule-bases for different database structures. So it will be possible to produce configurable natural language interface (NLI), which will reduce the time & effort required for producing interfaces separately.

2.5 Facilitator Agent

The facilitator agent is a server agent that is responsible for coordinating agent communication and control and for providing a global data store to its client agents. It maintains a registry of agent service and data declarations [4]. The facilitator agent breaks tasks into subtasks, matches subtasks to service providers, routes and collects information among distributed participants.

2.6 Interagent Communication Language (ICL)

ICL is the interface language shared by all agents, no matter what machine they are running on or what computer language they are programmed in. Every agent participating in the system defines and publishes a set of capabilities specifications expressed in the ICL, describing the services that it provides [5,6]. These establish a high-level interface to the agent, which is used by a facilitator in communicating with the agent and delegating service requests to agents. The capabilities are referred as "solvables". In this work parsing a token string is a solvable for the parser agent. Displaying the output on the screen is a solvable for the user-interface agent. While performing a task, an agent can request the service provided by other agents through the procedure "Solve( )"

3. Results

Some sample screen shots of the multilingual information retrieval system are given below. Since we have only developed a prototype of the system, we haven't populated the database considerably. Hence only few records are seen in the output. The User Agent displays the results at the client machines.

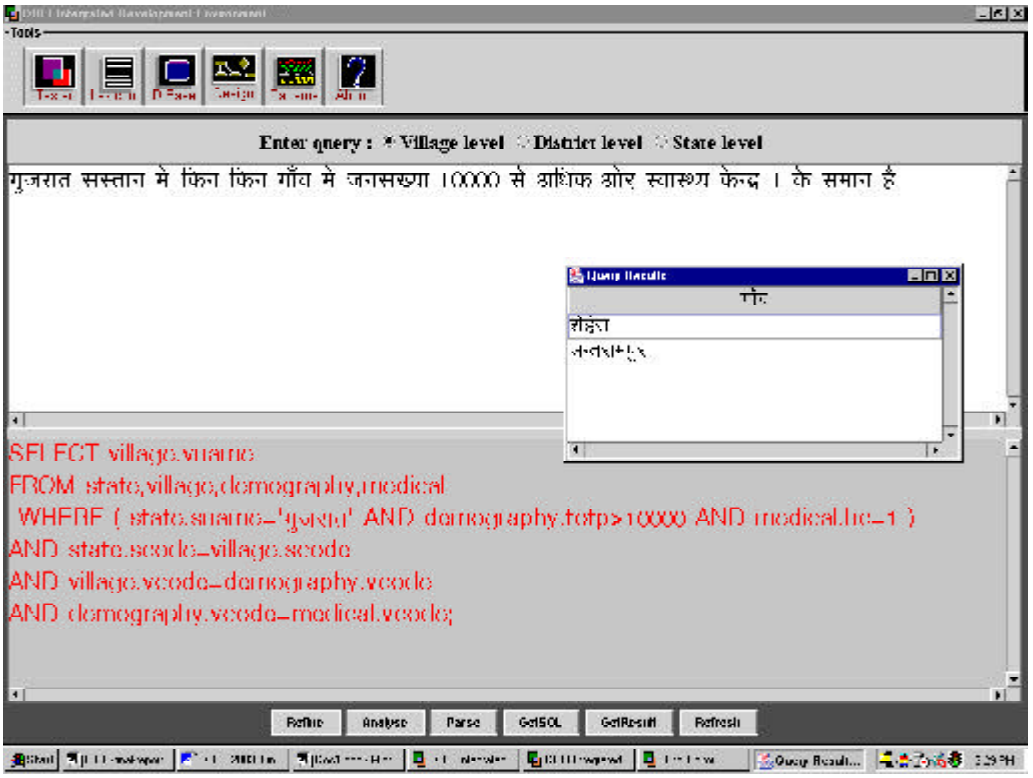
The screen shots display the results of queries given in Hindi and Malayalam. The user typed query and the corresponding SQL statement generated by the SQL Agent and the result obtained can be seen in the screen shots.

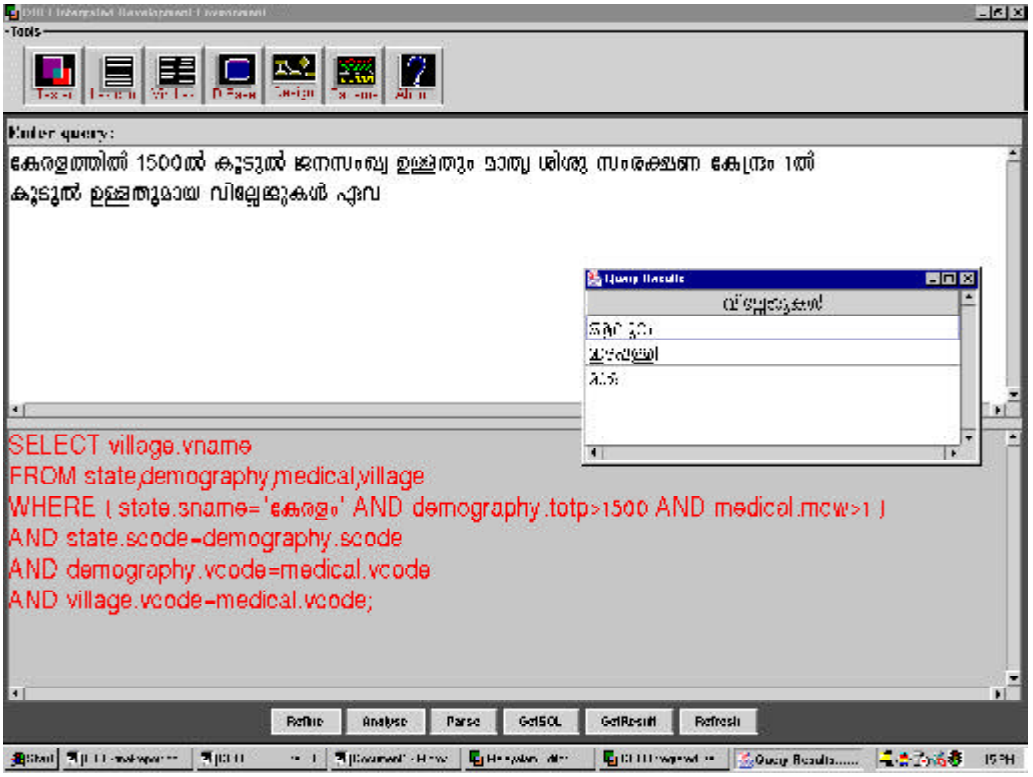
The Hindi and Malayalam queries are given below.

[illegible]

കേരളത്തിൽ 1500ൽ കൂടുൽ ന ംഖ്യ ഉള്ളതും മാതൃ ശിശു ംരക്ഷണ കേന്ദ്രം 1 ൽ

കൂടുൽ ഉള്ളതുമാമ വില്ലേ ുകൾ ഏവ





4. Conclusion & Future work

The advantage of this system is that the user can query the data store in his/ her own native language without knowing the complexity of the database structure and location of the database. The agents involved in the system are adaptive to the user language and the usage style. It is the facilitator agent who is planning and coordinating the sequence of tasks involved in query processing. The Open Agent Architecture is useful for building complex systems in which there are many heterogeneous components and in which flexibility and extensibility is important. The user interface can be made more friendly by adding agents capable of processing multimodal inputs like speech and gesture. Now the output is only in text form. It can be extended to include spatial information by integrating this system with spatial database. It adds more value to the result and can be effectively used by government bodies for resource planning. The common language framework used, makes this system adaptable to other regional languages of India .

5. Acknowledgement

This research work was sponsored by Indian Space Research Organization. We thank I.C Matieda of Space Applications Center, Ahmedabad for providing many analyses and suggestions that helped an efficient implementation.

6. References

1. Gerhard Weiss, *Multiagent Approach to Distributed Artificial Intelligence*, MIT Press 1999.
2. Bahrathi A, Chitanya V and Sangal R, *Natural Language Processing – A Paninian Perspective,* Printice-Hall of India, 1996
3. Michael N Hunhs, Munindar P Singh , *Readings in Agents*, Morgan Kaufmann 1997
4. Jeffrey M Bradshaw, Software Agents, AAAI Press, 1997
5. Nicholas R Jennings, Michael J Woodridge*, Agent Technology - Foundation, Application and Markets*, Springer , 1999
6. M. Sugumaran and P. Narayanasamy, *"An Intelligent Multiagent Meeting Schedule"r* in the proceedings of the International Conference on Artificial Intelligence in Engineering and Technology, Malaysia, 2002

About Authors



Dr. Sumam Mary Idicula is Reader in Computer Science, Cochin University of Science & Technology. She holds Ph.D in Computer Science from Cochin University of Science & Technology. She has contributed 18 Technical papers in the proceedings of International Conferences & Journals. Her area of Interest are Multilingual Computing, Intelligent Agent-based Systems, Software Engineering.
E-mail : sumam@cusat.ac.in



Sh. David Peter S is Reader in Computer Science Cochin University of Science & Technology. He holds M.Tech in Computer Science & Engineering from IIT, Chennai. He has to his credit 17 Technical papers in the proceedings of International Conferences & Journals. His fields of Interest is Intelligent Agent-based Systems, Software Engineering, Neural Networks.
E-mail : davidpeter@cusat.ac.in

A New Architecture for Brailee Transcription from Optically Recognized Indian Languages

Omar Khan Durrani K C Shet

Abstract

In order to bridge the digital divide between the sightless and sighted and to encourage literacy in them, we have designed an architecture for transcribing Braille from optically recognized Indian language. The system will help to convert masses of information in different Indian languages into a tactile reading form. The system mainly consists of OCR modules designed in an efficient manner to promote portability and scalability. In first section, we have introduced the importance and necessity of the work with successive sections clarifying briefly the properties of Braille and Indian scripts. We have also described the OCR work done with respect to Indian languages and the related work to our system. Finally, the System architecture is explained clearly followed by some conclusion and future work. The paper also identifies the needs to be fulfilled to percolate the benefit of the technology developed to the masses.

Keywords : Visually disabled, Digital divide, Braille Script, Indian Languages, Optical Character Recognition, Braille translation.

0. Introduction

Braille is a system of tactile reading and writing for visually disabled people. Each character or cell is made up of 6 (2*3) dot positions, 64 possible characters are available by using any one or a combination of dots. This system exists even today, 150 years after Louis Braille worked out its basics.

The sightless community in India and the developing world faces a tremendous hindrance in getting access to printed reading material which are scarce and communicating with the sighted community in writing, due to the difference in the script systems. Consequently, they are impaired in their educational opportunities as well as in the mainstream employment opportunities. The lack of readily available Braille material had restricted the literacy level to just three per cent of the visually impaired population. There was a pressing need for schools to generate Braille material through the computer, indigenously, to keep costs down, and in the various Indian languages to help the users. This also helps the visually impaired in teaching, learning, reading, writing and printing.

With technology making gigantic leaps and bounds, it is necessary for us to progress with this perception and try to fulfill the needs of forgotten sections of society, hence bridging the digital divide between various segments of the people. Keeping this as objective we have an system Architecture which takes input, the image file image preprocesses it and then by using the respective Optical Character Recognition (OCR) software to recognize the various composite characters and convert them into digitized text, which is then edited by a multilingual editor which includes facilities for Braille editing, Braille translation and multilingual screen reader module. The resultant Braille script can then be allowed for embossing with a suitable Braille printer.

Section 2 explains properties of Indian scripts, which are to be to taken care by the OCR engine for each Indian language. Section 3 illustrates the Braille script needed for Braille translation and Braille printing. Section 4 gives details about the work related to our system proposal. Section 5 describes the system architecture and Section 6 brings out the conclusion.

1. Properties of Indian Scripts

In India, there are eighteen official (Indian constitution accepted) languages, namely Assamese, Bangla, English, Gujarati, Hindi, Konkani, Kannada, Kashmiri, Malayalam, Marathi, Nepali, Oriya, Panjabi, Rajasthani, Sanskrit, Tamil, Telugu and Urdu... Twelve different scripts are used for writing these official languages. Examples of these scripts are shown in Fig. 1. Most Indian scripts originated from ancient Brahmi through various transformations [13]. Two or more of these languages may be written in one script. For example, Devnagari is used to write Hindi, Marathi, and Rajasthani, Sanskrit and Nepali languages, while Bangla script is used to write Assamese and Bangla (Bengali) languages. Apart from vowel and consonant characters, called basic characters, there are compound characters in most Indian script alphabet systems (except Tamil and Gurumukhi scripts), which are formed by combining two or more basic characters. The shape of a compound character is usually more complex than the constituent basic characters. In some languages, a vowel following a consonant may take a modified shape, depending on whether the vowel is placed to the left, right, top or bottom of the consonant. They are called modified characters. In general, there are about 300 character shapes in an Indian script [13].

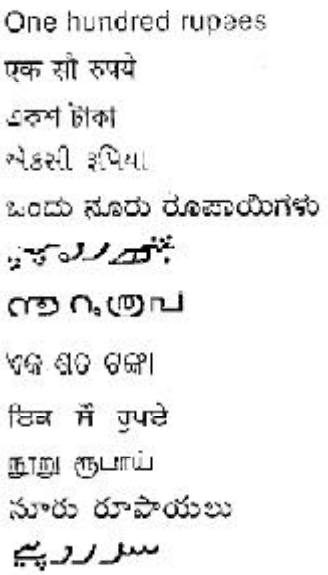


Fig. 1. Examples of 12 Indian scripts:

In some Indian script alphabets (like Devnagari, Bangla and Gurumukhi, etc.), it is noted that many characters have a horizontal line at the upper part. In Bangla, this line is called matra while in Devnagari it is called sirorekha. However, in this paper, we shall call it as head-line (see Fig. 2). When two or more characters sit side by side to form a word in the language, the headline portions touch one another and generate a big headline. Because of these, character segmentation from word for OCR is necessary. In some scripts, however, (like Gujarati, Oriya, etc.) the characters do not have headline.

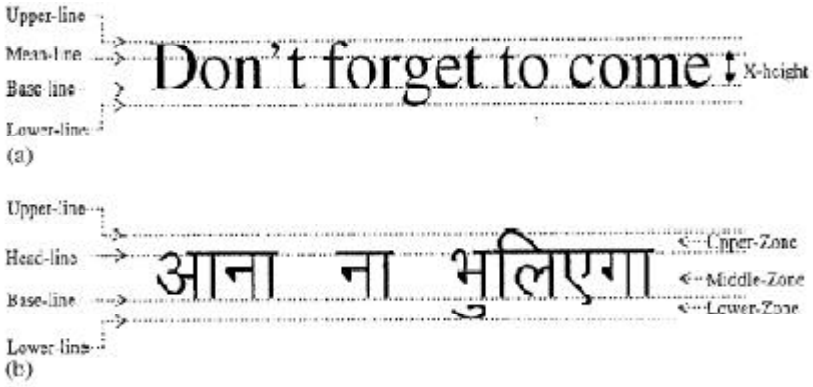


Fig. 2. Different zones of (a) English and (b) Devnagari text lines.

In most of the Indian languages, a text line may be partitioned into portion above the head-line, the middle-zone covers the portion of basic (and compound) characters below head-line and the lower-zone is the portion below base-line. Those text where script lines do not have headline, the mean-line separates upper- and middle-zone, while the base-line separates middle and lower-zone. An imaginary line, where most of the uppermost (lowermost) points of characters of a text line lie, is referred as mean-line (base-line). Examples of zoning are shown in Fig. 2. In this case, the head or mean-line along with base-line partition the text line into three zones.

2. Bharti Brailee

Bharti Braille is the standard prescribed in India for preparing Braille documents in all the Indian languages. The standard uses the six-dot system as in normal Braille but the cell assignments are corresponding to the aksharas of the Indian languages. Very simply, Braille is used as another script to text in all the Indian languages. This is possible on account of the phonetic nature of the languages where the writing system follow rules for displaying syllables rather than the basic vowels and consonants. The six dots Braille standard conforms to the following arrangement of dots, in three rows and two columns. The dots are numbered as indicated, below

1 o o 4

2 o o 5

3 o o 6

The six-dot system provides for displaying 64 different patterns. Of these, only 63 may be used for representing the aksharas. The 64th pattern is a cell without dots and is implied to represent the space character. In English Braille, the 63 different cells represent the letters of alphabet (26), ten punctuation marks, fourteen frequently used short letters and the rest assigned special meanings.

It may be noted that the assignment of meanings to each cell has no direct relationship to the set of displayable ASCII characters (96 in use). The meaning of a cell is to be interpreted in the context in which the cell is present, such as the cell preceding it, whether it appears in the beginning of a word, etc..

A sheet of printed Braille will have a series of cells embossed on thick paper and passing one's forefinger over each line of embossed cells can sense the embossing. A standard sheet of Braille has about forty cells per line and may contain 20 or more lines.

Bharti Braille is thus a system for writing syllables using a basic set of 63 shapes, each corresponding to a cell. Here, the most basic approach to writing syllables using generic consonants has been used. A syllable in Indian languages can take any one of the following forms.

(i) A pure vowel V. (ii) A pure consonant and a vowel CV. (iii) A conjunct with two or more consonants and a vowel CC..V.

Besides the syllables, special symbols are also used. These include modern punctuation marks as well.

In Bharti Braille, the basic vowels and consonants of the languages have been assigned individuals cells. Across the language of the country, between 13 and 18 vowels are in use and the consonants are between 33 and 37 in number. Thus more than 50 cells have been assigned for the basic vowels and consonants leaving the rest for special marks.

The cell assignment for a consonant assumes that the consonant has an implied vowel "a" as part of it. A pure consonant (also known as generic consonant) has no vowel and so to distinguish a basic consonant from its generic equivalent a special symbol is used in the writing systems. This is known as the halanth and its shape is a language specific ligature added to the shape of the basic consonant. Bharti Braille has set a part one cell for this purpose and this cell placed before the cell for a basic consonant turns it into generic consonant.

The idea here is that one can use this principle to write syllables in the CC..V form simply by concatenating the cells for each generic consonant.

The cell assignment corresponding to the basic vowels and consonants are similar to the assignments of the English alphabet where the sounds match. But only about 25 can be matched this way. Cells in standard Braille, which correspond to specific two letter contractions, have been chosen to take care of the aksharas such as the diphthongs and the aspirated consonants. In assigning the cells, a superset of the aksharas from all the Indian languages has been taken into consideration

Here (figure 3) are the assignments . It is also observed that some of the aksharas have been assigned two cells. The first of the two cells will invariably be a cell with just one dot, typically dot 5.

The understanding here is that the following cell has to be interpreted differently. Such schemes where a special symbol is employed to provide specific interpretation of the following character are common with computer systems and the special character is known as the escape character.

Bharti Braille confirms to the syllabic writing system followed for all the Indian languages and syllables are just written using the cells assigned for the consonants and vowels.

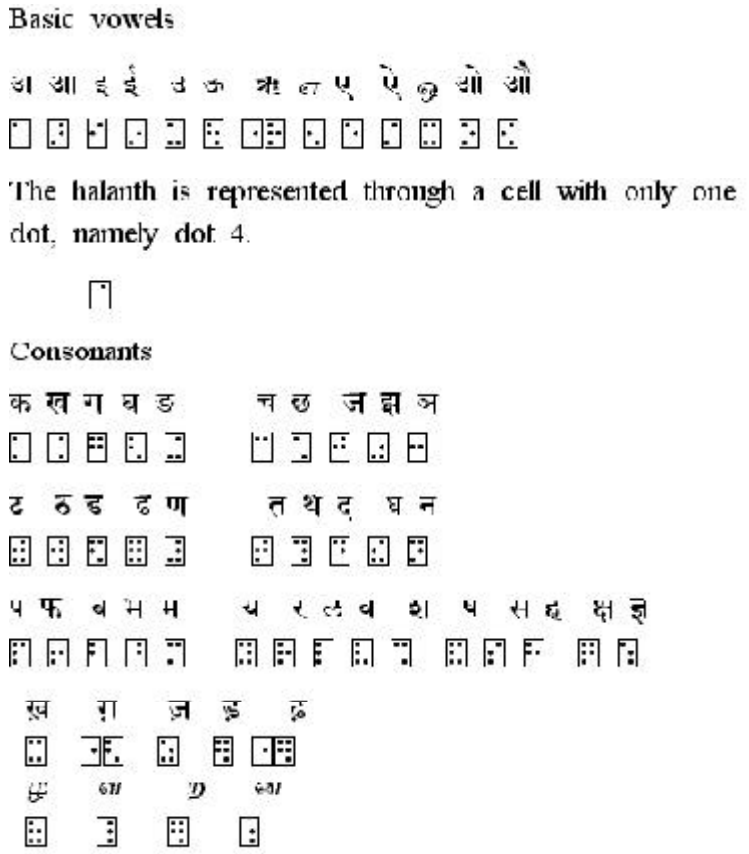
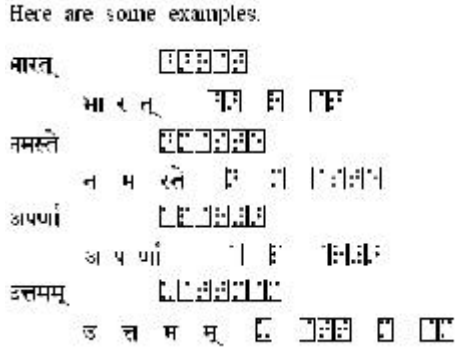


Figure 3

A pure vowel is always shown using the cell assigned. A basic consonant is always shown using the cell assigned for the consonant. A consonant vowel combination is shown using the respective cells. Normally a pure vowel will not follow a consonant and will appear only at the beginning of a word. However , there are many exception to this rule as explained in [4].



3. A view of OCR work on Indian Languages

In fact, there is not sufficient number of studies on Indian language character recognition. Most of the pieces of existing work are concerned about Devnagari and Bangla script characters, the two most popular languages in India. Some studies are reported on the recognition of other languages like Tamil, Telugu, Oriya, Kannada, Panjabi, Gujrathi, etc. [13]. Structural and topological features based tree classifier, and neural network classifiers are mainly used for the recognition of Indian scripts.

From the previous work done and experiences on various Indian script it is observed from [3,5,6,9,10,111,14,15] that the methods for preprocessing steps as well as method adopted for text and word separation can be common for all languages. Examples are Histogram based thresholding for binarization, Logical-smoothing approach for Noise cleaning, Hough transforms for skew detection, horizontal and vertical projection profiles for text and word separation are commonly used for Indian script. Structure and template based feature extraction and tree classification for Brahmi script (Devnagiri, Gurumukhi, Bangla etc.) are found to be better. For Dravidian script (Tamil, Telugu, Kannada and Malayalam) it varies due to some diverse nature among the scripts. Feature extraction using zoning and Simple Vector Machine (SVM), Nearest Neighbor (NN) classifications are found to be better (refer table 1).

4. Related Work

Galileo Reading System developed by the Robotron group has a scanner at the top and keypad at the front base. It is attached with speaker and floppy drive .It has an in-built hard disc, a serial port to get connected to the computer, printer etc. It is a multi-lingual machine. The machine operates by receiving commands through the keypad giving response through the speaker. The machine scans and stores the document as image in its buffer and recognizes it .The resultant image /text can then be either sent to hard disc, floppy disc, Embosser or to the computer. After recognition, it reads out the document till the end. Images or text can be copied from the floppy/hard disc/computer to the buffer for recognition and reading or vice versa. The recognized text can be sent to the translator after which we can have a Braille output through the embosser. As the Indian counter part, we have a system Drishti developed by LERC (Language Engineering Research Center) for Telugu and is currently being extended for other Indian scripts. Galileo is restricted to roman script; lacks built-in Braille translation from ASCII code and prices are exorbitant in the Indian context. Drishti has no built-in multi-lingual editing and text-to-speech facility as well as Braille translation.

5. System Architecture

The System has three main modules, the Preprocessing module, OCR engine and the Multi-lingual editor with Braille translation (See Figure 4b).

5.1 Preprocessing module

Along with Preprocessing we have added text and word separation in this module making it common for all the Indian languages. The methods for each step are selected as mentioned in the example of section 3. The input image under goes Binarisation, Noise cleaning, Skew detection and Skew correction, and gets segmented into lines and words. Binarisation is the process of converting the gray level image into a binary image with foreground as white and background as black. The skew may be caused while placing the paper on the scanner, or may be inherently present in the paper. Even with lot of care, some amount of skew is inevitable. After finding the skew angle; we need to correct the skew. Text and word separation involves breaking the text in the page to lines and words, which are required to identify the script.

5.2 Language Detection

The segmented text and words are then used by the language detector, which matches the language specific, shape characteristics shown below for Indian languages [2].

Bangla, Assamese - Triangular shapes with headline

Devanagari, Punjabi - Vertical Lines with headline

Gujrati - Vertical line with right bent at bottom but without headline

Oriya - Vertical line and inverted curves

Kannada - Horizontal and double bowl curves but no headline

Malayalam - Circular and no headline

Tamil - Multiple enclosed areas with some exceptional

Linear shape characters but no headline

Telugu - Circular with a tick mark sigh on top

This information (stored in the data base) is used to detect language regions, in the text region. With the help of all the information collated above, we can run different O.C.R algorithm (engine) for different languages.

5.3 The OCR engine

The OCR engine consists of four steps; character segmentation, feature extraction, classification and post processing. These steps differ for each Indian script. Therefore we have a separate OCR Engine for each Indian script (see figure 4 a).

Under character segmentation, individual words are processed to obtain the components for the scripts. For Devnagiri, Gurumukhi, Bangla etc segmentation involves the removal of sirorekha (the horizontal bar). This separates the constituent components from a word. For Telugu and Kannada, component extraction implies the separation of connected components. The individual connected components here are not distinct letters. They can also be modifiers. This results in splitting of the single word into many separated components. Once the sirorekha is removed, for Devnagiri like scripts, the top, middle and bottom zones are identified easily. Top zone gets automatically separated with the removal of sirorekha. Bottom zone is identified from the projections. Components in top and bottom zones for Hindi are part of vowel modifiers. Each of these components is then scaled to a standard size before feature extraction and classification.

Feature extraction captures the essential characteristics of the symbol to help in classification. It finds the amount of data by extracting relevant information, usually results in a vector of scalar values. Features are also normalized for distance measurements. Classification compares the feature vectors to the various models in the database and finds the closest match. The output after classification is transformed into a format like ASCII set or ISCII (Indian Script Code for Information Interchange).

5.4 Multi-lingual editor

The Multilingual-editor takes the input the ASCII or ISCII code and provides proof reading. It will also have utilities like screen reading and Braille translation. A provision for giving proof reading in Braille Script can also be provided. IIT, Chennai has also developed such software, which can be utilized in this module

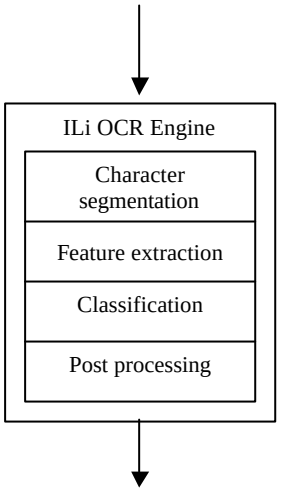


Figure 4 a : Expansion of OCR engine module.

Figure 4 b: System Architecture

6. Conclusions and Future Work

Using the methodology discussed above we can get better accuracy and output. This method has advantages like High efficiency, greater speed, less utilization of processing resources and less human intervention. With this system, Government can promote even higher education for the visually disabled. The work to bring this into a practical form has been started by us. We have already implemented the multilingual editor and extension to Braille translation is in progress hoping to finish in few months. Initially the work can be done for one or two fonts and later extended for multi font. Indian script OCR error correction module that corrects single character can be extended to post recognition error correction (spell check and morphological techniques with grapheme features of language script), which will improve OCR accuracy. A work in [5] has shown better results for Telugu and Hindi with SVM-PCA based OCR. Hence a study to adopt SVM-PCA for all Indian languages can be made keeping efficiency, accuracy and portability as objectives. Other aspects like Phonetic nature of the language and its reflection on the script, similarities and dissimilarities between script based on their origin and evolution, geometrical shape features for characterization of characters can also be addressed.

6.1 Acknowledgments

The authors wish to thank Prof. Anupam Basu of IIT, Kharagpur, Prof. Murali, PES, Mandya and Central Institute Indian Language, Mysore for suggestions and contributions.

Table 1: Different OCR systems on printed Indian script

Script	**System**	**Feature**	**Classification Technique**	**Accuracy claimed**
Devanagari	Pal and Chaudhuri [13]	Structural and template features	Tree classifier	96.5%
	Garain and Chaudhuri [13]	Run length-based template feature	Tree classifier	97.5%
Bangla	Chaudhuri and Pal [13]	Structural and template features	Tree classifier	96.8 %
Gurumukhi	Lehal and Singh [13]	Structural and topological features	Tree classifier	97.3%
Oriya	Chaudhuri et al.[13]	Structural and template features	Tree classifier	96.3%
Telugu	C.V Lakshmi and C Patvardhan [6]	Gradient based features	Symbol association information during segmentations	98%
Tamil	A G Ramakrishnan & Kaushik Mahata [10]	Geometric based moments	Based on spatial occupancy and Nearest neighbor	97%
Kannada	Ashwin and Sastry [11]	Zoning features	SVM classifier	97%
Bi-lingual Hindi-Telugu	C. V. Jawahar et al. [5]	Principle component transformation	SVM classifier	96%

7. References

1. Anupam Basu "Bharati Braille Information System: An Affordable Multilingual System for Empowering the Sightless Population", development by design, Bangalore, Dec 1-2, 2002.
2. Aditya Gokhale "Bi-lingual optical character recognition system For devanagari and english, Centre For Development Of Advanced Computing, GIST R&D Pune, Maharashtra, INDIA.
3. B B Chaudhuri, U Pal And M Mitra Automatic recognition of printed Oriya script Sadhana Vol. 27, Part 1, February 2002, pp. 23–34
4. Bharati Braille, a tutorial, http://acharya.iitm.ac.in/disabilities/index.html
5. C. V. Jawahar, M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran A Bilingual OCR for Hindi-Telugu Documents and its Applications Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
6. C. Vasantha Lakshmi and C. Patvardhan An optical character recognition system for printed Telugu text Pattern Anal Applic (2004) 7: 190–204.
7. Durre, K.P. (1990). "BrailleButler: A new approach to non-visual computer applications" Proceedings Third Annual IEEE Symposium on Computer-Based Medical Systems, University of North Carolina at Chapel Hill 1990. Los Alamitos, CA: IEEE Computer Society Press.
8. Directory of Aids and Appliances for the Visually Handicapped. Government of India, Rep. Ministry of Welfare, 1991.
9. G.S. Lehal, C. Singh, Feature extraction and classification for OCR of Gurmukhi script, Vivek 12 (1999) 2–12.
10. Kaushik Mahata and A. G. Ramakrishna, "A complete OCR for printed Tamil text ", Proc Tamil Internet 2001, Singapore, July 22-24, 2000, pp. 165-170.
11. T.V. Ashwin, P.S. Sastry, A font and size independent OCR system for printed Kannada documents using support vector machines, Sadhana 27 (2002) 35–58. [12] U. Garain, B.B. Chaudhuri, Segmentation of touching characters in printed Devnagari and Bangla scripts using fuzzy multifactorial analysis, IEEE Trans. Systems Man Cyber. Part C-32 (2002) 449–459.
12. U. Pal, B.B. Chaudhuri Indian script character recognition: a survey Pattern Recognition 37 (2004) 1887– 1899.
13. U. Pal and B.B. Chaudhuri On the development of an optical character recognition (OCR) system for printed Bangla script, Ph.D. Thesis, 1997.
14. U. Pal and BB choudhury, Printed Devnagari script OCR system, Vivek 10 (1997) 12–24.
15. Galileo reading system, http://www.sensorytools.com/ galileo.html.

About Authors



Sh. Omar Khan Durrani, Second year MTech student, Deptt. Of Computer Engineering, NITK., Surathkal, Karnataka.
E-mail : pcs03879@nitk.ac.in



Dr. K C Shet is a Professor in Department of Computer Engineering, NITK, Surathkal, Karnataka.
E-mail : kcshet@nitk.ac.in

A Document Reconstruction System for Transferring Bengali Paper Documents into Rich Text Format

Anirban Ray Chaudhuri Debnath Singh Mita Nasipuri Dipak Kumar Basu

Abstract

The transformation of a scanned paper document into an editable form suitable for further processing such as desktop publishing or archiving in a digital library is a complex process. It requires solutions to several problems – document analysis by acquiring knowledge of document layout by a Page Layout Analyzer (PLA), followed by document recognition, which mainly comprises text recognition by Optical Character Recognition (OCR). Besides these two, another important problem is document reconstruction by transforming content into an electronically editable format by keeping the original layout intact. Core OCR modules exist on different Indian scripts, but no such document reconstruction system is available for Indian scripts. The document reconstruction system reported in this paper is the first of its kind on Indian scripts and it addresses document reconstruction for Bengali document images. The system makes use of the knowledge of both document layout extracted by a PLA in a graphical user interface (GUI) and the results of text recognition steps performed by OCR for transformation of paper documents into Rich Text Format.

Keywords : Indian Scripts, Desktop Publishing, Page Layout Analysis, Optical Character Recognition, Document Reconstruction, Encoding Standard, Indian Language.

0. Introduction

For the digital archiving, the most rudimentary way to make scanned documents accessible is to insert the document images in an MSWord document or as an attachment to HTML pages, after having converted into supportable digital format (e.g., GIF or JPEG). In this way, the information present in the input document could be preserved. However, this approach presents several disadvantages.

1. Compressed raster images are still quite large and their transfer can be unacceptably slow.
2. The content can only be viewed but is not editable.
3. Information retrieval based on 'keywords query and searching' is not possible.
4. For desktop publishing, to access the content in editable form, manual entry of the textual content (e.g., title, abstract and even the whole document) is required.
5. For multi page documents, pages can be presented only in a sequential order, thus missing the advantages of a hypertext structure necessary for document browsing in a digital library.

In view of the recent advances in information and communication technologies viz., in the frontiers of information retrieval, fast and noiseless data transfer and document archiving, there is an urgent need for tools that are able to transform data presented on paper into an editable form. This demand has been considerably met for European scripts and some of the Asian scripts [1,2]. Now there is an increasing demand for the same for Indian scripts. Looking at the growing needs of the same for the last few years, government agencies like Department of Science & Technology and Ministry of Information & Technology

sponsored several turnkey projects at major R & D organizations and maximum emphasis has been given to Devanagari and Bengali, the 3rd and 4th most popular scripts of the world respectively. Recently commercial Optical Character Recognition (OCR) systems on Devnagari Script have also come up in the market. OCR systems also exist on other major North Indian scripts such as Gurumukhi, Marathi, Assamese and Oriya [3,4]. These OCR systems are still in a nascent form; particularly, document reconstruction part receives the minimum attention due to non-availability of appropriate tools such as fonts with standardized Font driver and editor support. These systems may be considered as core OCR modules as they can handle document images containing no other entities except text in a single column. These modules can save scanned documents in plain text format where no style sheet is associated. As a result, their appearance is not similar to the original document. However, for the sake of original document layout preservation, manual modification of Font-size and paragraph attributes like indentation, justification, line spacing, insertion of graphical components and regeneration of text in multicolumn are very tedious and time consuming.

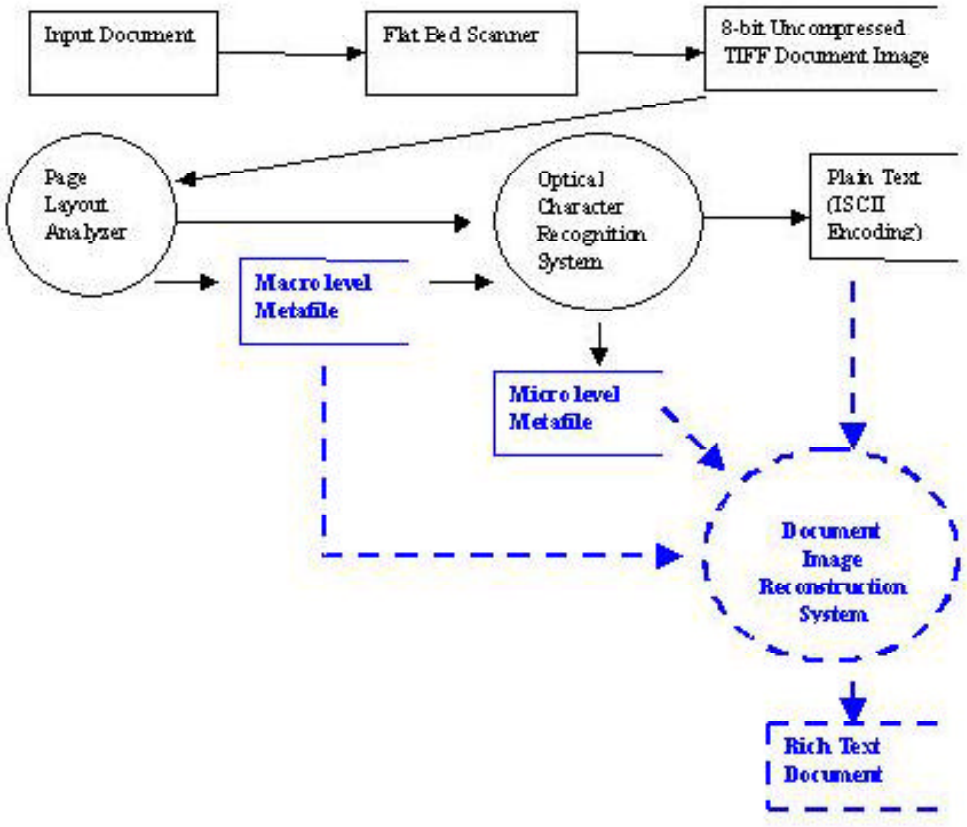


Figure 1: The automatic generation of the structured rich text documents from document images.

It may be noted that most of the North Indian scripts such as Devnagari, Marathi, Bengali, Gurumukhi and Assamese are based on Brahmi based scripts and have very similar characteristics. As a result, extension of one OCR system in any of these scripts could be easily adopted by the OCR systems of the other scripts.

In this paper, we report the status of the ongoing project, “Anulikhan” on Brahmi based script in the context of Bengali document reconstruction. The objective of the project is to upgrade the core OCR into an advanced document image processing system where original layout of the input document image, including document formatting could be preserved in an editable electronic version. The reported system on document reconstruction transforms the document images into Rich Text Format (RTF), the most acceptable format that can represent document of arbitrary complexity. The system preserves the document structure in the final output as a RTF file by aggregating textual, graphical, and document formatting layout that is extracted by the PLA and the core OCR.

The paper is organized as follows. Section 2 presents an overview of the complete document image processing system. In Section 3, implementation details of document reconstruction system are presented. Performance evaluation of the proposed system is provided in Section 4. Section 5 concludes the paper and indicates the directions of the future work.

1. The Complete System Overview

The Figure 1 illustrates the overall process of the automatic generation of the structured rich text documents from document images. Prior to document reconstruction, the Page Layout Analyzer (PLA) performs layout structure analysis on the input document image and classifies document entities into texts, images and rulers/separators. It first coverts the image from gray tone to two tone and corrects the skew, if required. For gray tone to two-tone conversion, we use a simple histogram based thresholding approach [5,6]. A small amount of skew in the document image can be automatically corrected with projection profile and Hough transformation techniques [7]. To handle arbitrary skew in multiple directions an intelligent Graphical User Interface (GUI), based on mouse controls, is designed. From the deskewed binary image, PLA localizes and extracts text regions as rectangular blocks containing moderate sized connected components satisfying homogeneity or compatibility in terms of headline, vertical line and other (textual) features particularly available in Brahmi based scripts [8]. Note that the headline is a horizontal line present in the upper part of the characters, establishing word-level connectivity; whereas, the vertical line lies just below the headline and spreading over the middle zone. These two features occur in most of the characters. Each homogenous textual block is then fed into the core OCR system. The core OCR system performs the following tasks: line-word-character segmentation, character recognition, followed by word formation. For line, word as well as character segmentation, a projection profile based technique is adopted, as shown in Figure 3 [9]. The character recognition module primarily uses a run-length encoding based template and closest target matching for character level recognition as described in [10]. The block-level output of the core OCR is coded in ISCII and stored in a plain text file. The plain text file is displayed in a plain text editor, embedded in the system and is used for online correction of the OCR output. To carryout the document reconstruction, two metafiles are primarily generated. One metafile is designed to store the information evolving out of the PLA module. Since it contains information at image/block/region level, this metafile may be designated as a macro level metafile. Each block generated by the PLA is described by the positional information of the two diagonally opposite corners. Also, based on the content type (textual or graphical information), each block is labeled as text, picture (halftone images or line drawings), horizontal line or vertical line. To store the character level and paragraph formatting information that are evolving out from OCR output, a metafile that might be designated as micro level metafile is designed. The ISCII encoded plain text file together with the two metafiles, described above, are fed into the document reconstruction system. Within the document reconstruction system, these two metafiles are integrated into a single metafile in a more compact and structured format. By using this metafile and the plain text file, the reconstruction system generates the final output in RTF. Figure 2 illustrates an overall process description of the system for document reconstruction.

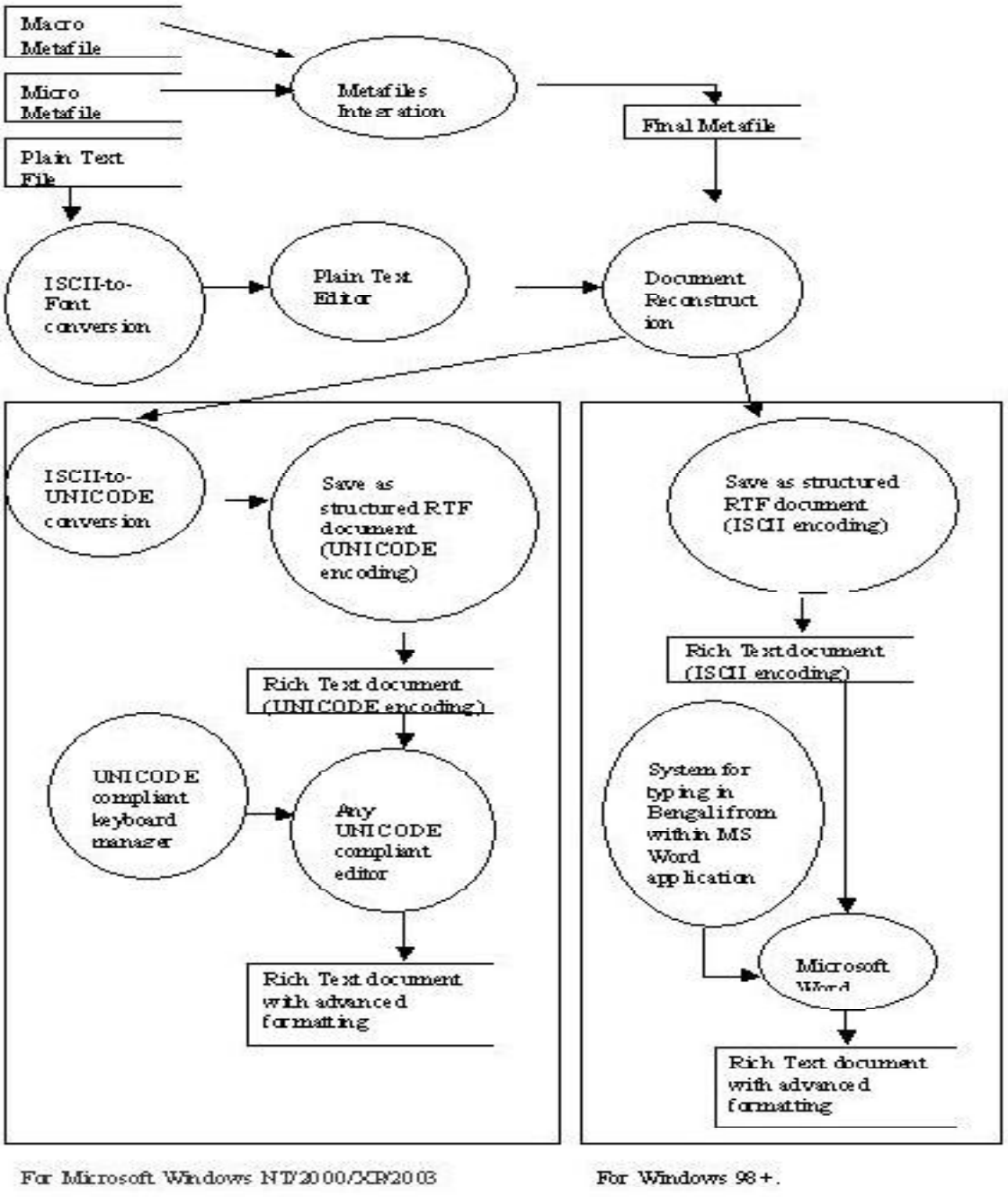


Figure 2: An overall system for document reconstruction.



Figure 3 : Profile based method for line-word to character segmentation where matra and vertical bar features are extensively used.

2. Document Reconstruction System

Since the document reconstruction system has much functionality to perform, the entire system is designed in several stages. It reduces the complexity of the overall problem, thereby achieving an effective modular solution.

The stages are :

1. Font generation and Font driver:
   (a) Designing Indian Standard Code for Information Interchange (ISCII) compliant Bengali True Type Font (TTF).
   (b) Designing UNICODE compliant Bengali Open Type Font (OTF).
   (c) Designing keyboard layout and font driver for writing with the True Type fonts.
2. Font Conversion:
   (a) ISCII file to TTF file and vice versa.
   (b) ISCII file to UNICODE file.
3. Online correction of OCR.
4. Integration of metafiles.
5. Rich text generation.
6. Proofreading of the final output:
   (a) Designing a module for typing in Bengali from within Microsoft Word for correction of the final output (designed for Win98+).
   (b) Designing an UNICODE compliant keyboard managing system for typing Bengali.

2.1 Font Design and Font Driver Generation

(a) Designing ISCII compliant Bengali TTF

A TTF is designed for writing in Bengali. This is required due to non-availability of Bengali fonts having glyphs that support all conjuncts with standardized Font driver. A glyph set is designed for all Bengali characters. Note that in 8-bit character encoding scheme a maximum of 256 combinations are available out of which 32 are used by the system itself (e.g., tab key, alt key and enter key.). Remaining 224 combinations are not enough to accommodate all Bengali characters. Each basic character is represented by a single glyph and most of the conjuncts are displayed programmatically by concatenating two or more glyphs. Furthermore, the font is made ISCII compliant by placing the glyphs of the basic characters in appropriate or designated code point specified by ISCII.

After determining the basic Bengali character set, the dimension of the font i.e., the font attributes such as ascent, descent, leading, font name and weight are determined. For generating TTF, we use a standard font designing software, Fontographer 4^{TM}.

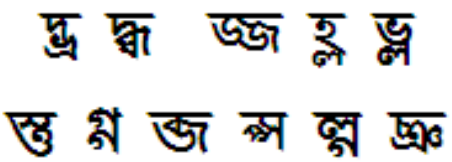


Figure 4: Some of the glyphs of conjuncts in the newly designed font that rarely occur in other popular fonts.

(b) Designing UNICODE compliant Bengali OTF

OpenType fonts are also referred to as TrueType Open version 2.0 fonts, because they use the TrueType 'sfnt' font file format [11]. We design an OpenType font using Microsoft Visual OpenType Layout Tool (VOLT) that provides an easy-to-use graphical user interface to add OpenType layout tables to fonts with TrueType outlines. It supports a wide range of substitution and positioning types. It also contains a proofing tool that helps correcting the result of applying layout table lookups. We use the tool to add OpenType layout tables to our designed Bengali True Type fonts. Prior to adding the tables, the font is made UNICODE compliant by placing the glyphs of the characters in appropriate or designated code point specified by UNICODE.

(c) Keyboard Layout and Font Driver Design for writing with the designed TTF



Figure 5. Typographical structure of a word.

Most of the keyboards available in India are designed for writing of English characters. For languages other then English, a soft keyboard layout design is necessary to determine which character of that language is assigned to which key. The keyboard layout is designed for our fonts and this becomes the

foundation for writing the font driver program. Note that this font driver manages the mapping of the keyboard key to Bengali characters by assigning each Bengali character to a keyboard key. One key can be used to display two different Bengali characters with Shift key pressed and released. When a user enters a key, the key is trapped and appropriate Bengali character is displayed. Link-logic is for the identification of conjunct characters. A sequence of three or more keystrokes is necessary for display of a conjunct character. Conjunct characters are stored in a separate file and on detecting the link key at runtime, appropriate conjunct character is searched from the file and displayed.

2.2 Font conversion

In the absence of any standardization, each Bengali font has its own keyboard layout. Texts are being stored in font dependent glyph codes and the glyph-coding scheme for these fonts is not same. As a result, one cannot exchange the electronic Bengali documents from one font to another as conveniently as in English. To alleviate this problem, font conversion utilities are developed. One of the utility converts text encoded in ISCII to our system supported font and vice-versa; while the other one coverts text data encoded in ISCII to UNICODE and vice-versa.

(a) ISCII-to-Font and vice-versa

ISCII is a standard encoding scheme for Indian languages and a convenient way of exchanging information. This module performs conversion operation from ISCII to system supported font and vice-versa. The converter program accepts an ISCII file as input and for each character, it replaces the ISCII character code with the specific font character code. Next, the data is saved in a file. For font to ISCII conversion, the logic is same as above but in the reverse order.

(b) ISCII-to-UNICODE and vice-versa

The obvious disadvantage of coding an ISCII encoded data to a particular font is that the font must have its own font driver. In worst case, for N supported fonts we require N Font drivers. A more standard solution to this problem is to encode data in UNICODE. Once data is encoded in UNICODE, we can smoothly switch it between UNICODE compliant Bengali Open Type fonts, without worrying for Font driver. Therefore, a converter module has been designed to convert data between ISCII and UNICODE.

2.3 Online correction of OCR

The OCR output (the plain text file) may contain some word errors due to incorrect character and/or word recognition. A word error can belong to one of the two distinct categories viz., non-word error and real word error. A non-word error makes a word meaningless, while a real word error means an error that results a valid word but not the intended one in the sentence., thus making the sentence syntactically or semantically incorrect. Though some attempt is found on non-word error detection [12], no work is available on real word-error detection for Indian scripts.

The phrase “online correction of the OCR” implies prior to document reconstruction, support of an editor is required that enables user to correct the non-word and real word errors present in the OCR output. Using the standard Windows controls, such as buttons, menus, scroll bars, and lists we design the Editor for the above purpose. As we provide support for various text editing operations including different paragraph and character formatting, find/replace string etc., this editor is also used as a standalone ISCII compatible plain text editor. To help the new user to write in Bengali efficiently and quickly an on-screen keyboard is also designed.

2.4 Integration of metafiles

A module is designed to synthesis the macro and micro level metafiles. This module extracts necessary information for document reconstruction from the metafiles, and stores into a single metafile in a more integrated and structured format. The metafile so formed could be designated as final metafile. Besides the necessary block/region level information, the final metafile also contains important character level and paragraph formatting information for text regions. Association information needed to relate a text block with its corresponding textual information contained within the plain text ISCII file is also maintained within the final metafile.

2.5 Rich text generation

A module is designed to restore the physical layout and logical structure of the original document, based on labeled regions and recognized texts that we obtain as the result of layout analysis and text recognition. Using the block/region information contained within final metafile and as per RTF specification 1.7 [13] we reconstruct pictures and rulers/separators. To reconstruct text frame/block, the primary task is to recognize the font and its size. As shown in Figure 5, text line images are composed of three typographical zones – The ascender (upper), the x height (mid) and the descender (lower) zones, which are delimited by four virtual horizontal lines. While the ascender and descender zones depend on the text content, the x height zone is always occupied regardless of the characters that occur. The x height is commonly called midzone distance, and its proportion in the text height differs from one typeface to another.

A statistical classifier is designed, that makes use of typographical attributes such as ascenders, descenders, midzone distance, matra thickness and stroke width obtained from a word image (stored in the final metafile) to recognize font-shape-size. Another crucial issue is the identification of paragraphs along with the paragraph indentation and justification. Once the font-shape-size and paragraph attributes are identified, the system reconstructs text frames from the final metafile and the plain text file. Note that at present, no algorithm is designed and implemented to find the reading order of the text regions in Indian documents. We adopt a GUI based approach to tackle this issue for any complex layout. We generate the reading sequence by placing and clicking the mouse in the text blocks in an order.

2.6 Proofreading of the final output

Once the image is converted to rich text document, there might be two problems that need to be solved.

- There might still be some errors in text recognition results even after correction in online.
- Users might need to append/modify texts as well as layout of the original document and export it to another data processing software.

To address the above issues we design:

- A module to correct the OCR results. It supports typing in Bengali using one of the supported fonts, from within Microsoft Word to correcting the text recognition results. It is designed for Win98+ and operates on ISCII encoded text.
- An UNICODE compliant keyboard managing system for typing Bengali anywhere on UNICODE compliant Windows, to correct the UNICODE encoded text recognition results.

3. Performance Analysis

To assess the absolute performance of the document reconstruction system, a large number of document images of Bengali pages having multi-column layout with pictures rulers etc., from various resources should be tested. We are currently testing with a few samples of one hundred and twenty pages. Fifty

scanned pages are directly generated from computer. Twenty pages are scanned from “Desh”, the most popular Bengali literature magazine. About fifteen pages of cutting are taken from “Anaandabazar Patrika’, which has world-wide the maximum readership as a daily Bengali newspaper; whereas, about ten pages are derived from “Bartaman”, another daily Bengali newspaper with the second largest readership. Rest is taken from popular Bengali books and other type of publications. Pages from hardcopies are scanned into 8-bit uncompressed tiff image at 300 dpi resolution. The input and output of the document reconstruction system at different stages of the algorithm are shown in Figure 6.

As far as document reconstruction system is concerned, the results are quite satisfactory. Text blocks, pictures and rulers are correctly reconstructed wherever PLA and OCR provide correct information. In a very few cases incorrect character level and paragraph formatting information result by the reconstruction module. The error evolving out of the PLA and text recognition system produces erroneous font metric estimation besides wrong font and font size recognition.

For about 3% of the data, incorrect blocks are generated by PLA mostly due to small paragraphs of two-three lines with poor text alignment. So far, we do not pay much attention to the core OCR. Only about ten percent of the sample document images are learnt for the template generation. Only our designed fonts are learned thoroughly. Consequently, the accuracy of OCR falls rapidly for documents whose character/ fonts are not well represented by the template database. Particularly, OCR results for old print documents (less than 3% of the pages) are unacceptable due to several problems including poor object background segmentation, touching characters, noise and lack of representation in the template database.

4. Conclusions and Direction of the Future Work

The presented advanced document image processing system transforms Bengali printed documents with its original layout into Rich Text Format. To the best

of our knowledge this is the first initiative on any Indian script. There are several significant benefits to this transformation: the transformed documents can be edited, reformatted, appended into other documents or converted into HTML version for the access via Internet more quickly than the original bitmap image. The user can manipulate the original document and be able to make hypertext document and automated information retrieval that is necessary for document browsing. The potential beneficiaries of our system are – newspapers (printed in Bengali script) and other publishing houses, libraries (digital library generation), offices looking for office automation (document archiving), linguistic community (for creating corpus) and blind people (as automated reading aid). As the development is based on VC++, all the code is easily portable. In addition, since the object-oriented concepts are used for development, the system can be easily expanded as per requirement. System encodes the output in 8-bit ISCII and UNICODE encoding format and hence, outputs can be viewed or edited with any third party editor that supports ISCII and/or UNICODE. However, the system is still far from its final shape. The limitations and future scopes of the present system are :

- ? Due to reduction in cost and rapid modernization, present publication media uses multiple colors for text and shading in the background. Simple thresholding techniques that are currently being used by OCR modules are not appropriate for these documents. Designing of a more robust segmentation module for object background detection that could take care of degraded documents as well is started.
- ? For better representation and easy editing, PLA module as well as the macro metafile specification should be upgraded by merging of block level information to column level so that all blocks in a single column could be merged. In addition, by its very design, PLA is not effective for scripts where characters within word are not connected as in case of English. Note that in modern

Bengali publications, embedded English text of a few lines often presents. In such a situation, a script identifier is also required with the PLA.

- More effort is required for upgrading the core OCR for better accuracy. Segmentation of touching characters should be provided with the module. In addition some word level error correction by post processing should be developed. For easy learning of new templates and efficient template database management, we are planning to provide a GUI.
- For better representation of the reconstructed document, more fonts should be designed.
- More enhanced GUI for online correction as well as block level correction would be very helpful for the end user.
- We have also started designing a similar advanced document reconstruction system for Devnagari script.

5. Acknowledgement

Science and Engineering Research Council of Department of Science and Technology, Govt. of India is duly acknowledged for the “DST Fast Track” awarded to the first author and the work presented in this paper is the outcome of that project.

6. References

1. Nagy G. (2000), Twenty years of document image analysis in PAMI, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, No.1, pp.38-62, 2000.
2. Ding X., Wen D., Peng L., Liu C. (2004), Document Digitization technology and its application for digital library in China, Proc. 1st International Workshop on document image analysis for Libraries (DIAL’04), IEEE Press.
3. Ministry of Communications & Information Technology, Government of India (2003), ViswaBharat January 03, http://tdil.mit.gov.in
4. Bansal V. and Sinha RM.K., (2001), A Devanagari OCR and A Brief Overview of OCR Research for Indian Scripts in Proceedings of STRANS01, held at IIT Kanpur.
5. Jain A. K. and Yu. B (1998), Document Representation and Its Application to Page Decomposition, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 20, No. 3, pp. 294-308.
6. O’Gorman L, and Kasturi R (1997) Document image analysis. IEEE Computer Society Press Executive Briefing Series, Los Alamitos, CA.
7. Cattoni, R., Coianiz, T., Messelodi, S., and Modena CM, (1998), ITC-IRST, Povo,Trento, Italy, Geometric Layout Analysis Technique for Document Image Understanding: A review,http://tev.itc.it/people/modena/Papers/DOCSEGstate.pdf
8. Ray Chaudhri A,. Mandal A, and Chaudhuri B. B., (2002), Page Layout Analyzer for Mulitingual Indian Documents, Proc. Language Engineering Conference, IEEE CS Press.
9. Chaudhuri B . B., Garain U. and Mitra M., Indian Statistical Institute (2003), On OCR of the most popular indian scripts: Devnagari and Bangla, TR/ISI/CVPR/03/2003.
10. Garain U. and Chaudhuri B.B., (1998), Compound Character Recognition by Run Number Based Metric Distance, Proc. SPIE Annual Symposium on Electronic Imaging, San Jose, USA, pp.90-97.
11. Microsoft (2003) Microsoft typrgraphy, http: //www.microsoft.com/typography/default.mspx
12. Chaudhuri B. B.and T. T. Pal (1998), Detection of word error position and correction using reversed word dictionary, Int. Conf. on Computational Linguistics, Speech and Document Processing, Calcutta
13. Microsoft(2003) RTF Specification 1.7, http://support.microsoft.com/kb/q86999/

About Authors

Dr. Ray Chaudhuri, Anirban born in Shatiniketan ('69), received Masters in Pure Mathematics ('91) and Ph.D. in Computer Science ('2000) from Visva-Bharati (A Central University, Shantiniketan) and Indian Statistical Institute (Kolkata) respectively. He started service as a Teacher at Patha-Bhavana, the school section of Visva-Bharati and afterwards in the capacity of a Lecturer at the same University, Research Scientist at Indian Statistical Institute and Post-Doctoral Research Associate at Coordinated Science Laboratory as well as at the Department of Electronics and Computer Engineering, University of Illinois at Urbana Champaign. In 2001, to have a more independent research and business-consultancy career, he left the permanent job. At present with a Fast Track Fellowship from Department of Science & Technology, Govt. of India, he is primarily attached with the Department of Computer Science and Engineering, Jadavpur University, Kolkata as a Principal Investigator in the Project "Anulikhon: Advanced Optical Character Recognition System for Bangla and Similar Brahmi Based Indian Scripts". His research interest includes Statistical Pattern Recognition in point pattern analysis and consistent estimation, Image Processing and Computer Vision in areas of automatic target recognition, remote sensing, volume visualization and document image analysis.

Sh. Singh Debnath, born in Kolkata (79), received a Bachelor in Science from Kolkata University 1999, and currently doing his Masters in Computer Application from DoE (B-level). He is also currently working as a Project Trainee at the Centre for Micro Processor Application, Department of Computer Science & Engineering, Jadavpur University.

Dr (Mrs.) Nasipuri Mita received her B.E.(Tel.E.), M.E. (Tel.E.) and Ph.D(Engg.) Degrees from Jadavpur University, Kolkata, India, in 1979,1981 and 1990 respectively. She is currently a Professor and Head of the Computer Science and Engg. Department of Jadavpur University. Her current research interest includes Computer Architecture, Image Processing, Pattern Recognition, Multimedia Systems, Bio-medical Signal processing etc. She had a large number of research publications in International/National Journals and International/National Conferences. She is a Senior Member of the IEEE, USA, Fellow, The Institution of Engineers (India), Fellow, West Bengal Academy of Science and Technology.

Dr. Basu Dipak Kumar received his B. (Tel.E.), M.E.(Tel.E.) and Ph.D (Engg.) degrees from Jadavpur University, Kolkata, India, in 1964,1966, 1969 respectively. He joined Electronics & Telecommunication Engg. Department of Jadavpur University as a faculty member in 1968. He is currently a Professor in the Computer Science and Engg. Department of the same University. His field of research interest includes Digital Electronics, Microprocessor Applications, Bio-medical Signal Processing, Knowledge Based Systems, Image Processing, Pattern Recognition, Multimedia Systems etc. He had a large number of research publications in International/National Journals and International/National Conferences. He is a former fellow of the AvH Foundation, Germany. He is a Fellow, the Institution of Engineers (India), Fellow, West Bengal Academy of Science and Technology and Senior Member of the IEEE, USA.

Natural Language Requirements to Executable Models of Software Components

V R Rathod S M Shah Nileshkumar K. Modi

Abstract

The UniFrame approach to component-based software development assumes that concrete components are developed from a meta-model, called the Unified Meta- component Model, according to standardized business domain models. Implicit in this development is that there is a Platform Independent Model (PIM) which is transformed into a Platform Specific Model (PSM) under the principles of Model-Driven Architecture. This paper advocates natural language as the starting point for developing the business domain models and the meta-model and shows how this natural language may be mapped through the PIM to PSM using a formal system of rules expressed in Two-Level Grammar. This allows software requirements to be progressed from business logic to implementation of components and provides sufficient automation that components may be modified at the model level, or even the natural language requirements level, as opposed to the code level.

Keywords : Natural Language Processing, Model-Driven Architecture

0. Introduction

Model-driven architecture (MDA) is an approach whereby software components are expressed using models, typically in UML. The basic approach is to define Platform Independent Models (PIMs) which express the business logic of components conforming to some domain (e.g. banking, telecommunications, etc.) and then to derive Platform Specific Models (PSMs) using a specific component technology (e.g. CORBA, J2EE, etc.). Business logic is typically expressed in natural language before a model is developed. Standardization of business domains and associated components is being undertaken by the Object Management Group (OMG). To facilitate the MDA approach to be used in practice, automated tools are needed to develop the business domain specifications from their requirements in natural language as well as to enable transformation from PIMs into PSMs. Furthermore, if MDA is to be used for constructing distributed software systems, then the models must consider not only functional aspects of business logic, but also non-functional aspects, which we call Quality-of-Service (QoS). QoS attributes are not currently considered in the MDA framework.

UniFrame is an approach for assembling heterogeneous distributed components, developed according to MDA principles, into a distributed software system with strict QoS requirements. Components are deployed on a network with an associated requirements specification, expressed as a Unified Meta-component Model (UMM) in the Two-Level Grammar (TLG) specification language. The UMM is integrated with generative domain models and generative rules for system assembly which may be automatically translated into an implementation which realizes an integration of components via generation of glue and wrapper code. Furthermore, the glue/wrapper code is instrumented to enable validation f the QoS requirements.

This paper describes a unified method of expressing business domain models in natural language, translating these into associated business logic rules for that domain, application of the business logic rules in building MDA PIMs, and maintaining these rules through development of PSMs. The complete mapping takes place using a formal system of rules expressed in Two-Level Grammar. This allows software requirements to be progressed from business logic to implementation of components and provides sufficient automation that components may be modified at the model level, or even the natural

language requirements level, as opposed to the code level. Section 2 describes our previous work with Two-Level Grammar and its use as a specification language. The application of this to Model-Driven Architecture is discussed in section 3. Finally we conclude in section 4.

1. From Natural Language Requirements to Formal Models

To achieve the conversion from requirements documents to formal models, several levels of conversions are required. First, the original requirements written in natural language are refined as a preprocessing of the actual conversion. This refinement task involves checking spellings, grammatical errors, consistent use of vocabularies, organizing the sentences into the appropriate sections, etc. The requirements are expected to be organized in a well-structured way, e.g. as laid out in or as a collection of use-cases, and be part of an ontological domain. Since we are allowing for specification of components that will be deployed in a distributed environment, Quality-of-Service attributes are also specified. Next the refined requirements document is expressed in XML format. By using XML to specify the requirements, XML attributes (meta-data) can be added to the requirements to interpret the role of each group of the sentences during the conversion. The information of the domain-specific knowledge is specified in XML. The domain-specific knowledge describes the relationship between components and other constraints that are presumed in requirements documents or too implicit to be extracted directly from the original documents. Then a knowledge base is built from the requirements document in XML using natural language processing (NLP) to parse the documentation and to store the syntax, semantics, and pragmatics information. In this phase, the ambiguity is detected and resolved, if possible. Once the knowledge base is constructed, its content can be queried in NL. Next, the knowledge base is converted with the information of the domain specific knowledge, into Two Level Grammar by removing the contextual dependency in the knowledge base. TLG is used as an intermediate representation to build a bridge between the informal knowledge base and the formal specification language representation. The name “two-level” in Two-Level Grammar comes from the fact that TLG consists of two context-free grammars interacting in a manner such that their combined computing power is equivalent to that of a Turing machine. Our work has refined this notion into a set of domain definitions and the set of function definitions operating on those domains. In order to support object-orientation, TLG domain declarations and associated functions may be structured into a class hierarchy supporting multiple inheritances.

Finally, the TLG code is translated into VDM++, an object-oriented extension of the Vienna Development Method, by data and function mappings. VDM++ is chosen as the target specification language because VDM++ has many similarities in structure to TLG and also has a good collection of tools for analysis and code generation. Once the VDM++ representation of the specification is acquired we can do prototyping of the specification using the VDM++ interpreter.

Also, we can convert this into a high level language such as Java or C++ or into a Rational Rose model in UML using the VDM++ Toolkit. Using XMI5 format, not only the class framework but also its detailed functionalities can be specified and translated into OCL (Object Constraint Language). The structure of the system is shown in Figure1.

2. Integration with Model-Driven Architecture

The method of translating requirements in natural language into UML models and/or executable code described in the previous section may be used to translate business logic into formal rules. Business domain experts from various application domains may express their specification in natural language and then our system translates this into Two-Level Grammar rules via natural language processing (NLP). These rules are encapsulated in a TLG class hierarchy defining a knowledge base with domain ontology, domain feature models (specifying the commonality and variability among the product instances in that domain), feature configuration constraints, feature interdependencies, business operational rules, temporal concerns, etc. TLG specifies the complete feature model including the structural syntax and various kinds of semantic concerns. For example, assume that our application domain is banking. The

business domain will then include a feature model of a bank, which includes specification of the various attributes and operations a bank will have, such as account creation and management, deposit, withdrawal and balance checking operations on individual accounts, etc.

In related work, we have investigated the construction of Generative Domain Models using the Generic Modeling Environment. This tool may also be extended with a natural language processor as a front end, i.e., by applying natural language processing to the business domain model (which is represented in natural language), which can then extract feature model representation rules and then interpret those rules to generate a graphical feature diagram. Platform Independent Models in MDA are based upon the business domains and associated logic for the given application. TLG allows these relationships to be expressed via inheritance. If a software engineer wants to design a server component to be used in bank account management systems, then he/she should write a natural language requirements specification in the form of a UMM (Unified Meta-component Model) describing the characteristics of that component. Our natural language requirements processing system will use the UMM and domain knowledge base to generate platform independent and platform specific UMM specifications expressed in TLG (which we will refer to as UMM-PI and UMM-PS, respectively). UMM-PI describes the bulk of the information needed to progress to component implementation. UMM-PS merely indicates the technology of choice (e.g. CORBA). These effectively customize the component model by inheriting from the TLG classes representing the business domain with new functionality added as desired.

In addition to new functionality, we also impose Quality-of-Service expectations for our components. Both the added functionality and QoS requirements are expressed in TLG so there is a unified notation for expressing all the needed information about components. The translation tool described in the previous section may be used to translate UMM-PI into a PIM represented by a combination of UML and TLG. Note that TLG is needed as an augmentation of UML to define business logic and other rules that may not be convenient to express in UML directly.

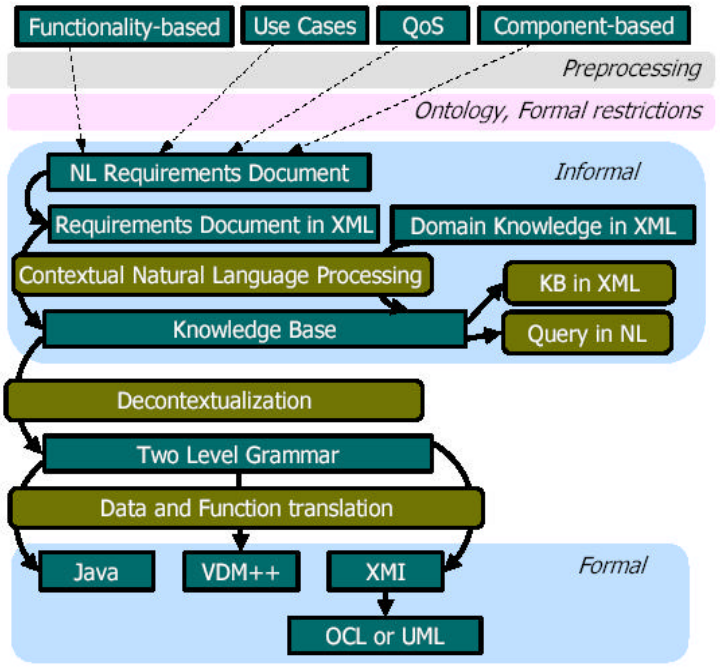


Figure1. Natural Language Requirements Translation into Executable Models

A Platform Specific Model is an integration of the PIM with technology domain specific operations (e.g. in CORBA, J2EE, etc.). These technology domain classes also are expressed in TLG. Each domain contains rules which are specific to that technology, including how to construct glue/wrapper code for components implemented with that technology and architectural considerations such as how to distinguish client code from server code. We express PSMs in TLG as an inheritance from PIM TLG classes and technology domain TLG classes. This means that PSMs will then contain not only the business-domain specific rules but also the technology domain specific rules. The PSM will also maintain the Quality-of-Service characteristics expressed at the PIM level (a related paper explores the rules for this maintenance in more detail and explores this issue for the QoS aspect of access control in particular).

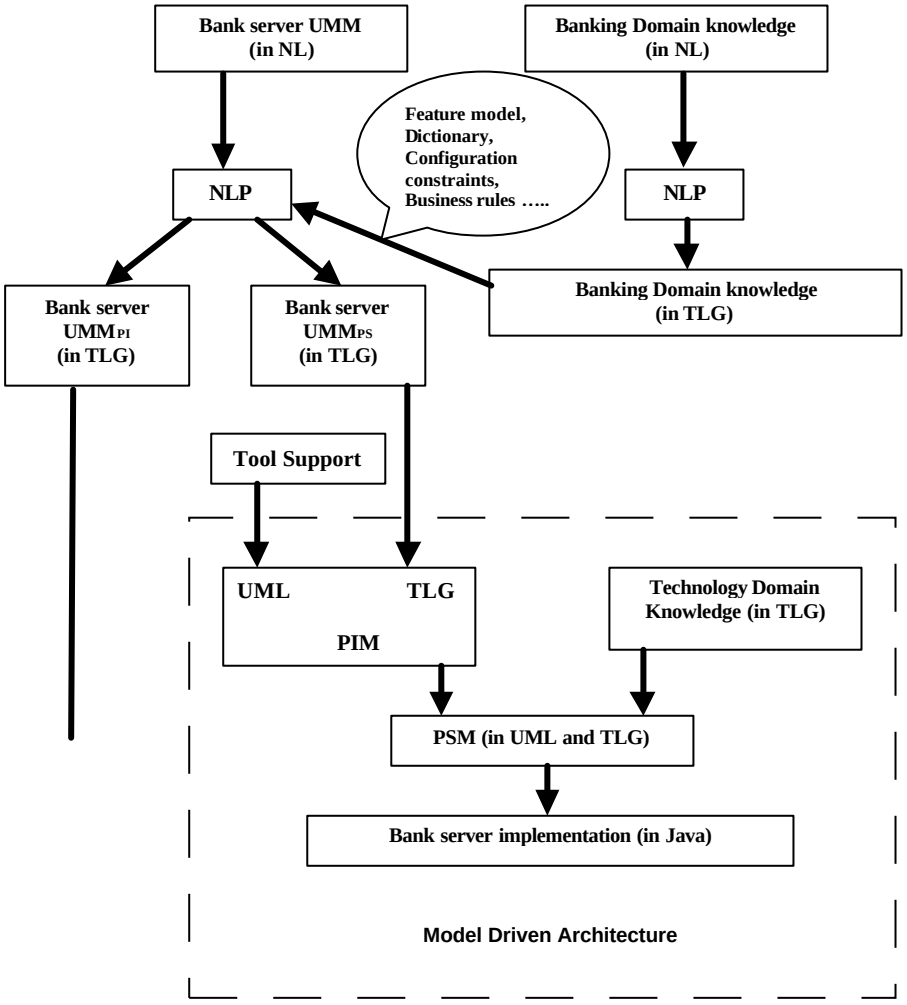


Figure 2. Integration of Two-Level Grammar with Model Driven Architecture

Since the model is expressed in TLG, it is executable in the sense that it may be translated into executable code in a high-level language (e.g. Java). Furthermore, it supports changes at the model level, or even requirements level if the model is not refined following its derivation from the requirements, since the code generation itself is automated.

Figure 2 shows the overall view of the model-driven development from natural language requirements into executable code for the banking example we have just described.

3. Conclusion

This paper has described an approach for unifying the ideas of expressing requirements in natural language, constructing Platform Independent Models for software components, and implementing the components via Platform Specific Models. The approach is specifically targeted at the construction of heterogeneous distributed software systems where interoperability is critical. This interoperability is achieved by the formalization of technology domains with rules describing how those technologies may be integrated together via the generation of glue and wrapper code. The processing of software requirements, construction of PIMs and PSMs, and specification of technology domain rules are all expressed in Two-Level Grammar, thereby achieving a unification of natural language requirements with the Model Driven Architecture approach.

For future work, we will investigate aspect-oriented technology as a mechanism or specifying crosscutting relationships across components and hence improving reusability of components and reasoning about a collection of components. Such aspects of components as functional pre/post conditions and QoS properties crosscut component modules and specification of these aspects spread across component modules. Preliminary work in defining an aspect oriented specification language is very promising.

We are also investigating the applicability of the UniFrame approach to real-time and embedded systems. Real-time constraints are already one of the Quality-of-Service parameters we are now validating. However, we expect that our current timing requirements will need refinement to be applicable in a true real-time setting. We are also looking at applying our modeling technology to the embedded system domain. Finally we are continuing our work in model-driven security to assure that security issues are maintained in migration from PIMs to PSMs.

4. References

1. “Two-Level Grammar as an Object-Oriented Requirements Specification Language,” Proc. HICSS-35, 35th Hawaii Int. Conf. System Sciences, 2002, http://www.hicss.hawaii.edu/HICSS_35/HICSSpapers/PDFdocuments/STDSL01.pdf.
2. “Formal Specification of Generative Component Assembly Using Two-Level Grammar,” Proc. SEKE 2002, 14th Int. Conf. Software Engineering Knowledge Engineering, 2002, pp. 209-212.
3. Burt, C. C., Bryant, B. R., Raje, R. R., Olson, A. M., Auguston, M., “Model Driven Security: Unification of Authorization Models for Fine-Grain Access Control,” to appear in Proc. EDOC 2003, 7th IEEE Int. Enterprise Distributed Object Computing Conf.
4. “Automating Feature-Oriented Domain Analysis ,” to appear in Proc. SERP 2003, 2003 Int. Conf. Software Engineering Research and Practice, 2003.
5. “Assembling Components with Aspect-Oriented Modeling/Specification,” to appear in Proc. WiSME 2003, UML 2003 Workshop Software Model Engineering.
6. “VDM++ - A Formal Specification Language for Object-Oriented Designs,” Proc. TOOLS USA ’92, 1992 Technology of Object-Oriented Languages and Systems USA Conf., 1992, pp. 263-278.
7. GME 2000 User’s Manual, Version 2.0. ISIS, Vanderbilt University, 2001.
8. IFAD, The VDM++ Toolbox User Manual, http://www.ifad.dk, 2000.

9. Jacobson, I., Booch, G., Rumbaugh, J., The Unified Software Development Process, Addison-Wesley, 1999.

10. Jurafsky, D., Martin, J., Speech and Language Processing, Prentice-Hall, 2000.

11. Kiczales, G., et al., “Aspect-Oriented Programming,” Proc. ECOOP ’97, European Conf. Object-Oriented Programming, 1997, pp. 220-242.

12. Lee, B.-S. and Bryant, B. R., “Contextual Knowledge Representation for Requirements

13. Documents in Natural Language,” Proc. FLAIRS 2002, 15th Int. Florida AI Research Symp., 2002, pp. 370-374.

14. Lee, B.-S. and Bryant, B. R., “Contextual Processing and DAML for Understanding Software Requirements Specifications,” Proc. COLING 2002, 19th Int. Conf. Computational Linguistics, 2002, pp. 516-522.

15. Lee, B.-S., Bryant, B. R., “Automation of Software System Development Using Natural Language Processing and Two-Level Grammar,” Proc. 2002 Monterey Workshop Radical Innovations Software and Systems Engineering in the Future, 2002, pp. 244-257.

16. Quatrani, T., Visual Modeling with Rational Rose 2000 and UML, Addison-Wesley, Reading, MA, 2000.

17. Raje, R. R., “UMM: Unified Meta-object Model for Open Distributed Systems,” Proc. ICA3PP, 4th IEEE Int. Conf. Algorithms and Architecture for Parallel Processing, 2000, pp. 454- 465.

18. Raje, R. R., Auguston, M., Bryant, B. R., Olson, A. M., and Burt, C. C., “A Unified Approach for the Integration of Distributed Heterogeneous Software Components,” Proc. 2001

19. Monterey Workshop Engineering Automation for Software Intensive System Integration, 2001, pp. 109-119.

20. Raje, R. R., Auguston, M., Bryant, B. R., Olson, A. M., Burt, C. C., “A Quality of Service-based Framework for Creating Distributed Heterogeneous Software Components,”Concurrency and Computation: Practice and Experience 14, 12 (2002), 1009-1034.

21. Warmer, J., Kleppe, A., The Object Constraint Language: Precise Modeling with UML, Addison-Wesley, 1999.

22. Wilson, W. M., “Writing Effective Natural Language Requirements Specifications,” Naval Research Laboratory, 1999.

23. Yang, C., Lee, B.-S., Bryant, B. R., Burt, C. C., Raje, R. R., Olson, A. M., Auguston, M., “Formal Specification of Non-Functional Aspects in Two-Level Grammar,” Proc.

24. UML 2002 Workshop Component-Based Software Engineering and Modeling Non-FunctionalAspects(SIVOES-MONA),2002,http://www-verimag.imag.fr/SIVOES-ONA/uniframe.pdf.

25. UML – Unified Modeling Language, http://www.omg.org/uml

27 CORBA – Common Object Request Broker Architecture, http://www.corba.org

28. J2EE – Java 2 Enterprise Edition, http://java.sun.com/j2ee

About Authors

Dr. V R Rathod is a Professor and Head in Department of Computer Science, Bhavnagar University, Bhavnagar, Gujarat.
E-mail : profvrr@rediffmail.com

Prof. S M Shah is working as Director in S. V. Institute of Computer Studies, S. V. Campus, Kadi. Gujarat.
E-mail : prof_smshah@yahoo.com

Mr. Nileshkumar K Modi is a Lecturer in S. V. Institute of Computer Studies, S. V. Campus, Kadi, Gujarat.
E-mail : tonileshmodi@yahoo.com

Analysis and Synthesis for Pyramid Based Textures

V Karthikeyani K Duraiswamy P Kamalakkannan

Abstract

This paper describes a method for synthesizing images that match the texture appearance of a given digitized sample. This synthesis is completely automatic and requires only the "target" texture as input. It allows generation of as much texture as desired so that any object can be covered. It can be used to produce solid textures for creating textured 3-d objects without the distortions inherent in texture mapping. It can also be used to synthesize texture mixtures, images that look a bit like each of several digitized samples. The approach is based on a model of human texture perception, and has potential to be a practically useful tool for graphics applications.

Keywords : Image Processing, Texture Analysis, Texture Sysnthesis, Graphic Analysis.

0. Introduction

Computer renderings of objects with surface texture are more interesting and realistic than those without texture. Texture mapping [15] is a technique for adding the appearance of surface detail by wrapping or projecting a digitized texture image onto a surface. Digitized textures can be obtained from a variety of sources, e.g., cropped from a photoCD image, but the resulting texture chip may not have the desired size or shape. To cover a large object you may need to repeat the texture; this can lead to unacceptable artifacts either in the form of visible seams, visible repetition, or both.

Texture mapping suffers from an additional fundamental problem: often there is no natural map from the (planar) texture image to the geometry/topology of the surface, so the texture may be distorted unnaturally when mapped. There are some partial solutions to this distortion problem [15] but there is no universal solution for mapping an image onto an arbitrarily shaped surface. An alternative to texture mapping is to create (paint) textures by hand directly onto the 3-d surface model [14], but this process is both very labor intensive and requires considerable artistic skill. Another alternative is to use computer-synthesized textures so that as much texture can be generated as needed. Furthermore, some of the synthesis techniques produce textures that tile seamlessly.

Using synthetic textures, the distortion problem has been solved in two different ways. First, some techniques work by synthesizing texture directly on the object surface (e.g., [31]). The second solution is to use solid textures[19, 23, 24]. A solid texture is a 3-d array of color values. A point on the surface of an object is colored by the value of the solid texture at the corresponding 3-d point. Solid texturing can be a very natural solution to the distortion problem: there is no distortion because there is no mapping. However, existing techniques for synthesizing solid textures can be quite cumbersome. One must learn how to tweak the parameters or procedures of the texture synthesizer to get a desired effect.

This paper presents a technique for synthesizing an image (or solid texture) that matches the appearance of a given texture sample. The key advantage of this technique is that it works entirely from the example texture, requiring no additional information or adjustment. The technique starts with a digitized image and analyzes it to compute a number of texture parameter values. Those parameter values are then used to synthesize a new image (of any size) that looks (in its color and texture properties) like the original. The analysis phase is inherently two-dimensional since the input digitized images are 2-d. The synthesis

phase, however, may be either two- or three- dimensional. For the 3-d case, the output is a solid texture such that planar slices through the solid look like the original scanned image. In either case, the (2-d or 3-d) texture is synthesized so that it tiles seamlessly.

1. Texture Models

Textures have often been classified into two categories, deterministic textures and stochastic textures. A deterministic texture is characterized by a set of primitives and a placement rule (e.g., a tile floor). A stochastic texture, on the other hand, does not have easily identifiable primitives (e.g., granite, bark, sand). Many real-world textures have some mixture of these two characteristics (e.g. woven fabric, woodgrain, plowed fields).

Much of the previous work on texture analysis and synthesis can be classified according to what type of texture model was used. Some of the successful texture models include reaction diffusion [31, 34], frequency domain [17], fractal [9, 18], and statistical/ random field [1, 6, 8, 10, 12, 13, 21, 26] models. Some (e.g., [10]) have used hybrid models that include a deterministic (or periodic) component and a stochastic component. In spite of all this work, scanned images and hand-drawn textures are still the principle source of texture maps in computer graphics.

This paper focuses on the synthesis of stochastic textures. Our approach is motivated by research on human texture perception. Current theories of texture discrimination are based on the fact that two textures are often difficult to discriminate when they produce a similar distribution of responses in a bank of (orientation and spatial-frequency selective) linear filters [2, 3, 7, 16, 20, 32]. The method described here, therefore, synthesizes textures by matching distributions (or histograms) of filter outputs. This approach depends on the principle (not entirely correct as we shall see) that all of the spatial information characterizing a texture image can be captured in the first order statistics of an appropriately chosen set of linear filter outputs. Nevertheless, this model (though incomplete) captures an interesting set of texture properties.

Computational efficiency is one of the advantages of this approach compared with many of the previous texture analysis/ synthesis systems. The algorithm involves a sequence of simple image processing operations: convolution, subsampling, upsampling, histograming, and nonlinear transformations using small lookup tables. These operations are fast, simple to implement, and amenable to special purpose hardware implementations (e.g., using DSP chips).

2. Pyramid Texture Matching

The pyramid-based texture analysis/synthesis technique starts with an input (digitized) texture image and a noise image (typically uniform white noise). The algorithm modifies the noise to make it look like the input texture (figures 2, 3, 4). It does this by making use of an invertible image representation known as an image pyramid, along with a function, match-histogram, that matches the histograms of two images. We will present examples using two types of pyramids: the Laplacian pyramid (a radially symmetric transform) and the steerable pyramid (an oriented transform).

2.1 Image Pyramids

A linear image transform represents an image as a weighted sum of basis functions. That is, the image, $I(x, y)$, is represented as a sum over an indexed collection of functions, $g_i(x, y)$:

$$I(x, y) = \sum_i y_i g_i(x, y);$$

where y_i are the transform coefficients. These coefficients are computed from the signal by projecting onto a set of projection functions, $h_i(x, y)$:

$$y_i = \sum_{x,y} h_i(x, y) I(x, y):$$

For example, the basis functions of the Fourier transform are sinusoids and cosinusoids of various spatial frequencies. The projection functions of the Fourier transform are also (co-)sinusoids.

In many image-processing applications, an image is decomposed into a set of subbands, and the information within each subband is processed more or less independently of that in the other subbands. The subbands are computed by convolving the image with a bank of linear filters. Each of the projection functions is a translated (or shifted) copy of one of the convolution kernels (see [28] for an introduction to subband transforms and image pyramids).

An image pyramid is a particular type of subband transform. The defining characteristic of an image pyramid is that the basis/ projection functions are translated and dilated copies of one another (translated and dilated by a factor or 2j for some integer j). The subbands are computed by convolving and subsampling. For each successive value of j, the subsampling factor is increased by a factor of 2. This yields a set of subband images of different sizes (hence the name image pyramid) that correspond to different frequency bands.

In an independent context, mathematicians developed a form of continuous function representation called wavelets(see [30] for an introduction to wavelets), that are very closely related to image pyramids. Both wavelets and pyramids can be implemented in an efficient recursive manner, as described next.

Laplacian Pyramid : The Laplacian pyramid [4, 5, 22] is computed using two basic operations: reduce and expand. The reduce operation applies a low-pass filter and then subsamples by a factor of two in each dimension. The expand operation upsamples by a factor of two (padding with zeros in between pixels) and then applies the same low-pass filter. A commonly used low-pass filter kernel (applied separably to the rows and columns of an image) is: 1/16 (1, 4, 6, 4, 1).

One complete level of the pyramid consists of two images, l^0 (a low-pass image), and b^0(a high-pass image), that are computed as follows:

l0 = Reduce(im)

b0 = im - Expand(l0),

where im is the original input image. Note that the original image can be trivially reconstructed from l^0 and b^0:

reconstructed-im = b0 + Expand(l0).

The next level of the pyramid is constructed by applying the same set of operations to the l^0 image, yielding two new images, l^1 and b^1. The full pyramid is constructed (via the make-pyramid function) by successively splitting the low-pass image l^i into two new images, l^{i+1} (a new low-pass image) and b^{i+1} (a new band-pass image).

The combined effect of the recursive low-pass filtering and sub/upsampling operations yields a subband transform whose basis functions are (approximately) Gaussian functions. In other words, the transform represents an image as a sum of shifted, scaled, and dilated (approximately) Gaussian functions. The projection functions of this transform are (approximately) Laplacian-of-Gaussian (mexican-hat) functions, hence the name Laplacian pyramid. Note that the pyramid is *not* computed by convolving the image directly with the projection functions. The recursive application of the reduce and expand operations yields the same result, but much more efficiently.

In the end, we get a collection of pyramid subband images consisting of several bandpass images and one leftover lowpass image. These images have different sizes because of the subsampling operations; the smaller images correspond to the lower spatial frequency bands (coarser scales). Note that the original image can always be recovered from the pyramid representation (via the collapse-pyramid function) by inverting the sequence of operations, as exemplified above.

Steerable Pyramid : Textures that have oriented or elongated structures are not captured by the Laplacian pyramid analysis because its basis functions are (approximately) radially symmetric. To synthesize anisotropic textures, we adopt the steerable pyramid transform [25, 29]. Like the Laplacian pyramid, this transform decomposes the image into several spatial frequency bands. In addition, it further divides each frequency band into a set of orientation bands.

The steerable pyramid was used to create all of the images in this paper. The Laplacian pyramid was used (in addition to the steerable pyramid, see Section 4) for synthesizing the solid textures. Fig 1 shows the analysis/synthesis representation of the steerable pyramid transform. The left-hand side of the diagram is the analysis part (make-pyramid) and the right hand side is the synthesis part (collapse-pyramid). The circles in between represent the decomposed subband images. The transform begins with a high-pass/low-pass split using a low-pass filter with a radially symmetric frequency response; the high-pass band corresponds to the four corners of the spatial frequency domain. Each successive level of the pyramid is constructed from the previous level's lowpass band by a applying a bank of band-pass filters and a low-pass filter.

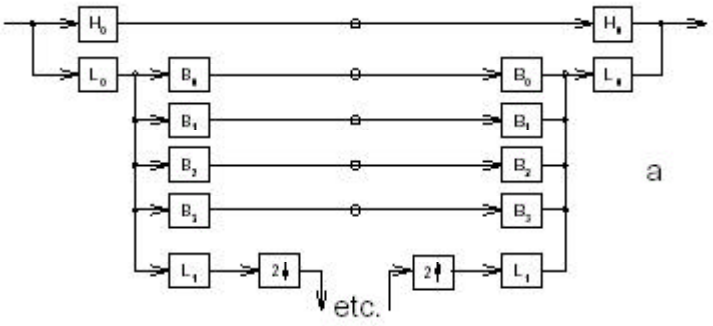


Fig 1: System diagram for the first level of the steerable pyramid. Boxes represent filtering and subsampling operations: H0 is a high-pass filter, L0 and Li are low-pass filters, and Bi are oriented bandpass filters.

The orientation decomposition at each level of the pyramid is “steerable” [11], that is, the response of a filter tuned to any orientation can be obtained through a linear combination of the responses of the four basis filters computed at the same location. The steerability property is important because it implies that the pyramid representation is locally rotation-invariant. The steerable pyramid, unlike most discrete wavelet transforms used in image compression algorithms, is non-orthogonal and over complete; the number of pixels in the pyramid is much greater than the number of pixels in the input image (note that only the low-pass band is subsampled). This is done to minimize the amount of aliasing within each subband. Avoiding aliasing is critical because the pyramid-based texture analysis/synthesis algorithm treats each subband independently.

The steerable pyramid is self-inverting; the filters on the synthesis side of the system diagram are the same as those on the analysis side of the diagram. This allows the reconstruction (synthesis side) to be efficiently computed despite the non-orthogonality. Although the steerable pyramid filter kernels are nonseparable, any nonseparable filter can be approximated (often quite well) by a sum of several separable filter kernels [25]. Using these separable filter approximations would further increase the computational efficiency.

Psychophysical and physiological experiments suggest that image information is represented in visual cortex by orientation and spatial-frequency selective filters. The steerable pyramid captures some of the oriented structure of images similar to the way this information is represented in the human visual system. Thus, textures synthesized with the steerable pyramid look noticeably better than those synthesized with the Laplacian pyramid or some other nonoriented representation. Other than the choice of pyramid, the algorithm is exactly the same.

2.2 Histogram Matching

Histogram matching is a generalization of histogram equalization. The algorithm takes an input image and coerces it via a pair of lookup tables to have a particular histogram. The two lookup tables are: (1) the cumulative distribution function (cdf) of one image, and (2) the inverse cumulative distribution function (inverse cdf) of the other image. An image’s histogram is computed by choosing a binsize (we typically use 256 bins), counting the number of pixels that fall into each bin, and dividing by the total number of pixels. An image’s cdf is computed from its histogram simply by accumulating successive bin counts.

The cdf is a lookup table that maps from the interval [0,256] to the interval [0,1]. The inverse cdf is a lookup table that maps back from [0,1] to [0,256]. It is constructed by resampling (with linear interpolation) the cdf so that its samples are evenly spaced on the [0,1] interval.

These two lookup tables are used by the match-histogram function to modify an image (im1) to have the same histogram as another image (im2):

```
Match-histogram (im1,im2)

im1-cdf = Make-cdf(im1)

im2-cdf = Make-cdf(im2)

inv-im2-cdf = Make-inverse-lookup-table(im2-cdf)

Loop for each pixel do

        im1[pixel] = Lookup(inv-im2-cdf, Lookup(im1-cdf, im1[pixel]))
```

2.3 Texture Matching

The match-texture function modifies an input noise image so that it looks like an input texture image. First, match the histogram of the noise image to the input texture. Second, make pyramids from both the (modified) noise and texture images. Third, loop through the two pyramid data structures and match the histograms of each of the corresponding pyramid subbands. Fourth, collapse the (histogram-matched) noise pyramid to generate a preliminary version of the synthetic texture. Matching the histograms of the pyramid subbands modifies the histogram of the collapsed image. In order to get *both* the pixel and pyramid histograms to match we iterate, rematching the histograms of the images, and then rematching the histograms of the pyramid subbands.

```
Match-texture(noise,texture)
   Match-Histogram (noise,texture)
   analysis-pyr = Make-Pyramid (texture)
   Loop for several iterations do
      synthesis-pyr = Make-Pyramid (noise)
      Loop for a-band in subbands of analysis-pyr
         for s-band in subbands of synthesis-pyr do
     Match-Histogram (s-band,a-band)
       noise = Collapse-Pyramid (synthesis-pyr)
       Match-Histogram (noise,texture)
```

Whenever an iterative scheme of this sort is used there is a concern about convergence. In the current case we have not formally investigated the convergence properties of the iteration, but our experience is that it always converges. However, stopping the algorithm after several (5 or so) iterations is critical. As is the case with nearly all discrete filters, there are tradeoffs in the design of the steerable pyramid filters (e.g., filter size versus reconstruction accuracy). Since the filters are not perfect, iterating too many times introduces artifacts due to reconstruction error.

The core of the algorithm is histogram matching which is a spatially local operation. How does this spatially local operation reproduce the spatial characteristics of textures? The primary reason is that histogram matching is done on a representation that has intrinsic spatial structure. A *local* modification of a value in one of the pyramid subbands produces a *spatially correlated* change in the reconstructed image. In other words, matching the *pointwise* statistics of the pyramid representation does match some of the *spatial* statistics of the reconstructed image. Clearly, only spatial relationships that are represented by the pyramid basis functions can be captured in this way so the choice of basis functions is critical. As mentioned above, the steerable pyramid basis functions are a reasonably good model of the human visual system's image representation.

If we had a complete model of human texture perception then we could presumably synthesize perfect texture matches. By analogy, our understanding of the wavelength encoding of light in the retina allows us to match the color appearance of (nearly) any color image with only three colored lights (e.g., using an RGB monitor). Lights can be distinguished only if their spectral compositions differ in such a way as to produce distinct responses in the three photoreceptor classes. Likewise, textures can be distinguished only if their spatial structures differ in such a way as to produce distinct responses in the human visual system.

2.4 Edge Handling

Proper edge handling in the convolution operations is important. For the synthesis pyramid, use circular convolution. In other words, for an image I(x, y) of size NxN, define: I(x, y) = I(x mod N, y mod N). Given that the synthesis starts with a random noise image, circular convolution guarantees that the resulting synthetic texture will tile seamlessly.

For the analysis pyramid, on the other hand, circular convolution would typically result in spuriously large filter responses at the image borders. This would, in turn, introduce artifacts in the synthesized texture. A reasonable border handler for the analysis pyramid is to pad the image with a reflected copy of itself. Reflecting at the border usually avoids spurious responses (except for obliquely oriented textures).

2.5 Color

The RGB components of a typical texture image are *not* independent of one another. Simply applying the algorithm to R, G, and B separately would yield color artifacts in the synthesized texture. Instead, color textures are analyzed by first transforming the RGB values into a different color space. The basic algorithm is applied to each transformed color band independently producing three synthetic textures. These three textures are then transformed back into the RGB color space giving the final synthetic color texture.

The color-space transformation must be chosen to decorrelate the color bands of the input texture image. This transformation is computed from the input image in two steps. The first step is to subtract the mean color from each pixel. That is, subtract the average of the red values from the red value at each pixel, and likewise for the green and blue bands. The resulting color values can be plotted as points in a three-dimensional color space. The resulting 3-d cloud of points is typically elongated in some direction, but the elongated direction is typically not aligned with the axes of the color space.

The second step in the decorrelating color transform rotates the cloud so that its principle axes align with the axes of the new color space. The transform can be expressed as a matrix multiplication, y = Mx, where x is the RGB color (after subtracting the mean) of a particular pixel, y is the transformed color, and M is a 3x3matrix.

The decorrelating transform M is computed from the covariance matrix C using the singular-value-decomposition (SVD). Let D be a 3xN matrix whose columns are the (mean-subtracted) RGB values of each pixel. The covariance matrix is: $C = DD^t$, where D^t means the transpose of D. The SVD algorithm algorithm decomposes the covariance matrix into the product of three components, $C = US^2U^t$. Here, U is an orthonormal matrix and S^2 is a diagonal matrix. These matrices (C, U and S^2) are each 3x3, so the SVD can be computed quickly. The decorrelating transform is: $M = S^{-1}U^t$, where S is a diagonal matrix obtained by taking the square-root of the elements of S^2.

After applying this color transform, the covariance of the transformed color values is the identity matrix. Note that the transformed color values are: $MD=S^{-1}U^tUSV^t = V^t$. It follows that the covariance of the transformed color values is: $V^tV = I$.

The color transform is inverted after synthesizing the three texture images in the transformed color space. First, multiply the synthetic texture's color values at each pixel byM^{-1}. This produces three new images (color bands) transformed back into the (mean subtracted) RGB color space. Then, add the corresponding mean values (the means that were subtracted from the original input texture) to each of these color bands.

3. Solid Textures

Pyramid-based texture analysis/synthesis can also be used to make isotropic 3-d solid textures. We start with an input image and a block of 3-d noise. The algorithm coerces the noise so that any slice through the block looks like the input image. The solid texture synthesis algorithm is identical to that described above, except for the choice of pyramid: use a 2-d Laplacian pyramid for analysis and a 3-d Laplacian pyramid for synthesis. As usual, match the histograms of the corresponding subbands. Note that since the Laplacian pyramid is constructed using separable convolutions, it extends trivially to three-dimensions.

We have obtained better looking results using a combination of Laplacian and steerable pyramids. On the analysis side, construct a 2-d Laplacian pyramid and a 2-d steerable pyramid. On the synthesis side, construct a 3-d Laplacian pyramid and construct steerable pyramids from all two-dimensional (x-y, x-z, and y-z) slices of the solid. Match the histograms of the 3-d (synthesis) Laplacian pyramid to the corresponding histograms of the 2-d (analysis) Laplacian pyramid. Match the histograms of each of the many synthesis steerable pyramids to the corresponding histograms of the analysis steerable pyramid. Collapsing the synthesis pyramids gives four solids (one from the 3-d Laplacian pyramid and one from each set of steerable pyramids) that are averaged together.

4. Texture Mixtures

Fig 5 shows some texture mixtures that were synthesized by choosing the color palette (decorrelating color transform) from one image and the pattern (pyramid subband statistics) from a second image. One can imagine a number of other ways to mix/combine textures to synthesize an image that looks a bit like each of the inputs: apply match-texture to a second image rather than noise, combine the high frequencies of one texture with the low frequencies of another, combine two or more textures by averaging their pyramid histograms, etc.



Fig 2: (Left) Input digitized sample texture: burled mappa wood. (Middle) Input noise. (Right) Output synthetic texture that matches the appearance of the digitized sample. Note that the synthesized texture is larger than the digitized sample; our approach allows generation of as much texture as desired. In addition, the synthetic textures tile seamlessly.

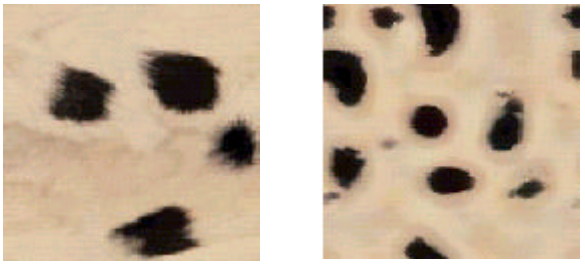


Fig 3: left image is original and right image is synthetic

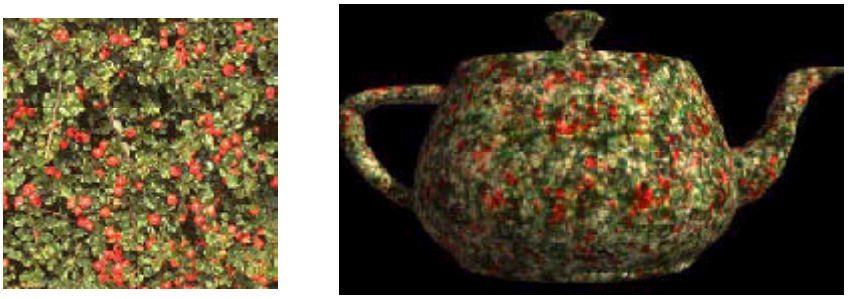


Fig 4: left image is original and right image is synthetic.

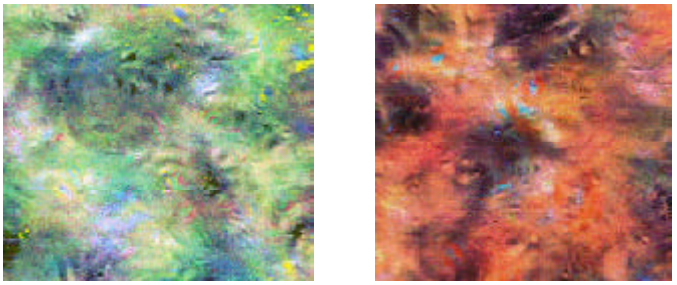


Fig 5: Texture mixtures synthesized by choosing the color palette from one image and the pattern from a second image.

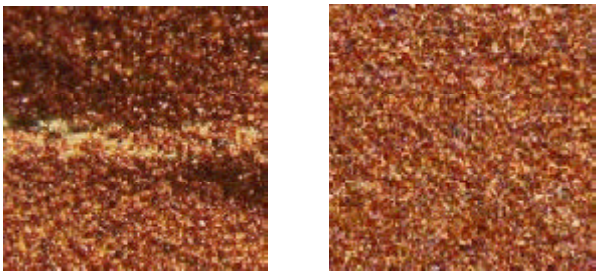


Fig 6: (Left) Inhomogoneous input texture produces blotchy synthetic texture. (Right pair) Homogenous input.

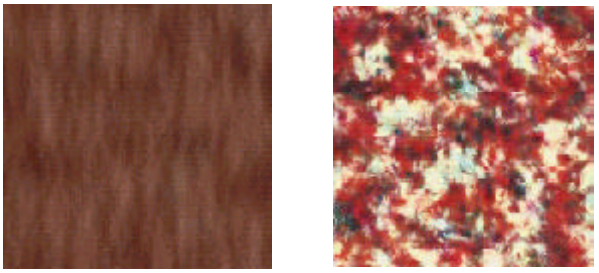


Fig 7: Examples of failures: wood grain and red coral.

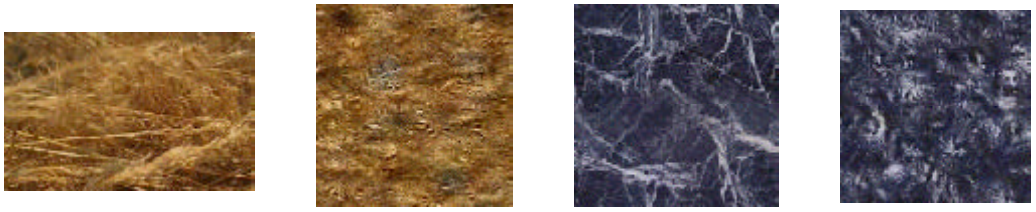


Fig 8: More failures: hay and marble.

5. Conclusion

This paper presents a technique for created a two- or three-dimensional (solid) texture array that looks like a digitized texture image. The advantage of this approach is its simplicity; you do not have to be an artist and you do not have to understand a complex texture synthesis model/procedure. You just crop a textured region from a digitized image and run a program to produce as much of that texture as you want.

6. References

1. C.Bennis, and A.Gagalowicz, “2-DMacroscopic Texture Synthesis”, Computer Graphics Forum 8, 291–300, 1989.
2. J.R.Bergen, “Theories of Visual Texture Perception”, Spatial Vision, D. Regan, Ed. CRC Press, pp. 114–133, 1991.
3. J.R.Bergen, and E.H.Adelson, “Early Vision and Texture Perception”, Nature 333, 363–367, 1988.
4. P.Burt, “Fast Filter Transforms for Image Processing”, Computer Graphics and Image Processing 16, 20–51,1981.
5. P.J.Burt, and E.H.Adelson, “A Multiresolution Spline with Application to Image Mosaics”, ACM Transactions on Graphics 2, 217–236, 1983.
6. R.Chellappa, and R.L.Kashyap, “Texture Synthesis Using 2-D Noncausal Autoregressive Models”, IEEE Transactions on Acoustics, Speech, and Signal Processing 33, 194–203, 1985.
7. C.Chubb and M.S.Landy, “Orthogonal Distribution Analysis: A New Approach to the Study of Texture Perception”, Computational Models of Visual Processing, M. S. Landy and J. A. Movshon, Eds. MIT Press, Cambridge, MA, pp. 291–301, 1991.
8. G.C.Cross, and A.K.Jain, “Markov Random Field Texture Models”, IEEE Transactions on Pattern Analysis andMachine Intelligence 5, 25–39, 1983.
9. A.Fournier, D.Fussel, and L.Carpenter, “Computer Rendering of StochasticModels”, Communications of the ACM 25, 371–384,1982.
10. J.M.Framcos, A.Z.Meiri, and B.Porat, “A Uni-fied TextureModel Based on a 2DWold-Like Decomposition”, IEEE Transactions on Signal Processing 41, 2665– 2678,1993.
11. W.T.Freeman, and E.H.Adelson, “The Design and Use of Steerable Filters”, IEEE Transactions on Pattern Analysis and Machine Intelligence 13, 891–906,1991.
12. A.Gagalowicz, “Texture Modelling Applications”, The Visual Computer 3, 186–200,1987.
13. A.Gagalowicz, and MA, S. D, “Sequential Synthesis of Natural Textures”, Computer Vision,Graphics, and Image Processing 30, 289–315,1985.
14. P.Hanrahan, and P.Haeberli, “Direct WYSIWYG Painting and Texturing of 3D Shapes”, Proceedings of SIGGRAPH 90. In Computer Graphics, vol. 24, ACM SIGGRAPH, pp. 215–223, 1990.
15. P.S.Heckbert, “Survey of Texture Mapping”, IEEE Computer Graphics and Applications 6, 56–67,1986.
16. M.S.Landy, and J.R.Bergen, “Texture Segregation and Orientation Gradient”, Vision Research 31, 679–691,1991.
17. J.P.Lewis, “Texture Synthesis for Digital Painting”, Proceedings of SIGGRAPH 84. In Computer Graphics, vol. 18, ACM SIGGRAPH, pp. 245–252,1984.
18. J.P.Lewis, “Generalized Stochastic Subdivision”, ACM Transactions on Graphics 6, 167–190,1987.
19. J.P.Lewis, “Algorithms for Solid Noise Synthesis. Proceedings of SIGGRAPH 89. In Computer Graphics, vol. 23, ACM SIGGRAPH, pp. 263–270,1989.
20. J.Malik, and P.Perona, “Preattentive Texture Discrimination with Early Vision Mechanisms”, Journal of the Optical Society of America A 7, 923–931,1990.

21. T.Malzbender, and S.Spach, “A Context Sensitive Texture Nib”, Communicating with Virtual Worlds, N. M. Thalmann andD. Thalmann, Eds. Springer-Verlag,New York, pp. 151–163,1993.

22. J.M.Ogden, E.H.Adelcon, J.R.Bergen, and P.J.Burt, “Pyramid-Based Computer Graphics”, RCA Engineer 30, 4–15,1985.

23. D.R.Peachy, “Solid Texturing of Complex Surfaces”, Proceedings of SIGGRAPH 85. In Computer Graphics, vol. 19, ACM SIGGRAPH, pp. 279–286,1985.

24. K.Perlin, “An Image Synthesizer”, Proceedings of SIGGRAPH 85. In Computer Graphics, vol. 19, ACM SIGGRAPH, pp. 287–296,1985.

25. P.Perona, “Deformable Kernels for Early Vision”, IEEE Transactions on Pattern Analysis and Machine Intelligence, To appearMay 1995.

26. K.Popat, and R.W.Picard, “Novel Cluster-Based Probability Model for Texture Synthesis, Classification, and Compression”,In Proceedings of SPIE Visual Communications and Image Processing, pp. 756–768,1993.

27. D.L.Ruderman, and W.Bialek, “Statistics of Natural Images: Scaling in the Woods”, Physical Review Letters 73, 814–817,1994.

28. E.P.Simoncelli, and E.H.Adelson, “Subband Transforms”, Subband Image Coding, J. W. Woods, Ed. Kluwer Academic Publishers, Norwell, MA, 1990.

29. E.P.Simoncelli, W.T.Freeman, E.H.Adelson, and D.J.Heeger, “Shiftable Multi-Scale Transforms”, IEEE Transactions on Information Theory, Special Issue on Wavelets 38, 587–607,1992.

30. G.Strang, “Wavelets and Dilation Equations: A Brief Introduction”, SIAM Review 31, 614–627,1989.

31. G.Turk, “Generating Textures on Arbitrary Surfaces Using Reaction-Diffusion”, Proceedings of SIGGRAPH 91. In Computer Graphics, vol. 25, ACM SIGGRAPH, pp. 289– 298,1991.

32. M.R.Turner, “Texture Discrimination by Gabor Functions”, Biological Cybernetics 55, 71–82,1986.

33. L.Williamsm, “Pyramidal Parametrics”, Proceedings of SIGGRAPH 83. In Computer Graphics, vol. 17, ACM SIGGRAPH, pp. 1–11,1983.

34. A.Witkin, and M.Kass, “Reaction-Diffusion Textures”, Proceedings of SIGGRAPH 91. In Computer Graphics, vol. 25, ACM SIGGRAPH, pp. 299–308,1991.

About Authors

Mr. V Karthikeyani is a Research Scholar in Department of Computer Science, K. S. R. College of Technology, Tiruchengode, Tamilnadu
Email : karthi_vajram@yahoo.com

Dr. K. Duraiswamy is a Principal in K. S. R. College of Technology, Tiruchengode, Tamilnadu

Mr. P Kamalakkannan is a Research Scholar in Department of Computer Science, K. S. R. College of Technology, Tiruchengode, Tamilnadu

Critical Challenges in Natural Language Processing

Veena A Prakashe

Abstract

In this paper, the author attempts to enlist some of the basic bottlenecks that pose challenges while designing the automation of any natural language understanding system. In the beginning, some background material on the study of language and an overview of linguistics is presented for the benefit of the reader who might be new to the fields of artificial intelligence and cognitive science. Natural language systems are also discussed briefly so as to give a better insight into the processing of natural languages by computer systems. Then the three major threats or challenges of natural language processing, viz. knowledge acquisition from natural language; interaction with multiple underlying systems; and partial understanding of multi-sentence and fragments of language are discussed.

Keywords : Natural Language Processing.

0. Introduction

What is natural language? Natural language is any language that humans learn from their environment and use to communicate with each other. Whatever the form of the communication, natural languages are used to express our knowledge and emotions and to convey our responses to other people and to our surroundings. Natural languages are usually learned in early childhood from those around us. Children seem to recognize at a surprisingly early age the value of structure and uniformity in their utterances. Words, phrases, and sentences replace grunts, whines, and cries and better serve to convince others to recognize the child's needs. Natural languages can be acquired later in life through school, travel, or change in culture, but with very few exceptions, all humans in all cultures learn to communicate verbally in the language natural to their immediate environment.

In contrast to natural languages, artificial languages are languages created by humans to communicate with their technology, for example, computer programming languages.

Human process natural languages whenever they read Shakespeare, dictate a business letter, or tell a joke. Sign language is used by the hearing impaired to communicate thoughts and feelings with others and replaces the language they are unable to hear. Despite the different forms of language in each of these situations, aspects of the language used are similar. Whether language is spoken or written, message has a structure and the elements of language relate to each other in recognizable ways. Verbal communication or speech is characterized by the sounds which almost every human is capable of producing. Whether each person learns to produce a particular sound is determined by the languages learned rather than the anatomical speech production mechanisms. which are approximately the same for all normal humans. Speech is produced by stringing together individual human sounds in recognized patterns. The study of these patterns of sounds is called *phonology*. The study of the structure of language units and their relationships is called *syntax*. Phonology and syntax are both important parts of the field of linguistics.

Linguists are also concerned with semantics, the study of the relationship between the linguistics structures used and the meanings intended; in other words, how does what we say or write relate to what we mean? It is not enough for a sentence to be correct in form; it must also make sense. For example, the sentence.

The tree sang the chair.

Would not in ordinary discourse, be a meaningful sentence, even though it is grammatically reasonable. The noun phrase, The tree, can be the subject of a sentence; sang is obviously a verb; and the chair is a noun phrase which can serve as a direct object of a verb. But trees do not sing, and nothing that sings, sings chairs. So, how can this string of words be a sentence, even an unreasonable one?

We recognize that the following sentences have the same structure.

> The tree sang the chair.
>
> The students finished the exam.

The new, young vice-president in change of financial affairs in the company established extraordinary regulations concerning the procedures for reporting exceptional situations in the payroll department.

Thus, something in language makes us aware of similarities among sentences despite the variance in subject matter. The systems used by linguists to describe these similarities are called grammars. The term grammar is also used to refer to the methods taught in school such as diagramming sentences. These methods are designed to show the relationships among the various structures within sentences. Grammars consist of the elements allowed within sentences and the rules for putting these elements together. For example, the structure of some sentences can be described as a noun phrase followed by a verb phrase. In this case, the elements are the sentence, a noun phrase, and verb phrase. A rule to express their relationship could be written as:

SENTENCE ✍ NOUN PHRASE + VERB PHRASE

Obviously further definition of these elements would be required to describe noun phrases, verb phrases, and their components, and each of the components would have to be defined. This process of redefinition of the grammatical constructs would be continued until the elements were defined as specific words. The words in a grammar are called the vocabulary. The grammar is made up of the rules and the vocabulary along with the meanings associated with the vocabulary.

Besides linguists' use of grammars for describing language, grammars are used by logicians and formal philosophers to study formal languages. A formal grammar is essentially a set of rules and a list of elements upon which the rules can be applied.

Other logicians have sought to represent natural language by means of prepositional logic. All sentences in the language considered are written as propositions and can be manipulated according to the rules of formal logic. The statement, *All teenagers drive cars*, could be rewritten in logical notation as:

FOR-ALL (x) (EXISTS (y)(TEENAGER(x) (CAR(y) AND DRIVE(x,y))))

When a sentence has been thus transcribed, the rules of logic can be applied to test the validity of any references to the information contained in the sentence. However, prepositional logic can express only a subset of all sentences in any natural language. The method only applies to statements about which the truth or falsity can be known.

Cognitive psychologists, concerned with how humans think, approach language from a different perspective than linguists and logicians. They view language as a representational medium for thought rather than viewing language as an independent phenomenon.

Writing personal letters is a good example of social communication. Language used for social purposes follows the same rules as other language, but frequently is highly formulaic. The same phrases are used over and over in similar situations, often losing their meaning somewhat, yet still serving their basic function. For example,

> I love you.
>
> Hello, how are you? Fine, thank you, and you?
>
> Thank you very much for the gift. I like it a lot.
>
> Hey, bro. Wha's hap'nin'?

Dealing with language of this sort requires different analysis techniques from other language. The meaning behind the words seems to be of a different nature than the content of language used to convey information. Yet the notion of language as social interaction still fits the paradigm of the human information processing system.

1. Text Processing

Much of the information processed by computers is text, data of the type generally called character or alphanumeric. In natural language processing, all written material is text.

2. Characteristics of Text

Dealing with text is both simpler and harder than manipulating numeric data. In terms of the physical characteristics, it is simpler in that text is linear; the first character is handled, then the second, then the third, until the last is reached. At that point the data is processed. But in logical terms, text is quite slippery. Generally, in the computer, numeric data is represented in a specific form; a number is given a fixed quantity of bits and a set format. All integers occupy the same amount of storage in memory in a particular computer, as do real numbers. (Of course, extended precision may increase the amount, but it is still a fixed amount.) Text, on the other hand, is made up of words and names and other strings of characters, which are many different lengths, and thus require differing amounts of memory for storage.

Connected text, such as this paragraph which you are reading, can be handled as a linear string of characters, then broken up into words of varying length, which could then be processed. A word is defined in text processing as the string of characters, usually alphabetic, that fall between delimiters: blanks, commas, periods, parentheses, and any other allowable punctuation marks that indicate the end of a word. This definition covers some forms of text besides connected text, such as business letters and mailing lists containing names and addressed, or bibliographic data with various fields separated by specific punctuation marks. Much of this type of data is not referred to as natural language because it is not in sentence form. However, parts of some fields look very much like natural language, such as titles of books. This type of textual data has been the primary object of text editors and word processing systems without much concern for the language involved.

3. MARC Format and WEBMARC

A format designed to handle the various problems of dealing with text was developed by the U.S. Library of Congress MARC (MAchine Readable Catalog) Project in 1967. The MARC format has been used for a variety of library projects including communication and information exchange among the many libraries with machine readable information. Donald Sherman adapted the MARC format for recording dictionary data, specifically Webster's Seventh Collegiate Dictionary (known as W7), which was originally recorded

in machine readable form by the Lexicographical Project at SDC. WEBMARC, Sherman's version of W7 in MARC format, contains 68,657 entries, each stored as a variable-length record representing the information about one word in the dictionary.

In WEBMARC, the leader is the first 24 characters in each record and contains fixed-length fields recording the record length, status (F for full record), source of data (W7), record extension number (usually, 0, for non-extended record, 1, 2,... for entry requiring more space than one physical record), an address pointer to the data part of the record, and a record identification number. The record directory follows the leader and is made up of a series of fixed-length (12 character) segments containing a tag identifying each part of the lexical entry, the address of the first character in that part, and the length of that part. The tag fields are three- digit numbers, the first digit of which identifies the type of the field.

The WEBMARC format illustrates several of the methods described for recording variable data, such as dictionary entries and bibliographical citations. It is not especially efficient in that no data compression is used and many of the record fields take up more space than required.

4. Design of Natural Language Systems

A natural language system designed to understand and manipulate language should be capable of accepting input in natural language text, storing knowledge related to the application domain, drawing inferences from that knowledge, answering questions based on the knowledge, and generating responses.

5. General Description of NLS

NLS is a knowledge based natural language understanding system. It processes natural language input and generates appropriate output. The knowledge base for this system is precompiled; in other words, a knowledge domain exists before execution begins. This knowledge base (KB) preserves both hierarchical and prepositional information about the data stored. The input accepted by the system includes statements to be paraphrased, i.e., restated to assure proper understanding; statements which represent knowledge to be learned, i.e., added to the KB; and questions to be answered by accessing the KB. The system outputs appropriate responses to input, as well as paraphrases of statements and answers to questions.

The major modules of NLS include the Parser, the Understander, and the Generator. The parser accepts the input string and maps in into an internal structure compatible with the KB. The generator maps from the internal structure to the output string. The understander module accesses the KB for various purposes; to obtain knowledge, to draw inferences, or to add knowledge to the KB. These three functions interact to accomplish various tasks.

6. Critical challenges for natural language processing

This paper identifies the problems that we believe must block widespread use of computational linguistics.

Knowledge acquisition from natural language (NL) texts of various kinds, from interactions with human beings, and from other sources. Language processing requires lexical, grammatical, semantic, and pragmatic knowledge.

Interaction with multiple underlying systems to give NL systems the utility and flexibility demanded by people using them. Single application systems are limited in both usefulness and the language that is necessary to communicate with them.

Partial understanding gleaned from multi-sentence language, or from fragments of language. Approaches to language understanding that require perfect input or that try to produce perfect output seem doomed to failure because novel language, incomplete language, and errorful language are the norm, not the exception.

7. State-of-the-art

The limitations of practical language processing technology have been summarized as follows:

Domains must be narrow enough so that the constraints on the relevant semantic concepts and relations can be expressed using current knowledge representation techniques, i.e. primarily in terms of types and sorts. Processing may be viewed abstractly as the application of recursive tree rewriting, including filtering out tree out matching a certain pattern.

Handcrafting is necessary, particularly in the grammatical components of systems (the component technology that exhibits least dependence on the application domain). Lexicons and axiomatizations of critical facts must be developed for each domain, and these remain time-consuming tasks.

The user must still adapt to the machine, but, as the products testify, the user can do so effectively.

Current systems have limited discourse capabilities that are almost exclusively handcrafted. Thus current systems are limited to viewing interaction, translation, and writing and reading text as processing a sequence of either isolated sentences or loosely related paragraphs. Consequently, the user must adapt to such limited discourse.

It is traditional to divide natural language phenomena (and components of systems designed to deal with them) into three classes:

Syntactic phenomena- those that pertain to the meaning of a sentence and the order of words in the sentence, based on the grammatical classes of words rather than their meaning.

Semantic phenomena- those that pertain to the meaning of a sentence relatively independent of the context in which that language occurs.

Pragmatic phenomena- those that relate the meaning of a sentence to the context in which it occurs. This context can be linguistic (such as the previous text or dialogue), or nonlinguistic (such as knowledge about the person who produced the language, about the goals of the communication, about the objects in the current visual field, etc.).

8. Knowledge acquisition for language processing

It goes without saying that any NLP system must know a fair amount about words, language, and some subject area before being able to understand language. Currently, virtually all NLP systems operate using fairly laboriously hand-built knowledge bases. The knowledge bases may include both linguistic knowledge (morphological, lexical, syntactic, semantic, and discourse) and nonlinguistic knowledge (semantic world knowledge, pragmatic, planning, inference), and the knowledge in them may be absolute or probabilistic. (Not all of these knowledge bases are necessary for every NLP system).

9. Types of knowledge acquisition

Just as there are many kinds of knowledge, there are a number of different ways of acquiring that knowledge:

Knowing by being pre-programmed - this includes such things as hand-built grammars and semantic interpretation rules.

Knowing by being told - this includes things that a human can "tell" the system using various user-interface tools, such as semantic interpretation rules that can be automatically built from examples, selectional restrictions, and various lexical and morphological features.

Knowing by looking in up - this means using references such as an online dictionary, where one can find exactly the information that is being sought.

Knowing by using source material – this means using references such as an encyclopedia or a corpus of domain-relevant material, from which one might be able to find or infer the information being sought; it may also mean using large volumes of material as the source of probabilistic knowledge (e.g., bank is more likely to mean a financial institution than the side of a river).

Knowing by figuring it out - this means using heuristics and the input itself (such as the part of speech of words surrounding an unknown word). Knowing by using a combination of the above techniques- this may or may not involve human intervention.

10. Interfacing to multiple underlying systems

Most current NL systems, whether accepting spoken or typed input, are designed to interface to a single homogeneous underlying system; they have a component geared to producing code for that single class of application systems, such as a relational database (Stallard, 1987; Parlance User Manual, Learner User Manual.) These systems take advantage of the simplicity of the semantics and the availability of a formal language (relational calculus and relational algebra) for the system's output.

The challenge is to recreate a systematic, tractable procedure to translate from the logical expression of the user's input to systems that are not fully relational, such as expert system functions, object-oriented and numerical simulation systems, calculation programs, and so on. Implicit in that challenge is the need to generate code for non-homogeneous software applications- those that have more than one application system.

The norm in the present generation of user environments is distributed, networked applications. A seamless, multi-model, NL interface should make use of a heterogeneous environment feasible for users and, if done well, transparent. Otherwise, the user will be limited by the complexity, idiosyncrasy, and diversity of the computing environment.

Such interfaces will be seamless in at least two senses:

The user can state information needs without specifying how to decompose those needs into a program calling the various underlying systems required to meet those needs. Therefore, no seams between the underlying systems will be visible.

The interface will use multiple input/output modalities (graphics, menus, tables, pointing, and natural language). Therefore, there should be no seams between input/output modalities.

Although acoustic and linguistic processing can determine what the user wants, the problem of translating that desire into an effective program to achieve the user's objective is a challenging, but solvable problem.

In order to deal with multiple underlying systems, not only must our NL interface be able to represent the meaning of the user's request, but it must also be capable of organizing the various application programs at its disposal, choosing which combination of resources to use, and supervising the transfer of data among them.

Partial understanding of fragments, novel language, and errorful language

It is time to move away from dependence on the sentence as the fundamental unit of language. Historically, input to NL systems has often had to consist of complete, well-formed sentences. The systems would take those sentences one at a time and process them. But language does not always naturally occur in precise sentence-sized chunks. Multi-sentence input is the norm for many systems that must deal with newspaper articles or similar chunks of text. Subsentence fragments are often produced naturally in spoken language and may occur as the output of some text processing. Even when a sentence is complete, it may not be perfectly formed; errors of all kinds, and new words, occur with great frequency in all applications.

11. Multi-sentence input

Historically, computational linguistics has been conducted under the assumption that the input to a NL system is complete sentences (or, in the case of speech, full utterances) and that the output should be a complete representation of the meaning of the input. This means that NL systems have traditionally been unable to deal well with unknown words, natural speech, language containing noise or errors, very long sentence (say, over 100 words), and certain kinds of constructions such as complex conjunctions.

12. Errorful language; including new words

Handling novel, incomplete, or errorful forms is still an area of research. In current interactive systems, new words are often handled by simply asking the user to define them. However, novel phrases or novel syntactic/ semantic constructions are also an area of research. Simple errors, such as spelling or typographical errors resulting in a form not in the dictionary, are handled in the state-of-the-art technology, but far more classes of errors require further research.

The state-of-the-art technology in message understanding systems is illustrative. It is impossible to build in all words and expressions ahead of time. As a consequence, approaches that try for full understanding appear brittle when encountering novel forms or errorful expressions.

The state of the art in spoken language understanding is similarly limited. New words, novel language, incomplete utterances, and errorful expressions are not generally handled. Including them poses a major roadblock, for they will decrease the constraint on the input set, increase the perplexity of the language model, and therefore decrease reliability in speech recognition.

The ability to deal with novel, incomplete, or errorful forms is fundamental to improving the performance users can expect from NLP systems.

13. Conclusion

We feel that knowledge acquisition, interaction with multiple underlying systems, and techniques for partial understanding are the three solvable problems that will have the most impact on the utility of natural language processing. The norm in the present generation of user environments is distributed, networked application. A seamless, multi-modal, natural language system should make use of a heterogeneous environment feasible for users, otherwise, the users may be limited by the complexity, idiosyncrasy, and diversity of the computing environments.

14. References

1. Bates, M, and Weischedel, R.M. (1993). Challenges in Natural Language Processing. Cambridge University Press. pp3-33.
2. Harris, Mary Dee. (1985). Introduction to Natural Language Processing. Reston Publishing. Company, Inc. pp55-66.
3. Bates, M., Boisen, S., and Makhoul, I (1991). "Developing an Evaluation Methodology for Spoken Language Systems", DARPA Speech and Natural Language Workshop, Hidden Valley, PA, Morgan Kaufmann Publishers, pp. 102-108.
4. Bobrow, R., Ingria, R., and Stallard, D. (1991). "Syntactic and Seminatic Knowledge in the DELPHI Unification Grammar," DARPA Speech and natural Language Workshop, Hidden Valley, PA, Morgan Kaufmann Publishers, pp. 230-236.
5. Neal, J., and Walter, S. (editors). (1991). Natural Language Processing Systems Evaluation. Workshop, Rome Laboratory.
6. Weischedel, R. M., Carbonell, J., Grosz, B., Marcus, M., Perrault, R., and Wilensky, R. (1990). Natural Language Processing, Annual Review of Computer Science, Vol.4, pp. 435-452.

About Author

Mrs. Veena A Prakashe is presently working as Information Scientist in Nagpur University Library. She holds M.Sc. (CSc.), MLISc, Diploma in German Language. She looking after the Computerization and Networking of Nagpur University Library, UGC- Info-Net Project. Research Experience : Worked on various Expert systems/ KBS projects of DoE and MEF, GoI, in CEERI, Pilani and NEERI, Nagpur. (Both CSIR Laboratories). Publications : a) A Book titled "DBASE III PLUS" in 1992, published by Pitamber Publishing House, New Delhi, financed by DoE, GoI. b) 9 papers published in the Conf. Proc. of various National Conferences. Membership : A life member of IWSA (Indian Women Scientist Association)

E-mail : sh_veena@hotmail.com

UNL Nepali Deconverter

Birendra Keshari Sanat Kumar Bista

Abstract

This paper discusses about the Interlingua approach of machine translation, especially the Nepali generator part of Interlingua based machine translation in which the Interlingua used is UNL (Universal Networking Language). Nepali is the national language of Nepal, a country in Indian sub continental region. UNL is an Interlingua proposed by United Nations University/Institute Of Advanced Studies, Tokyo, Japan to remove language barrier and digital divide in the World Wide Web. This paper describes about the architecture and design of UNL Nepali Deconverter (Generator), that has been implemented using a tool called DeCo*, a language neutral generator. Nepali sentences are generated using information present in Nepali language at different linguistic levels. Information like case relations, case markers etc. in Nepali sentences can be generated from morphological level itself since Nepali is a morphologically rich language.*

Keywords : Universal Networking Language, Machine Translation, Nepali Language.

0. Introduction

There are several trends in Machine Translation Systems. Interlingua approach is one of them. In this approach the source language sentences are first analyzed and converted to an intermediate form called Interlingua, which is an equivalent semantic form of the source language. The Interlingua representation is then analyzed using source-target language dictionary and grammar to generate the target language sentences. In UNL based system, Enconverter analyzes the source language to produce UNL and Deconverter generates the target language. UNL as such has been designed as a standard Interlingua (Uchida and Zhu., 2002). Enconverter and Deconverter provide language neutral framework for source language analysis and target language generation. UNL is going to be the future language for computer (Uchida and Zhu., 2002) . This paper describes the Deconverter module for Nepali.

While English follows SVO pattern, Nepali follows SOV pattern. Nepali is a free word order language. This is due to the reason that in Nepali, thematic case relations of nouns and pronouns, number, tense, gender and honor markers of verbs are conveyed by suffixes.

1. Universal Networking Language (UNL)

UNL is an artificial digital language that represents meaning sentence by sentence. The representation is in logical form. Such logically formed expressions can be viewed as a semantic net or an acyclic directed hyper graph where a node can be a graph itself. Each node represents UW (Universal Word) or concepts. The arc represents relation between the two concepts. So, UNL can also be viewed as a set of binary relations between the concepts.

For example the Nepali sentence in transliterated form, ‘Ram kaathmaandu yunivarsithimaa padhcha’, which means ‘Ram reads in kathmandu university’, can be expressed using following UNL expressions:

{unl}
[S]
agt(read.@entry.@present,Ram)
plc(read.@entry.@present,university)
mod(university,kathmandu)
[/S]
{/unl}
The above UNL expressions can be represented as a graph in the following way.

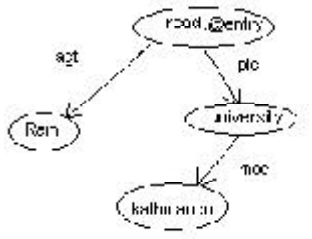


Figure 1. Graph representation of above UNL expressions.

In the above example read, student etc. are UWs and agt, plc and mod are UNL relations. The symbols starting with '@' character like @entry and @present are called UNL attributes.

UWs are based on English words but they are made unambiguous by their position in UNL Knowledge Base (KB). UNL KB maintains the hierarchy of concepts that are universal i.e. the concepts is not any language, culture or tradition specific (Uchida and Zhu., 2002) . Furthermore, restricted UW is used to avoid ambiguity by restricting the concept. For example the word "book" can represent two concepts; 'a thing' or 'act (of booking)'. So, it is disambiguated by using restricted UW like book(icl>do) or book(icl>thing).

Relations represent the semantic role such as agent, object, condition, and, place, co-agent etc. that UWs play. Attributes are attached to UWs to express the objectivity of the sentence. There are several such relations and attributes specified *UNL Specification of UNL Center* (UNL Center, 2003). UW's and relations express subjectivity of the sentence. More information about UNL can be found in Deconverter Specification of UNL Center (UNL Center, 2000), UNL Specification of UNL Center (UNL Center, 2003), (Uchida and Zhu., 2002) and many others.

1.1 UNL Benefits

Once the information is converted to UNL form, it becomes language neutral and it can be converted to other different languages. Thus, it can be used for information exchange between languages. Information in a source language can be converted to UNL using source language Deconverter and then using Enconverter of target language, UNL can be enconverted in to that language.

Since, UNL is in logical form, knowledge processing can be done unambiguously to produce useful and desired results.

2. UNL Nepali Deconverter

A tool called *DeCo* has been designed by UNU/IAS as a language independent generator that provides synchronously a framework for morphological and syntactic generation and word selection for natural collocation. Its structure has been shown in figure 2.

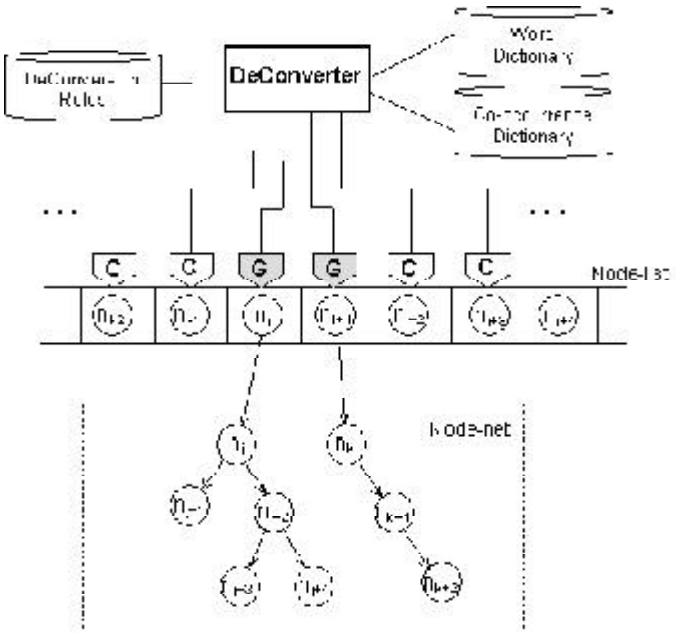


Figure 2. Deconverter Structure.

It can deconvert both context-sensitive and context-free languages. It uses target language specific Word Dictionary, Co-occurence Dictionary and Deconversion rules to generate the target language. So, developing a Deconverter for a language means developing dictionaries and writing deconversion rules, which are understood by the *DeCo* and these are language dependent. The structure of a *DeCo* has been shown in figure 2.

Each entry in Word Dictionary includes native language Head Word, corresponding UW, and the attributes. Attributes include grammatical and semantic attributes. An example of an entry in Nepali Language Word Dictionary Attributes can be:

[kitaaba] “book(icl>thing)” (N,C,INANI,PHY)<N,0,0>;

In the above example [kitaaba] is the Nepali Head Word, book(icl>thing) is UW and (N,C,INANI,PHY) is the attribute list.

First, the deconversion rules are converted into binary format and then binary format rules are loaded. The UNL expressions are converted in to semantic net called Node-net. The UWs are replaced with corresponding native language Head Words. If it is not possible to unambiguously decide the correct Head Word for a given UW, Co-occurence dictionary is used. Co-occurence dictionary contains more semantic information for proper word selection without the ambiguity. But the use of Co-occurence dictionary is optional. We have not used Co-occurence dictionary for UNL Nepali Deconverter.

Node-net represents the hyper graph (a representation of UNL expressions) that has not yet been visited. Each node contains certain attributes initially loaded from the Language Dictionary and sometime generated by *DeCo* during runtime. These attributes can be read or deleted or new attributes can be added. This is governed by deconversion rules. Each node in the Node-net is traversed and inserted in to the Node-list.

Node-list shows the current list of nodes that the Deconverter can look at through its windows. Node-list includes two generation windows circumscribed by condition windows. At the initial stage before any deconversion rule application there are three nodes in the Node-list, *Sentence Head* node, *Entry* node and *Sentence Tail* node. This is explained in *Deconverter Specification of UNL Center* (UNL Center, 2000). The generation occurs at the generation windows, when the conditions in the condition windows are satisfied.

The result of rule application is operation on the nodes in Node-list like changing attributes, copy, shift, delete, exchange etc. and/or insertion of nodes from Node-net to Node-list. The rule application halts when either Left Generation Window reaches the *Sentence Tail* node or Right Generation Window reached the *Sentence Head* node. If post-editing is required the Deconverter will start applying post editing rules. Post editing rule has not been used for UNL Nepali Deconverter. At the end, the nodes in the Node-list represent the generated sentence. More information about *DeCo* can be found in *Deconverter Specification of UNL Center* (UNL Center, 2000).

3. Architectural Design

There are basically two modules for UNL Nepali Deconversion; Syntax Planning Module and Morphology Generation Module. The overall architecture and structure of Nepali Deconverter has been shown in figure 3.

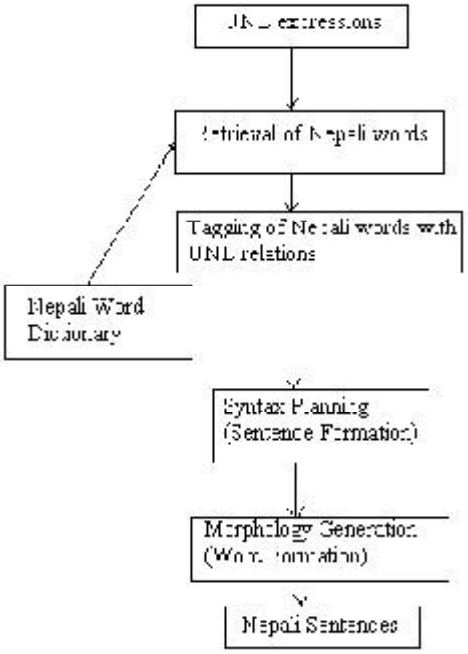


Figure 3. Nepali Deconverter Structure.

3.1 Syntax Planning Module

This module is responsible for Nepali sentence formation by syntax planning. In UNL relation *r*el(UW1,UW2), UW1 is the parent node and UW2 is the child node. We plan the syntax, by deciding which child to insert first and at what position (left or right) with respect to other child of its parent. This is done by creating a (M+1)x(M+1) priority matrix where M is the total number of relations. We write the relation labels in the first row and first columns. Each Mij can be 'L', 'R' or nothing (we represented it by '-'), where i is the row number and j is the column number.

Mij = 'L' means that the child labelled with relation label in row *i* is to be inserted in to the Node-List to the left of the child labelled with relation label in column*j*. Similarly, Mij = 'R' means that the child labelled with relation label row*i* is to inserted in to the Node-List to the right of the child labelled with relation in column *j*. Mij = - means the position with respect to each other is not applicable. A rank for each relation label is calculated by adding the number of 'R' in the row of each relation label. The higher the value of the rank the further right from the main verb is the corresponding word.

	agt	obj	ben	Rank
agt	-	L	L	0
obj	R	-	R	2
ben	R	L	-	1

Table 1. Priority Matrix

The above priority matrix, table 1, considers only three relations and suggests that child of relation agt is the leftmost element, child of ben is the middle element and child of obj is the rightmost element. Let's plan the syntax of the following UNL expression according to the rule from above table.

(Ram bought an apple for you)
{unl}
[S]
agt(buy.@entry.@past,Ram)
obj(buy.@entry.@past,apple.@def)
ben(buy.@entry.@past,you)
[/S]
{/unl}

According to above table the child of agt is 'Ram'. So, it will be the leftmost node. The child of ben is 'you' so, it will be the middle node and similarly, the child of obj, 'apples' will be the last node. So, the syntax generated will look like; *Ram(le) timi(rolaagi) syaauu kin(yo).* The morphemes, which are later generated during morphology generation phase, are shown inside "()".

Since, the syntax depends upon the sentence type, there are set of syntax planning rules as described above, for each sentence type. The sentence type is determined by checking the attributes attached to the *entry node.*

4. Morphology Generation Module

This module is responsible for proper word formation though morphology generation. This module generates most of the words. This module handles noun, verb and adjective morphology generation. This module not only inflects the root words, but also introduces conjunctions, case markers and any other new words if necessary.

The morphological rules are governed by UNL relations and attributes. Morphological rules due to UNL relations are called relation label morphology. Some relation label morphology rules have been shown in table 2. These rules introduce affixes. For example; relation ben appears in relation, suffix 'kolaagi' is added to the child. Sometimes new words are introduced. For example; if two UWs are related by relation and, new Nepali word 'ra' is introduced which has same meaning as 'and' in English.

Relation	**Definition**	**Word to be introduced**
agt	a thing that initiates an action	"le"
and	conjuctive relation between two concepts	"ra"
bas	thing used as basis for expressing degree	"bhandaa"
ben	indirectly related beneficiary	"kolaagi"
cao	thing not in focus	"sita"
con	non-focused event or state that conditions a focused event or state	"yadi"
fmt	range between two things	"samma" "dekhi"
gol	final state of an object	"laaii"
ins	instrument to carry out an event	"le"
met	means to carry out an event	"sita" "le"
opl	a place in focus affected by an event	"maa"
or	Disjunctive relation between two concepts	"athawaa"
per	basis or unit or proportion	"prati"
plc	Place where an event occurs	"maa"
pos	possessor of a thing	"ko" "kaa"
pof	concept of which a focused thing is a part	"ko
rsn	reason why an event or a state happens	"legardaa"
src	initial state of an object or an event	"baata"
tmt	Time an event ends or a state becomes false	"samma"
tmf	Time an event occurs or a state becomes true	"dekhi"
via	intermediate place or state of an event	"bhaera"

Table 2. UNL relations and Nepali affixes/words to be introduced.

UNL attributes, which expresses information like aspect, tense, number, gender, speaker's view point etc. also play an important role in morphology generation. For example, the attribute @pl means plural. When a noun has an attribute @pl, suffix 'haruu' is added to the stem (noun/pronoun). Similarly, if @not is attached to a verb, the verb needs to be negated. Suffix 'na'is added at the end of the main predicate verb to negate it.

4. Conclusion

This paper has described the development of UNL Nepali Deconverter, a Nepali language generator. Techniques of syntax planning and morphology generation have been used. Syntax planning has been done by studying the syntactic structure of the Nepali sentences. Morphology has been generated by the effect of UNL relations and attributes on Nepali word morphology. Most of the information has been generated at morphological level. The current Nepali Deconverter can deconvert moderately complex UNL expressions. Due to lack of standard UNL test data the system has yet not been formally evaluated. The size of the dictionary is small (only about 500 entries). However the size of the dictionary can be increased in the similar manner.

Nepali Deconverter can be coupled with other language Enconverter to develop a complete Machine Translation system. It can be used for future UNL Nepali viewer.

5. References

1. Dave S., Parikh J. and Bhattacharya P. (2002). Interlingua Based English Hindi Machine Translation and Language Divergence. Journal of Machine Translation, Volume 17.
2. Dhanabalan T. and Geetha T.V.(2003). UNL Deconverter For Tamil.
3. Uchida H. and Zhu M.(2001). The UNL Beyond MT. United Nations University, Tokyo.
4. UNL Center.(2000). Deconverter Specification. UNDL Foundation.
5. UNL Center.(2000). Enconverter Specification. UNDL Foundation.
6. UNL Center. (2003). UNL Specification. UNDL Foundation.
7. Uchida H. and Zhu M. (1999).A gift for a millennium . United Nations University.

About Authors

Birendra Keshari is a graduate in Computer Engineering from Kathmandu University. He is currently employed as a Teaching and Research Assistant in the Department of Computer Science and Engineering, Kathmandu University. Mr. Birendra is involved in Language Computing Research from past one and half year and is also a member of Language Processing Research Unit at Kathmandu University (www.ku.edu.np/cse/unl). He is also a member of the Nepali Language Computing Project at Madan Puraskar Pustakalaya. His general research interest is in Natural Language processing (especially Nepali Language Computing), Artificial Intelligence and Logic Programming.
E-mail : birendra@ku.edu.np

Sanat Kumar Bista, is an Assistant Professor in the Department of Computer Science and Engineering at Kathmandu University, Dhulikhel, Nepal, where he has been involved in teaching and research related to Computer Science and Information Technology. He currently leads the Language Processing Research Unit (LPRU) at Kathmandu University. He is a project leader from Kathmandu University for "Nepali Language Computing Project", being carried out in collaboration with Madan Puraskar Pustakalaya, Nepal (http://mpp.org.np) as a part of the PAN localization project(http://www.panl10n.net). Sanat's main research interests lie in the area of Localization, Multilingual Computing and Digital Libraries.
E-mail : nepal.sanat@ku.edu.np

Preprocessing Algorithms for the Recognition of Tamil Handwritten Characters

N Shanthi K Duraiswamy

Abstract

Handwriting has continued to persist as a means of communication and recording information in day-to-day life even with the introduction of new technologies. Handwriting is a skill that is personal to individuals. Recognition of characters is an important area in machine learning. Widespread acceptance of digital computers seemingly challenges the future of handwriting. However, in numerous situations, a pen together with paper or a small notepad is much more convenient than a keyboard. Handwriting data is converted to digital form either by scanning the writing on paper or by writing with a special pen on an electronic surface such as a digitizer combined with a liquid crystal display. The two approaches are distinguished as offline and online handwriting respectively. It is necessary to perform several document analysis operations prior to recognizing text in scanned documents. This paper presents detailed analysis of various preprocessing operations performed prior to recognition of Tamil handwritten characters and the results are shown.

Keywords : Indian Language, Handwriting Recognition.

0. Tamil Language

Tamil which is a south Indian language, is one of the oldest languages in the world. It has been influenced by Sanskrit to a certain degree[2]. But Tamil is unrelated to the descendents of Sanskrit such as Hindi, Bengali and Gujarati. Most Tamil letters have circular shapes partially due to the fact that they were originally carved with needles on palm leaves, a technology that favored round shapes. Tamil script is used to write the Tamil language in Tamil Nadu, SriLanka, Singapore and parts of Malaysia, as well as to write minority languages such as Badaga. Tamil alphabet consists of 12 vowels, 18 consonants and one special character (AK). Vowels and consonants are combined to form composite letters, making a total of 247 different characters and some Sanskrit characters. The complete Tamil alphabet and composite character formations are given in [5]. The advantage of having a separate symbol for each vowel in composite character formations, there is a possibility to reduce the number of symbols used by the alphabet.

1. Steps involved in Handwriting Recognition System

The major steps involved in recognition are shown in Fig.1. They are

1. Preprocessing and segmentation
2. Feature Extraction
3. Classification
4. Post processing

This paper presents about the first stage of handwriting recognition system known as preprocessing and the various steps to be performed before the recognition of Tamil handwritten characters.

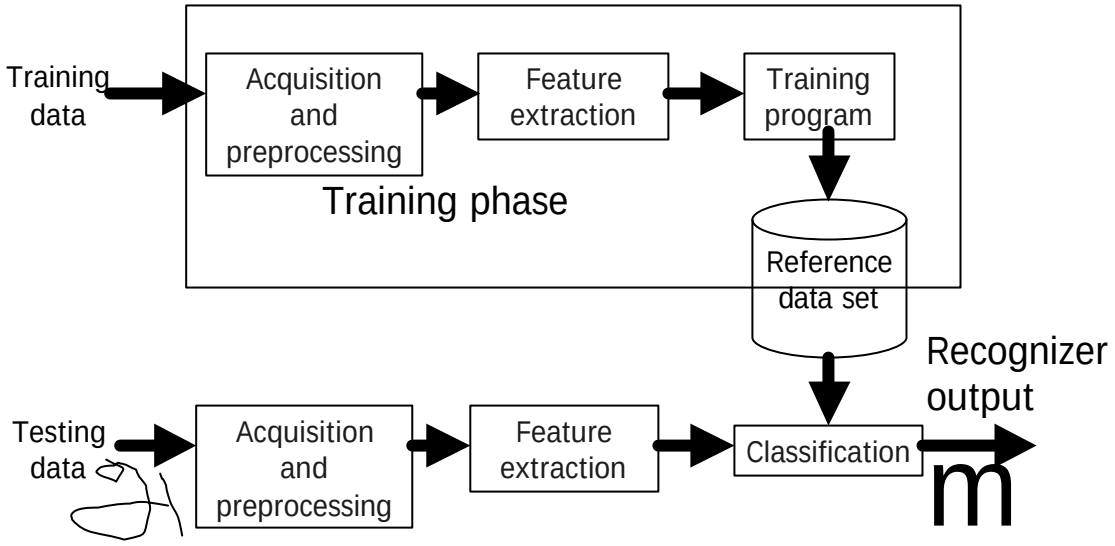


Fig.1. Steps in handwriting recognition system

2. Preprocessing

The raw input of the digitizer typically contains noise due to erratic hand movements and inaccuracies in digitization of the actual input. Original documents are often dirty due to smearing and smudging of text and aging [1]. In some cases, the documents are of very poor quality due to seeping of ink from the other side of the page and general degradation of the paper and ink. Preprocessing is concerned mainly with the reduction of these kinds of noise and variability in the input. The number and type of preprocessing algorithms employs on the scanned image depend on many factors such as paper quality, resolution of the scanned image, the amount of skew in the image and the layout of the text. Some of the common operations performed prior to recognition are: thresholding, the task of converting a gray-scale image into a binary black-white image; skeletonization, reducing the patterns to thin line representation; line segmentation, the separation of individual lines of text; and character segmentation, the isolation of individual characters [3].

3. Thresholding

The task of thresholding is to extract the foreground from the background. A number of thresholding techniques have been previously proposed using global and local techniques. Global methods apply one threshold to the entire image while local thresholding methods apply different threshold values to different regions of the image[Leedham]. The histogram of gray scale values of a document image typically consists of two peaks: a high peak corresponding to the white background and a smaller peak corresponding to the foreground. So the task of determining the threshold gray-scale value is one of determining as optimal value in the valley between the two peaks. Here Otsu's method of histogram-based global thresholding algorithm is used and is described below [6].

4. OTSU's Method for Image Thesholding

An image is a 2D grayscale intensity function, and contains N pixels with gray levels from 1 to L. The number of pixels with gray level i is denoted f_i, giving a probability of gray level i in an image of

$p_i = f_i / N$ (1)

In the case of bi-level thresholding of an image, the pixels are divided into two classes, C_1 with gray levels [1, …, t] and C_2 with gray levels [t+1, …, L]. Then, the gray level probability distributions for the two classes are

C_1: $p_1/ù_1(t), \ldots. p_t/ù_1(t)$ and

C_2: $p_{t+1}/ù_2(t), p_{t+2}/ù_2(t),\ldots, p_L/ù_2(t),$

where $ù_1(t) = \sum_{i=1}^{t} p_i$ (2)

and

$ù_2(t) = \sum_{i=t+1}^{L} p_i$ (3)

Also, the means for classes C_1 and C_2 are

$ì_1 = \sum_{i=1}^{t} i\, p_i / ?_1(t)$ (4)

and

$ì_2 = \sum_{i=t+1}^{L} i\, p_i / ?_2(t)$ (5)

Let $ì_T$ be the mean intensity for the whole image. It is easy to show that

$ù_1ì_1 + ù_2ì_2 = ì_T$ (6)

$ù_1 + ù_2 = 1$ (7)

Using discriminant analysis, Otsu defined the between-class variance of the thresholded image as

$ó_B^2 = ù_1 (ì_1 - ì_T)^2 + ù_2 (ì_2 - ì_T)^2$ (8)

For bi-level thresholding, Otsu verified that the optimal threshold t* is chosen so that the between-class variance $ó_B^2$ is maximized; that is,

$t^* = \text{Arg} \quad \text{Max} \{ó_B^2(t)\}$ (9)

1d" t< L

5. Skeletonization

Skeletonization is the process of peeling off a pattern as many pixels as possible without affecting the general shape of the pattern. In other words, after pixels have been peeled off, the pattern should still be recognized. The skeleton hence obtained must be as thin as possible, connected and centered. When these are satisfied the algorithm must stop. A number of thinning algorithms have been proposed and are being used. Here Hilditch's algorithm is used for skeletonization [4].

Consider the following 8-neighborhood of a pixel p_1

P9	P2	P3
P8	P1	P4
P7	P6	P5

Consider a decision is to be taken whether to peel off P1 or keep it as part of the resulting skeleton. For this purpose the 8 neighbors of P1 in a clock-wise order and two functions are defined.

B(P1) = number of non-zero neighbors of P1

A(P1) = number of 0,1 patterns in the sequence P2,P3,P4,P5,P6,P7,P8,P9,P2

The algorithm consists of performing multiple passes on the pattern and on each pass, the algorithm checks all the pixels and decide to change a pixel from black to white if it satisfies the following four conditions.

2 <= B(P1) <= 6

A(P1) = 1

P2.P4.P8=0 or A(P2)!=1

P2.P4.P6=0 or A(P4)!=1

Stop when nothing changes. Hilditch's algorithm is a parallel-sequential algorithm. It is parallel because at one pass all pixels are checked at the same time and decisions are made whether to remove each of the checked pixels. It is sequential because this step just mentioned is repeated several times until no more changes are done.

6. Line Segmentation

Segmentation of handwritten text into lines, words, and characters has many sophisticated approaches. This is in contrast to the task of segmenting lines of text into words and characters, which is straight forward for machine-printed documents. It can be accomplished by examining the horizontal histogram profile.

7. Character Segmentation

Line separation is usually followed by a procedure that separates the text line into characters. Vertical histogram profile is used to separate the characters.

8. Experimentation and Results

The input image and the results of various preprocessing algorithm is shown below.

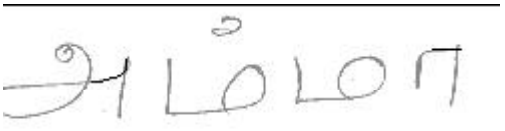


Fig.2 Original image

Fig.2 shows the original image which is used for the process of preprocessing. Data sample was collected and it was scanned using a flat-bed scanner at a resolution of 100 dpi and stored as 8-bit gray scale images.

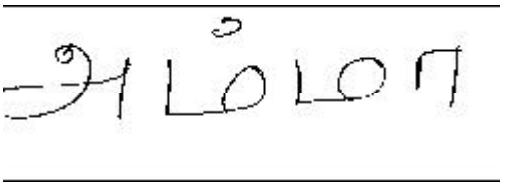


Fig.3 Thresholded image

Fig.3 shows the binary image after applying Otsu's global thresholding method to the image shown in Fig.1.

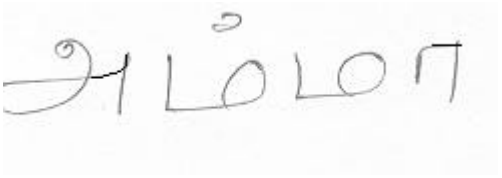


Fig.4 Skeleton of the image

Fig.4 is the skeleton of the image obtained after applying Hilditch's skeletonization algorithm.

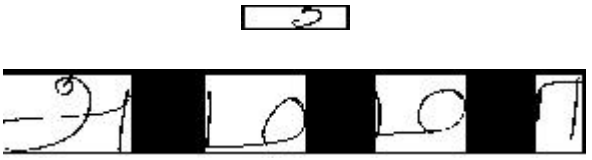


Fig.5 Segmented image

Fig.5 is the image obtained after applying the segmentation algorithm to the skeleton of the image.

9. Conclusion

This paper presents number of preprocessing algorithms that has to be performed before the process of feature extraction for Tamil handwritten character recognition system and the results are shown. The result shows that the algorithms are working reasonably well with sufficient accuracy. This work can be further extended by including few other preprocessing activities like smoothing of images, Slant correction, edge linking and size normalization. The preprocessed image can be given as input to the feature extraction phase.

10. References

1. Leedham et.al., "Comparison of some thresholding algorithms for text/background segmentation in difficult document images", ICDAR 2003.
2. S.Hewavitharana, H.C.Fernando, "A two stage classification approach to Tamil Handwriting recognition", Tamil Internet 2002, California, USA, pp.118-124
3. Srihari, "Online and Offline handwriting recognition: A comprehensive survey", IEEE PAMI, Vol.22, No.1, Jan.2000.
4. C.J.Hilditch, "Comparison of thinning algorithms on a parallel processor", Image Vision Computing, pp.115-132,1983.
5. P.Chinnuswamy and S.G.Krishnamoorthy, "Recognition of hand printed Tamil characters", Pattern recognition Vol.12, pp141-152,1980.
6. N.Otsu, "A threshold selection method from grey level histogram", IEEE Transaction. Syst. Man Cyber., vol.9 no.1, 1979, pp. 62-66.

About Authors

N Shanthi is a Assistant Professor in K. S. Rangasamy College of Technology
E-mail : shanthimoorthi@yahoo.com

Dr. K Duraiswamy is a Principal in K. S. Rangasamy College of Technology
E-mail : ksrctt@yahoo.com

Performance of Memoized- Most- Likelihood Parsing in Disambiguation Process

Maya Ingle M Chandwani

Abstract

In this paper, we first present a memoized parsing method for reducing the efforts of computation in parsing the strings/ sentences of a formal' natural language. We then discuss the statistical parsing that extracts the maximum/ most likelihood parse amongst the several parses of a string/ sentence in formal and natural domain as the most appropriate representative in disambiguation process. We integrate the statistical and memoized parsing together to achieve an efficient parsing technique. This integrated approach allows us to obtain the memoized-most-likelihood parse. Memoized-most-likelihood parse has an additional performance strength in the sense that it is highly useful further in parsing semantics.

Keywords : Natural Language Processing, Disambiguation, Statistical Parsing, Character Recognition

0. Introduction

Ambiguity and efficiency have always been the two important issues related to the parsing process as well as disambiguation process. There may exist a large number of possible derivation tree structures for a text of any language (formal or natural) and may require a large searching space. Probabilistic/ statistical techniques have been widely used to draw a maximum likelihood parse (or most likelihood parse) [1][2] whereas memoization technique in parsing allows effectively the scanning and understanding of a derivation tree structure using sub-tree criteria with a certain amount of efficiency [3][4]. The process of computing memoized-most-likelihood parse based on memoized probabilistic parsing technique has been presented in this paper. The memoized-most-likelihood parse helps us in drawing the most appropriate semantics of a formal language text (i.e. strings structure) as well as a natural language text (i.e. the sentences of English language) at the time of parsing itself in efficient manner thereby providing the ease in disambiguation process.

First, we discuss the memoization employed in parsing the strings of formal language and the sentences of natural language in Section 2. Section 3 describes the probabilistic parsing used to select the most likelihood parse amongst the several ones in formal and natural language domain both. The performance of memoized-most-likelihood parse is presented in Section 4. Finally, we conclude briefly in Section 5.

1. Memoized Parsing

Memoization is one of the techniques employed parsing algorithms to speed-up the parsing process. It reduces the re-computation efforts in producing the left parse (or right parse) of the strings (or sentences) of formal languages (or natural languages) using CKY-parsing algorithm. The effectiveness of memoization in parsing the strings of formal language contributes more significant as compared to the parsing of the natural language sentences. Memoization technique in CKY-algorithm (henceforth memoized CKY-algorithm) reduces the re-computation of sub-parses in parsing the sentences of natural language in which the repeated occurrences of same phrase structure exist. Also, it plays a significant role in parsing the strings of formal language particularly when the repeated sub-strings occur in an input string [5]. The bottom-up approach has been used to construct the recognition matrix and the top-down approach to produce the left parse of an input string in the algorithm [6].

1.1 Memoization in Formal Languages

The parses of the strings contain distinct as well as non-distinct sub-trees. We describe the performance of memoized-CKY-algorithm in brief considering both of these cases in this section.

Case I: Distinct sub-trees : The string *w*=abaabaaba of length n=9 is recognized by the grammar G_1 with *T*={*a, b*} and *N*={*S, A, B*} as shown in Table 1 (a), whose left parse tree is built-up and shown in Figure 1 using CKY-top-down parsing. Main algorithm executes the function *lookup-parse*(0, 9, *S*) such that the parse lists produced are either stored in lookup table *tab* or looked up from look-up table. The process of execution continues until all the recursive functions involved during parsing are executed and their corresponding *tab* values become non-null. Table 2 shows the actual visualization of execution of algorithm for parsing this string with parameters of current lookup-parse, status of *tab*, the current parse list at each stage and the parameters of lookup-parse functions which are used in current function recursively. The "looked-up" counts (entries in the status of tab column) compute the optimal degree of memorization which in this case is 5.

Case II: Non-distinct sub-trees : There exists three non-distinct sub-trees of order 1, 2 and 4 respectively in the parse tree of a string *w*=aaaaaaaa generated by some grammar G as shown in Figure 2. Using the algorithm, these non-distinct sub-trees are maintained separately in the look-up table as shown in Table 3 producing a parse tree with optimal degree of memoization 3.

1.2 Memoization in Natural Languages

The Memoized-CKY-algorithm performs effectively for parsing of sentences of natural language in which repeated occurrences of same phrase structure exist [7]. The effectiveness of an algorithm has been explained by considering a specific domain of sentences of a natural language grammar G_2 with production rules for parsing as follows:

Table 1(a) Grammar G_1

1SAB 2SBB 3AAB 4ACC 5Aa 6BBB 7Bb 8BCA 9Cb 10CBA 11CAA

Table 1(b) Rule probabilities of G_1

1	2	3	4	5	6	7	8	9	10	11
0.5743	0.4257	0.1062	0.2017	0.6921	0.0201	0.4848	0.4951	0.6217	0.1212	0.2571

Table 1(c) Relative probabilities of strings recognized by G_1

length of string=2

sno.	string	parses	parse list	relative prob
1	ab	1	1,5,7,	1.000000
2	bb	1	2,7,7,	1.000000

length of string=3

Sno.	String	Parses	Parse list	Relative prob
1	aba	1	1,5,8,9,5,	1.000000
2	abb	2	1,5,6,7,7, 1,3,5,7,7,	0.159507 0.840493
3	bab	1	2,8,9,5,7,	1.000000
4	bba	1	2,7,8,9,5,	1.000000
5	bbb	3	2,7,6,7,7, 2,6,7,7,7, 1,4,9,9,7,	0.041282 0.041282 0.917437

length of string=4

Sno.	String	Parses	Parse list	Relative prob
1	aaaa	1	1,5,8,11,5,5,5,	1.000000
2	aaab	1	2,8,11,5,5,5,7,	1.000000
3	aabb	1	1,4,11,5,5,9,7,	1.000000
4	abaa	1	1,5,8,10,7,5,5,	1.000000
5	abab	2	1,5,8,9,3,5,7, 1,3,5,8,9,5,7,	0.500000 0.500000
6	abba	2	1,5,6,7,8,9,5 ,5,	0.840493
7	abbb	3	1,5,6,7,6,7,7, 1,5,8,9,4,9,9, 1,3,3,5,7,7,7,	0.001825 0.947492 0.050682
8	baaa	1	2,7,8,11,5,5,5,	1.000000
9	baab	2	1,4,9,11,5,5,7, 2,8,10,7,5,5,7,	0.599351 0.400649
10	baba	1	2,8,9,5,8,9,5,	1.000000
11	babb	3	2,8,9,5,6,7,7, 2,6,8,9,5,7,7, 1,4,10,7,5,9,7,	0.188418 0.188418 0.623165
12	bbaa	1	2,7,8,10,7,5,5,	1.000000
13	bbab	3	2,7,8,9,3,5,7, 2,6,7,8,9,5,7, 1,4,9,10,7,5,7,	0.550224 0.104420 0.345356
14	bbba	3	2,7,6,7,8,9,5, 2,6,7,7,8,9,5, 1,4,9,9,8,9,5,	0.041282 0.041282 0.917437
15	bbbb	4	2,7,6,7,6,7,7, 2,7,8,9,4,9,9, 2,6,6,7,7,7,7, 1,3,4,9,9,7,7,	0.001567 0.813367 0.001567 0.183499

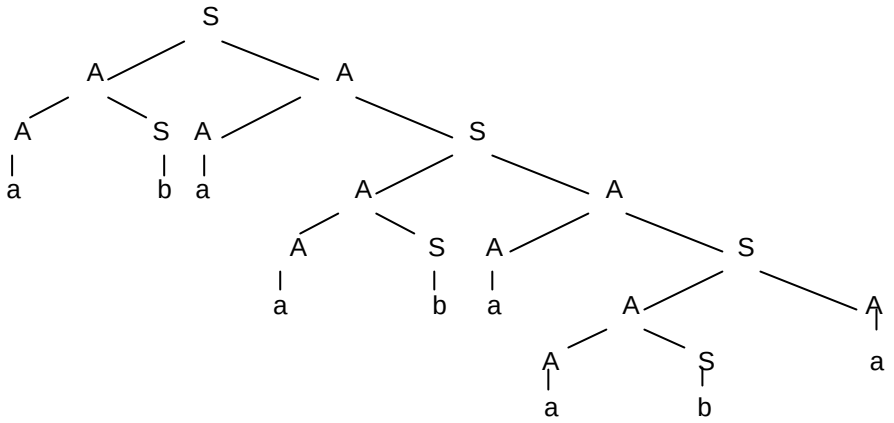


Fig. 1 Parse tree of a string w=abaabaaba showing effect of Memoization

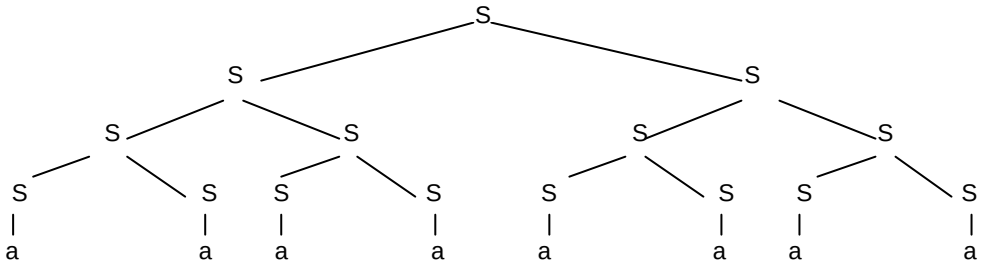


Fig. 2 Parse tree of a string w=aaaaaaaa with non-distinct terminating sub-trees

Table 2: Execution of Memoized CKY-Parsing (Using Distinct sub-trees)

Current Function	**Tab values**	**Current parse list (pt)**	**Functions used in Current Function**
(0, 9, S)	Null	{2}	(0, 2, A), (2, 9, A)
(0, 2, A)	Null	{2, 4}	(0, 1, A), (1, 2, S)
(0, 1, A)	{6}	{2, 4, 6}	—
(1, 2, S)	{3}	{2, 4, 6, 3}	—
(0, 2, A)	{4,6,3}	,,	—
(2, 9, A)	Null	{2, 4, 6, 3, 4}	(2, 3, A), (3, 9, S)
(2, 3, A)	Looked up	{2,4,6,3,4,6}	—
(3, 9, S)	Null	{2, 4, 6, 3, 4, 6, 2}	(3, 5, A), (5, 9, A)
(3, 5, A)	Looked up	{2, 4, 6, 3, 4, 6, 2, 4, 6, 3}	—
(5, 9, A)	Null	{2,4,6,3,4,6,2,4,6,3,4}	(5, 6, A), (6, 9, S)
(5, 6, A)	Looked up	{2, 4, 6, 3, 4, 6, 2, 4, 6, 3, 4, 6}	—

(6, 9, S)	Null	{2, 4, 6, 3, 4, 6, 2, 4, 6, 3, 4, 6, 2}	(6, 8, A), (8, 9, A)
(6, 8, A)	Looked up	{2, 4, 6, 3, 4, 6, 2, 4, 6, 3, 4, 6, 2, 4, 6, 3}	—
(8, 9, A)	Looked up	{2, 4, 6, 3, 4, 6, 2, 4, 6, 3, 4, 6, 2, 4,6,3,6}	—
(6, 9, S)	Non-null	,,	—
(5, 9, A)	Non-null	,,	—
(3, 9, S)	Non-null	,,	—
(2, 9, A)	Non-null	,,	—
(0, 9, S)	Non-null	,,	—

Table 3: Execution of Memoized CKY-Parsing for a string w=aaaaaaaa

(Using Non-distinct sub-trees)

Current Function	Tab Values	Functions Used in Current Function
(0, 8, S)	Null	(0, 4, S) (4, 8, S)
(0, 4, S)	Null	(0, 2, S) (2, 4, S)
(0, 2, S)	Null	(0, 1, S) (1, 2, S)
(0, 1, S)	Non-null	—-
(1, 2, S)	Looked-up*	—-
(0, 2, S)	Non-null	—-
(2 , 4, S)	Looked-up**	—-
(0, 4, S)	Non-null	—-
(4, 8, S)	Looked-up***	—-

***** Parse tree of order 1 is looked up from look-up table.

****** Parse tree of order 2 is looked up from look-up table.

******* Parse tree of order 4 is looked up from look-up table.

Table 4: Functions with their equivalent looked-up functions

No.	Current functions	Looked-up functions
1	(2, 4, NP)	(5, 7, NP), (8, 10, NP), (11, 13, NP)
2	(2, 3, *det)	(5, 6, *det), (8, 9, *det) etc.
3	(3, 4, *n)	(6, 7, *n) (9, 10, *n) (12, 13, *n)

Grammar G_2:

1. S®NP.VP
2. S®S.PP
8. *n ®I| telescope| man| hill| garden
9. *v®saw| cut| read

3. NP ®NP.PP
4. NP ®*det.*n
5. NP ®*n
6. PP ®*prep.NP
7. VP ®*v.NP
10. *det®a| the| an
11. *prep®with| on| from

A large number of procedures have been encountered with various parameters during top-down recursive in CKY-parsing algorithm as the sentence "*I saw a man with a telescope in the garden from the hill*" contains the three prepositional phrases as "*with a telescope*", "*in the garden*" and "*from the hill*". Some of the procedures share their output at lexical level and above also, thereby avoid the re-computation efforts as shown in Table 4.

2. Statistical Parsing

The statistical and structural information may be integrated together to rank the various parses of a string of a formal language as well. To derive all possible parses of a sentence/ string is not a difficult problem but it is crucial to rank these parses according to some criteria [8]. Probabilistic or statistical parsing has been widely used to resolve the various kinds of ambiguity. The sentence "*send for us the timely news*" seems to be ambiguous when "*us*" is interpreted as United States. In these situations, it is possible to compute the probabilities of various rules of grammar and to select the most likelihood parse amongst the various parses.

There exist three types of probabilistic grammars namely probabilistic context-free grammar, probabilistic context-sensitive grammar and probabilistic transformational grammar. Further, three types of probabilistic weighting of CFG are possible, they are namely, Suppes type weighting, Salomma type weighting and probabilistic CFG with derivation weighting (d_w grammar) [9]. We have used Suppes type weighting in probabilistic context-free grammar G_1 with N={*S, A, B, C*}; T={a, b}; augmented by rule probabilities (as listed in parentheses) in a set of production rules as follows:

1. Block	**2. Block**
1. S®®AB (0.50)	*3. A®®CC (0.33)*
2. S®®BB (0.50)	*4. A®®AB (0.33)*
	5. A®®a (0.34)
3. Block	**4. Block**
6. B®®BB (0.33)	*9. C®®b (0.34)*
7. B®®b (0.34)	*10. C®®BA (0.33)*
8. B®®CA (0.33)	*11. C®®AA (0.33)*

The probability of a sentence becomes negligible as its length increases. At the same time, a sentence may possess a finite number of parses. Therefore, the relative probabilities among ambiguous derivation trees are used. This measure gives the likelihood of each derivation amongst all possible derivations. A derivation with highest relative probability denotes the most appropriate parse of a string**.** An algorithm "Estimate" computes the rule probabilities and estimates the resultant probabilities by considering all possible parses (left and right parses both) of the strings of various lengths thereby producing the most correct and maximum likelihood parse of an ambiguous string [5].

2.1 Case Studies

The grammar for formal language G_1 with its rule probabilities is listed in Table 1(a) and Table 1(b) whereas the grammar for natural language G_2 given in Table 5(a) With the help of these probabilities, the language ambiguity has been quantified and pointed out the most likelihood parse of an ambiguous sentence/ string of natural/ formal language thereby improving the performance of top-down parsing technique. We discuss these cases in the following section.

Case I : Parsing in natural language domain : Using an algorithm "*Estimate*" the unambiguous sentences possess relative probability as unity whereas for the ambiguous sentences, the relative probabilities may be less than unity. These probabilities have been shown in Table 5(b). The values of relative probabilities signify the different interpretations each. The sentence "*I saw a man with a telescope in the garden*" is recognized by the grammar but it possesses the structural ambiguity [10]. There exists five distinct parses as shown in Table 5(b). The most likelihood parse possess the highest relative probability i.e. 0.990032. A parse tree of the sentence is shown in Figure 3(a) and its interpretation is given by Venn diagram as shown in Figure 3(b). The ambiguity increases sufficiently if we augment more prepositional phrases in the above sentence.

Case II : Parsing in formal language domain : The resultant rule probabilities for each of the grammatical rules and lexical rules of various grammars computed by an algorithm "*Estimate*" uses both left well as right parses. The resultant probability represents the probability that a L.H.S non-terminal is used in formation of any string *w* recognized by grammar G. The resultant rule probabilities of various rules of grammar G_1 have been shown in Table 6. Thus, resultant probability of rule 1 in G_1 i.e. *S*®*AB* is 0.5743 meaning that the sequence *AB* forms a sentence *S* with a probability 0.5743. Similarly, Table 7 shows some of ambiguous strings of length 3, 4 and 5 recognized by the grammar G_1 with their corresponding relative probabilities. It is observed that the increase in length of string causes increase in number of parses. A string *w*= abbbb of length n=5 recognized by the grammar G_1 and has five parses with relative probabilities 0.000164, 0.448343, 0.085086, 0.023982 and 0.442425 respectively as shown in Table 7. The relative probability of first parse is all most zero whereas the second parse possesses the relative probability as the highest amongst all other parses. Thus, the probabilistic parsing method owns its usefulness not only in sentence/ string disambiguation but also in improving the performance of top-down parsing when used for obtaining the hints for reordering the rules according to the rule probabilities.

3. Memoized - Most - Likelihood Parsing

The statistical and memorized parsing may be integrated together to produce memoized likelihood parse of a string/ sentence of a formal/ natural language. The performance of such parsing has been investigated and proved to be the best in this section.

Some structures of sentences in the language or sub-strings of a string in formal language occur frequently. And out of them only few sub-structures need to be computed during parsing whereas other are looked-up thereby saving the computational efforts. Similarly, there may exist the number of parses of a string/ sentence of a language. And out of them, the most likelihood parse has to be selected as it is the best structural representation of that sentence. While parsing in natural language domain, it has been found that the structural ambiguity is possessed in the sentence "*I saw a man with a telescope in the garden*" recognized by the grammar shown in Table 5(a) having five distinct parses as shown in Table 5(b). We consider the best structural representation of this sentence as the most likelihood parse with the highest relative probability as 0.990032. Instead of top-down CKY parsing, a Memoized-CKY-parsing algorithm is used with probabilistic grammars. In the maximum-likelihood parse as shown in Figure 3(a), there exist three distinct terminating sub-parse trees for NP-phrases. Since the algorithm produces

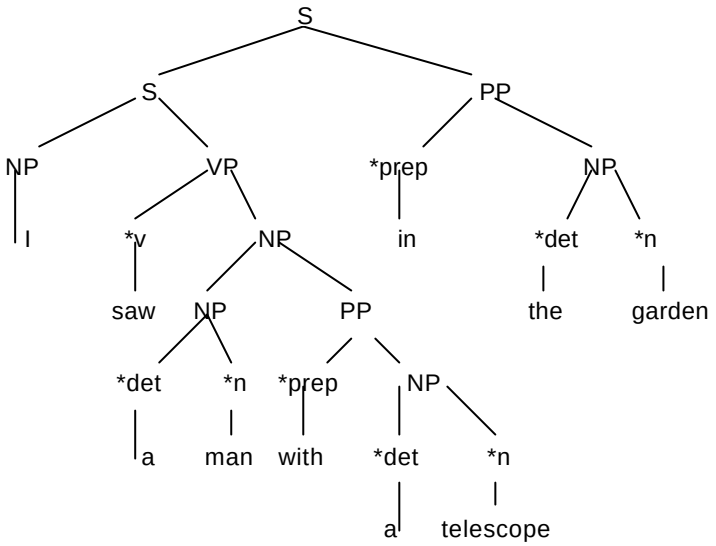
a left parse of a sentence, first NP-phrase will be computed while others are looked-up having the optimal degree of memoization thereby speeding-up the procedure of parsing with best output. Similarly, the effectiveness of Memoized-CKY-parsing using probabilistic grammars throws the attention towards the memorized-most-likelihood parse of a string *w*= abbbb recognized by the grammar G_1. Consider the second parse of this string with relative probability 0.442425 as shown in Table 7. During parsing, it stores some of the results of *lookup-parse* with specific parameters whereas some are looked up maintaining the optimal degree of memoization. . Also, it has been observed that amongst the several parses, more than one parse may have the maximum relative probability. In such a situation specially, Memoized-most-likelihood parsing technique helps us in selecting the most appropriate parse on the basis of optimal degree of memoization otherwise we choose a parse arbitrarily along with an additional performance strength. Thus, the memoized parsing in conjunction with probabilistic parsing concept establishes its improved performance in sentence/ string disambiguation in the sense that it is highly useful further in parsing semantics.

Table 5(a) Grammar G_2

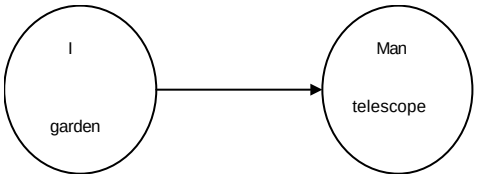
1SNPVP 2SSPP 3NPNPPP 4NP*det*n 5 NP*n 6PP*prepNP 7VP*v NP 8*nI 9*n man 10*ntelescope 11*ngarden 12*nhill 13*vsaw 14*deta 15*detthe 16*prepwith 17*prepin 18*prepon

Table 2(b) Relative probabilities of sentences recognized by Grammar G

I saw a man	
1,5,8,7,13,4,14,9,	1.000000
I saw a man with a telescope	
1,5,7,13,3,4,14,9,6,16,4,14,10,	0.316812
2,1,5,8,7,13,4,14,9,6,16,4,14,10	0.683188
I saw a man with a telescope on the hill	
1,5,7,13,3,4,14,9,6,16,3,4,14,10,6,17,4,15,11,	0.000025
2,1,5,8,7,4,13,4,14,9,6,16,3,4,14,10,6,17,4,15,11,	0.004959
2,1,5,8,7,13,3,4,14,9,6,16,4,14,10,6,17,4,15,11,	0.004959
1,5,8,7,13,3,3,4,14,9,6,16,4,14,10,6,17,4,15,11,	0.000025
2,2,1,5,8,7,13,4,14,6,16,4,14,10,6,17,4,15,11,	0.990032
I saw a man with a telescope on the hill in the garden	
1,5,8,7,13,3,4,14,9,6,16,3,4,14,10,6,17,3,4,15,11,6,18,4,15,12,	0.046872
2,1,5,8,7,13,4,14,9,6,16,3,4,14,10,6,17,3,4,15,11,6,18,4,15,12,	0.101078
2,1,5,8,7,13,3,4,14,9,6,16,4,14,10,6,17,3,4,15,11,6,18,4,15,12,	0.101078
2,1,5,8,7,13,3,4,14,9,6,16,3,4,14,10,6,17,4,15,11,6,18,4,15,12,	0.101078
1,5,8,7,13,3,3,4,14,9,6,16,4,14,10,6,17,3,4,15,11,6,18,4,15,12,	0.046872
1,5,8,7,13,3,3,4,14,9,6,16,3,4,14,10,6,17,4,15,11,6,18,4,15,12,	0.046872
2,1,5,8,7,13,4,14,9,6,16,3,3,4,14,10,6,17,4,15,11,6,18,4,15,12,	0.101078
2,2,1,5,8,7,13,3,4,14,9,6,16,4,14,10,6,17,4,15,11,6,18,4,15,12,	0.217969
2,1,5,8,7,13,3,3,4,14,9,6,16,4,14,10,6,17,4,15,11,6,18,4,15,12,	0.101078
2,2,2,1,5,8,7,13,4,14,9,6,16,4,14,10,6,17,4,15,11,6,18,4,15,12,	0.136026



(a) Most likelihood parse of "*I saw a man with a telescope in the garden*"



(b) A Venn diagram of above parse **Figure - 3**

Table 6 Rule probabilities of grammar G_1

RuleNo.	**ResultantParse**
1	0.5743
2	0.4257
3	0.1062
4	0.2017
5	0.6921
6	0.0201
7	0.4848
8	0.4951
9	0.6217
10	0.1212
11	0.2571

Table 7 Relative prob. of some ambiguous parses of strings recognized by grammar G_1

Length	String	No. of Parses	Parse list	Relative probability
3	abb	2	1,5,6,7,7	0.159507
			1,3,5,7,7	0.840493
	bbb	3	2,7,6,7,7	0.041282
			2,6,7,7,7	0.041282
			1,4,9,9,7	0.917437
4	abab	2	1,5,8,9,3,5,7	0.500000
			1,3,5,8,9,5,7	0.500000
	abbb	3	1,5,6,7,6,7,7	0.001825
			1,5,8,9,4,9,9	0.947492
			1,3,3,5,7,7,7	0.050682
	bbbb	4	2,7,6,7,6,7,7	0.001567
			2,7,8,9,4,9,9	0.813367
			2,6,6,7,7,7,7	0.001567
			1,3,4,9,9,7,7	0.183499
5	aaaab	2	1,5,8,11,5,5,3,5,7	0.503322
			1,4,11,5,5,11,5,5,7	0.496678
	abaab	3	1,5,8,9,4,11,5,5,9	0.720507
			1,3,5,8,10,7,5,5,7	0.108660
			2,8,11,3,5,7,5,5,7	0.170833
	abbab	4	1,5,6,7,8,9,3,5,7	0.022739
			1,5,8,9,4,10,7,5,9	0.333353
			1,3,3,5,7,8,9,5,7	0.119819
			2,8,11,5,4,9,9,5,7	0.524090
	abbbb	5	1,5,6,7,6,7,6,7,7	0.000164
			1,5,8,9,3,4,9,9,7	0.448343
			1,5,6,7,8,9,4,9,9	0.085086
			1,3,3,3,5,7,7,7,7	0.023982
			1,4,11,5,4,9,9,9,7	0.442425

4. Conclusion

Memoized-CKY-parsing with probabilistic parsing plays an effective role not only in disambiguation process but it also provides the most appropriate structural representation of the sentence at the same time. As compared to the performance of most-likelihood parse, definitely memoized-most-likelihood parse has an additional performance strength in the sense that it is highly useful further in parsing semantics.

5. References

1. Fujisaki, T., Jelinek, F. C., Black, E. and Niehino, T. A., “Probabilistic Parsing Method for Sentence Disambiguation”, International parsing Workshop, pp. 85-94, 1989.

2. Corazza, A., “Parsing Strategies for the Integration of Two Stochastic Context-Free Grammars”, IWPT 2003, 8 th International Workshop of Parsing Technologies, 23-25 April 2003.

3. Foth, K., Menzel, W., “Subtree Parsing to speed up Deep Analysis”, IWPT 2003, 8 th International Workshop of Parsing Technologies, 23-25 April 2003.

4. Coremen, T. H., Lieserson, C. E., Rivest, R. L., Introduction to Algorithms, Prentice Hall of India, 1999.

5. Ingle, M., Ph.D Thesis on “Computing Research Investigations in Natural Language Processing”, DAVV, Indore, Dec 2002.

6. Harrison, M. A., Introduction to Formal Language Theory, Addison-Wisley Pub., 1978.

7. Ingle, M. and Chandwani, M. “Memoized CKY-algorithm for Natural Languages,” Journal of Institute of Engineers (India), Vol. 83, pp. 12-15, 2002.

8. Charniak, E., “Statistical Parsing with a Context-free Grammar and Word Statistics”, American Association for Artificial Intelligence, pp. 589-603, 1997.

9. Klein, W. and Dittmar, N., Developing Grammars : The acquisition of German syntax by foreign workers, Springer-Verlag, Berlin H, New York, 1979.

10. Ingle, M. and Chandwani, M., “The Algorithm “Disambiguate” for Resolving Structural Ambiguity”,37 th National Convention 2002 of CSI, Harnessing and Managing Knowledge, IISC, Banglore, 2002.

About Authors

Dr. (Mrs.) M Ingle is System Analyst & Reader in School of Computer Science, D.A.V.V., Indore, Madhya Pradesh.
E-mail : maya_ingle@rediffmail.com

Dr. M Chandwani is a Director in Institute of Engineering and Technology, D.A.V.V., Indore, Madhya Pradesh.
E-mail : chandwanim@rediffmail.com

A New Contour Based Invariant Feature Extraction Approach for the Recognition of Multi-lingual Documents

Manjunath Aradhya V N Hemantha Kumar G Shivakumara P Noushath S

Abstract

Now a day, developing a single OCR system for recognizing multi-lingual documents becomes essential to enhance the ability and performance of the existing document analysis system. Hence in this paper, we present a new technique based on contour detection and distance measure for recognizing multi-lingual characters comprising south Indian languages (Kannada, Tamil, Telugu, Malayalam, English Upper case, English Lower case, English Numerals and Persian Alphanumeric). Proposed method finds boundary for a character using contour detection and the result of contour detection is given to feature extraction scheme to obtain distinct and invariant features for identifying different characters of different languages. The method extracts invariant features by computing distance between the centroid and the pixels of contour of character image.

We compare the experimental results of proposed method with result of existing methods to evaluate the performance of the method. Based on experimental results it is realized that the proposed method gives 100% accuracy with minimum expense and time. In addition, the method is invariant to Rotation, Scaling and Translation transformations (RST).

Keywords : Contour detection, Distance Measure, Invariant features, Character recognition, OCR

0. Introduction

In some situations like border places of state or countries and places where the different people meet together, the document may contain different languages. To understand such documents, there are methods in literature called hybrid OCR system. To build hybrid OCR we need to have different OCR systems for different languages. This is time consuming and it is not economically feasible. In addition to this, the system suffers from the following drawbacks. The methods fail to segment the words from the line containing different words of different languages. The method also fails to segment the different characters of different languages present in a single word. However, this kind of document is obvious in the place of railway station where we use the reservation form to reserve the seat, advertisements and any label of the product released by company.

Hence there is a necessity of developing novel technique to meet the above requirements. Therefore in this paper we present novel concept of single OCR for recognizing the multi-lingual documents. The proposed method involves only feature extraction scheme, which identifies the language through character recognition. The advantage of this concept is that it requires less computations, time and complexity compared to hybrid OCR system where it has different schemes for recognizing different languages and segmentation algorithm for segmenting the languages from single document.

1. Related Literature

In this section, we give the related literature for recognizing the characters of different languages.

(Pal .U and Chaudhuri. B.B, 2004) have proposed a review of the OCR work done on Indian Language Scripts. They discussed different methodologies applied in OCR development in International and national scenario. However, in this paper they have not addressed the problem of Indian languages like Kannada, Tamil etc.

(Pal .U and Chaudhuri. B.B, 2002) have proposed technique to identify different script lines from multi-script documents. In this paper, they have addressed the problem of development of an automatic technique for the identification of printed Roman, Chinese, Arabic, Devangari and Bangla text lines from a single document. However, the method works only at text line level but not word level and character level. In addition method fails to identify the text lines of south Indian language documents since the structure of the ext lines is almost similar.

(Pal .U and Chaudhuri. B.B, 2001) have proposed method to identify the machine printed and hand written text lines in the single document. In this paper, they have presented a machine-printed and hand-written text classification scheme for Bangla and Devangari, the two most popular Indian scripts. However, the method works for only two languages and the method fails to identify the south Indian languages.

(Pal .U et al, 2003) have introduced water reservoir concept to segment the touching numerals. In this paper, they have developed a new technique for automation segmentation of unconstrained handwritten connected numerals. The method fails to identify the characters as the number of character increase. In addition, the method has given accuracy about 94.8%.

(Chew Lim Tan et al, 2002) have proposed a method for image document text without OCR. Documents are segmented into character objects. Image features namely, the Vertical Traverse Density (VTD) and Horizontal Traverse Density (HTD), are extracted. An n-gram based document vector is constructed for each document based on these features. The method is language independent. The method works particularly if document images are of similar fonts and resolution such as in a corpus of newspaper.

(Pal .U et al, 2000) have proposed a technique to deal with an OCR (Optical Character recognition) error detection and correction technique for a highly inflectional language, Bangla, the second-most popular language in India and fifth-most popular in world. The technique is based on morphological parsing. The method is limited to only Bangla characters.

(Nagabhushan. P and Radhika M. Pai, 1999) have proposed modified region decomposition method and optimal depth tree in the recognition of non-uniform sized characters. However, the method is limited to Kannada characters only. In addition, the method is found to be computationally expensive.

(Masayoshi Okamoto and Kazuhiko Yamamoto, 1999) have proposed an on-line character recognition method that simultaneously uses both directional features, otherwise Known as off-line features, and direction-change features, which designed as on-line features. The method works for online character recognition. The method fails for Kannada characters.

(Anil K. Jain et al, 1995) have given survey on Feature extraction methods for character recognition. They have given an overview of feature extraction methods for off-line recognition of segmented (isolated) characters. We have found that no algorithms are reported in the paper to recognize the characters of south Indian languages.

(Rejean Plamondon and Sargur N. Srihari, 2000) have given survey on online and offline hand written recognition. They have described the nature of handwritten languages, how it is transduced into electronic data, and the basic concept behind written language recognition algorithms. The method works for only English characters.

(Hemantha Kumar et al, 2004) proposed a method based on construction of concentric rings for the recognition of Malayalam characters. This method draws concentric circles on the character and extracts feature values such as number of black and white pixels from each of the rings drawn. The method is invariant to rotation but variant to scaling. Further, the method requires more than 6 features to recognize the characters of different languages. Hence the method becomes computationally expensive.

(Hemantha Kumar et al, 2003) proposed a method based on distance measure and directional codes for the recognition of alphanumeric characters. The method works based on Euclidean distance measure and City-block distance measure for both thick and thinned characters. However, the method is invariant to rotation of 45-degree inclinations but it is variant to scaling. Further, the method is said to computationally expensive since the method involves two features.

Further (Hemantha Kumar et al, 2004) proposed method based on Polar Transformation. This method maps the spatial coordinate of the image to polar coordinates of polar domain. The features are extracted by counting Number of black and white pixels in each ring. The method is invariant to RST Transformation but it has less accuracy.

From the above discussion, it is revealed that to the best of our knowledge no single algorithm is reported for the recognition of characters of multi-languages.

In this paper, we introduced novel concept of single OCR system based on contour detection and distance measure to recognize the multi-lingual documents. The proposed method has three stages. In first stage, we develop an algorithm to obtain contour (boundary) of a character image. New invariant feature extraction scheme for identifying the different contours of different languages results in second stage. The contour is obtained by performing the mask designed. The features are extracted for each contour by computing distance between the centroid and black pixels of the contour. The square of number of black pixels in the contour divided by the sum of all distances gives invariant feature for each contour. In the last stage, we have designed database using linear search tree and binary search tree to study the performance of the proposed method.

The rest of the paper is organized as follows: Section 3 discusses the proposed methodologies. Section 4 presents the experimental results. Section 5 and 6 presents Comparative study and Conclusions respectively.

2. Proposed Methodology

This section presents a technique for developing single recognition system for multi-lingual documents. Proposed methodology is divided into two sub sections. Algorithm for boundary detection using contour is presented in section 3.1. A new feature extraction scheme is introduced in section 3.2 to obtain invariant features for recognizing multi-lingual documents. In this work, we have considered the following data set for designing the single OCR for six languages including English upper case letters, lower case letters and numerals (ref. Fig 1 – Fig 8). In this method, we have assumed that the characters are in isolated form.

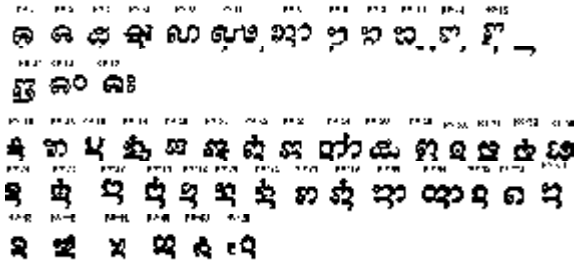


Fig. 1 Fifty Alphabets of Kannada Language

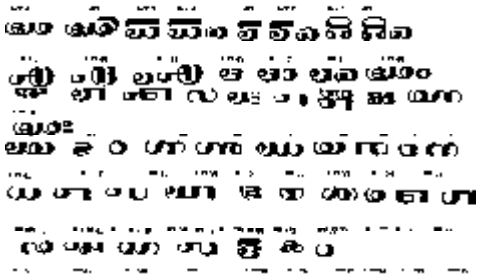


Fig. 2 Fifty-Two Alphabets of Malayalam Language

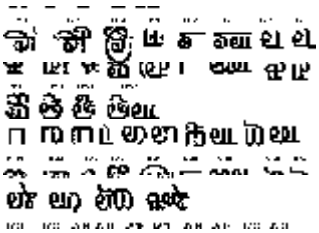


Fig. 3 Thirty-Five Alphabets of Tamil Language



Fig. 4 Fifty Alphabets of Telugu Language



Fig. 5 Twenty-Six Alphabets of English Upper Case Letters



Fig. 6 Twenty-Six Alphabets of English Lower Case Letters

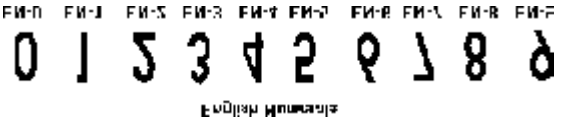


Fig. 7 Ten Numerals of English Language



Fig. 8 Forty-Two Letters of Persian Language

2.1 Contour Detection (CD)

In this sub section, we present a new preprocessing contour detection technique to obtain boundary for a character. This technique reduces the number of black pixels of width of the character boundary. Moreover it works even width of boundary of character varies. In order to obtain boundary for a character we have designed 3X3 mask, which is given in Fig. 9, where P_2, P_3, P_4, P_5, P_6, P_7, P_8 and P_9 are the eight neighbor pixels of P_1 pixel. The technique assumes that the 1 (white) represents background color and 0 (Black) represents foreground color.

P2	P3	P4
P9	P1	P5
P8	P7	P6

Fig 9 Mask used to find contour

The technique deletes border pixel when P_2,P_3, P_4, P_5,P_6, P_7, P_8 and P_9 are 0. That is the center pixel (P_1) changes into 1 when the above condition is satisfied. This condition is satisfied not at the inside and outer boundary of the character image. Hence it removes pixel, which are present inside the character image. This procedure is repeated until no further changes in the boundary of a character. If the condition is not satisfied then the mask move to next pixel of the character image. The result of the technique is shown in Fig. 10.

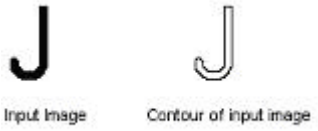


Fig 10 Result of contour detection

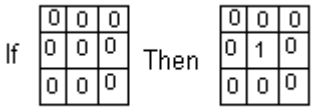
Algorithm: Contour Detection (CD)

Input: Character image

Output: Contour Image

Method Begins

Step 1: For each pixel of the character image, employ the following rule.



Step 2: Repeat the procedure until no further changes in the

Method ends

2.2 Invariant Features (IF) Extraction

For any variations in the character boundary, the CD algorithm gives single pixel boundary of the character. The result of the CD taken as input for feature extraction approach is presented in this section. The method computes centroid of the character by finding X_{min}, X_{max} and Y_{min}, Y_{max} coordinates of the contour. Next method estimates the distance (D) between the centroid (C) and the pixel of the contour. Further, the method finds Sum of all Distances (SD) and number of black pixels (N). The ratio square of N to SD is feature which is invariant to image transformation such as rotation, Scaling and Translation (RST). The steps involved in algorithm are given in Fig. 11.

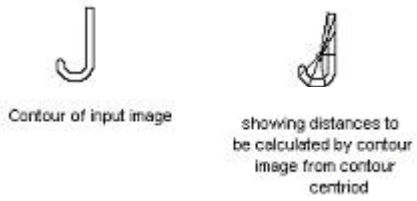


Fig 11 Feature extraction procedure

Algorithm: Invariant Feature (IF) Extraction

Input: Contour Image

Output: Invariant Features

Method Begins

Step 1: Find out the centriod (C) of the contour image.

Step 1.1: Sum (Sx) of all X coordinates is $Sx = \sum_{i=1}^{n} Xi$ where n is the number of black pixels.

Step 1.2: Sum (Sy) of all Y coordinates is $Sy = \sum_{i=1}^{n} Yi$

Step 1.3: X coordinate of the C is $= Cx = \frac{Sx}{N}$, where N is the total number of black pixels

Step 1.4: Y coordinate of the C is $Cy = \frac{Sy}{N}$.

Step 2: Find out the D from C to every black pixel using Euclidean Distance (ED).

$EDi = \sqrt{(Cx - Xi)^2 + (Cy - Yi)^2}$ For i = 1 to n, where Xi and Yi are the coordinates of black pixels.

Step 3: $SD = \sum_{i=1}^{n} EDi$

Step 5: $IF = \frac{N^2}{SD}$

Method ends

3. Experimental Results

In this section, we present the experimental results for evaluating the efficiency of the proposed contour based method. We have considered accuracy and computations as decision parameters to establish superiority of the proposed method. Accuracy of the method depends on number of characters recognized correctly out of 291 characters data set. Computations depend on the dominant operation involved in the method. In this method the dominant operation is pixel searching. In order to know the accuracy and recognition rate we have designed database using linear search tree and binary search tree [Jean. Paul. Tremblay and Paul. G. Sorenson, 1988] we have given comparative study in next section. Further, we have also shown that the proposed method is invariant to RST. 1.4 G Hz processor system is used for the experimentation purpose in this work. The values of accuracy and computations of the proposed method are tabulated in Table 1.

From Table 1 it is clear that the proposed method gives 100% accuracy and it takes 58201 computations for all 291 characters. We have experimented the proposed method to show that the method is invariant to rotations of any degree. In Fig. 12, we have given some samples of rotations of character image to tell that the method gives same feature values for different rotations. The corresponding features are tabulated in Table 2. From Table 2 and Fig. 13 it is clear that the proposed method is invariant to rotation transformation. Similarly for scaling also we have experimented the proposed method for different resolutions, which is given in Fig. 14. The corresponding values are tabulated in Table 3. From Table 3 and Fig. 15 it is observed that the method is invariant to scaling after 100dpi. In addition to this, we also concluded that the performance of the method degrades when the low-resolution image is given.

Table 1 Accuracy and computations of proposed method for 291 characters.

Contour Based Method	
Accuracy	Computation
100%	58201

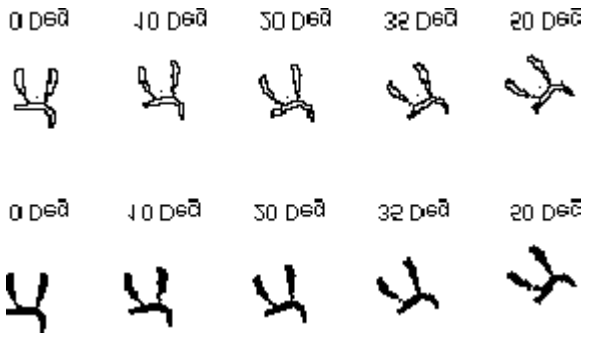


Fig 12. Different rotations of character image

Table 2 Features for different rotations

Degree	F
0	14.69
10	14.67
20	14.70
35	14.628
50	14.721

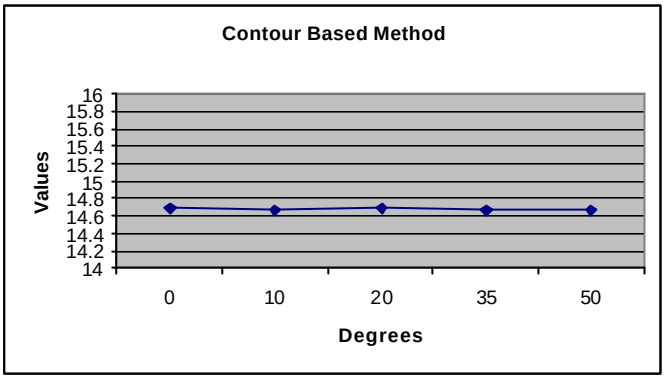


Fig 13 Graph Features v/s degrees

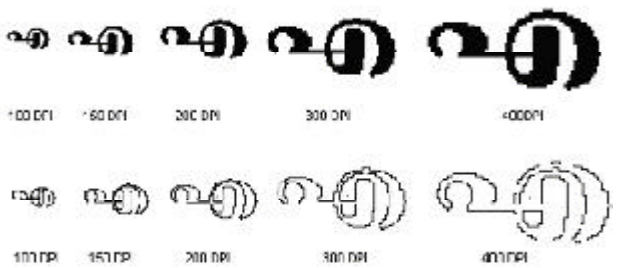


Fig 14 Different resolutions of character image

Table 3 Features for different dpi

Resolution	F
100	20.51
150	22.56
200	22.01
300	22.10
400	22.013

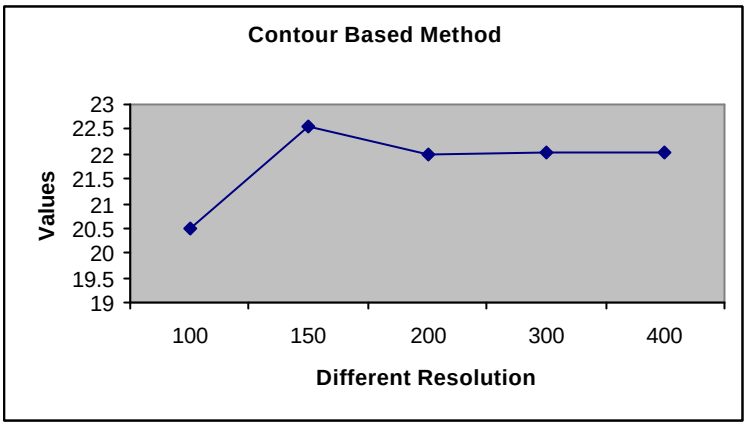


Fig 15 Graph Features v/s dpi

4. Comparative Study

In this section, we have given comparative study of proposed method with [Hemantha Kumar et al., 2004, Hemantha Kumar et al., 2004, Hemantha Kumar et al., 2003] method to evaluate the performance of the proposed method in terms of accuracy in recognition and number of computations. In order to compare the methods we have chosen parameters such as Time required to recognize the 291 characters using Linear Search Tree (TLST), Time using Binary Search Tree (TBST), computation involved in searching expected features in the database of 291 characters. In addition to this we have also consider the invariance property as parameter apart from the accuracy in recognition and number of computations.

From Table 4, it is noticed that the proposed method gives 100% accuracy and takes less computations (ref. Fig. 16 and Fig. 17) when compared to Polar Transformation Method (PTM) and Ring Projection Method (RPM) since the proposed method involves contour detection, which reduces the number of black pixels. However, Distance measure Method (DM) takes less computation compared to proposed method since DM involves only eight directions. The proposed method is competitive with respect to time parameter (ref. Fig. 18 and Fig 19) compared to RPM, PTM and DM in case of both linear and binary search trees. This is because the PTM involves feature vector containing 15 features, RPM involves feature vector containing 6 features and DM involves feature vector containing 2 features. As the parameter computations is concerned, the number of computations required to search feature using linear search tree and binary search tree, the proposed method gives better results than existing methods (Ref. Fig. 20 and Fig. 21). This is because of the feature vector having number of features.

Table 4 Values of the parameters based on experimental results

Name of the Methods	Linear Search		Binary Search		Accuracy in	Computations recognition
	Time	Computa-tions to search a FV	Time	Computa-tions to search a FV		
PTM	5.05 m	120539475	4.16 m	52890275	67%	81480
DM	2.07	24726852	1.53	6547500	100%	37910
RPM	4.54	74180556	3.54	19642500	100%	4431930
Contour Based Method (proposed method)	2.20 m	12363426	1.60 m	3273750	100%	58201

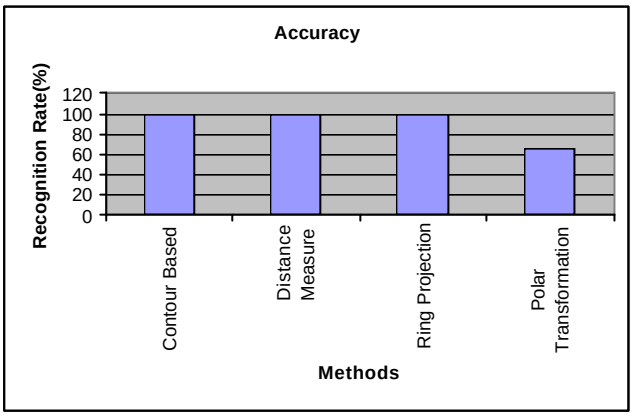


Fig 16 Graph for Accuracy v/s methods

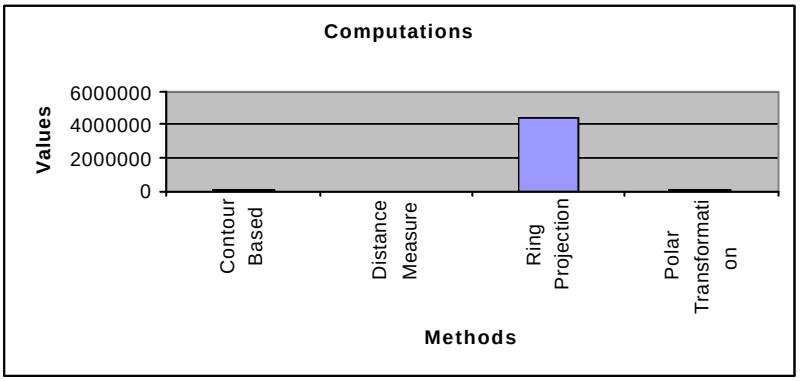


Fig 17 Graph for computations v/s methods

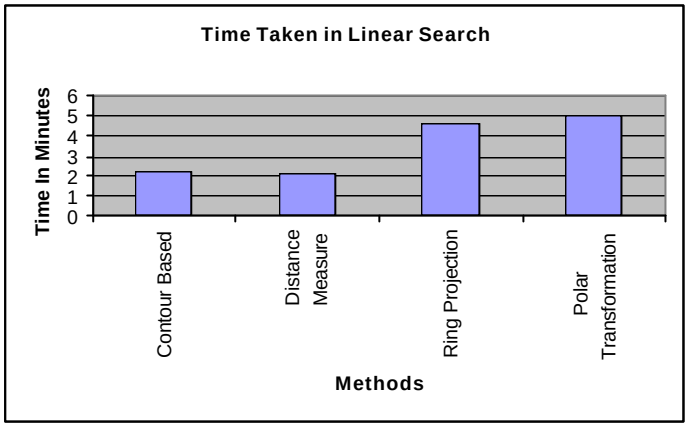


Fig 18 Graph for Time LST v/s methods

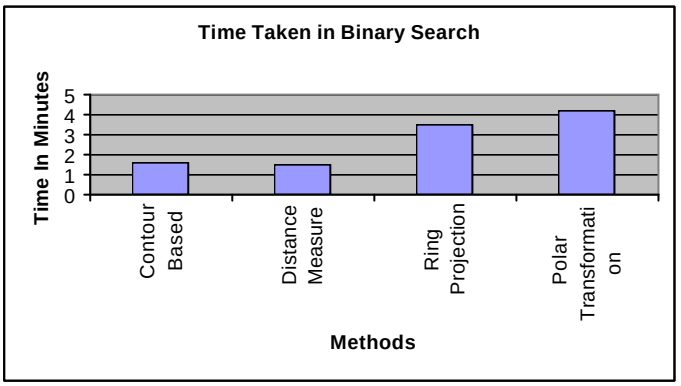


Fig 19Graph for Time BST v/s methods

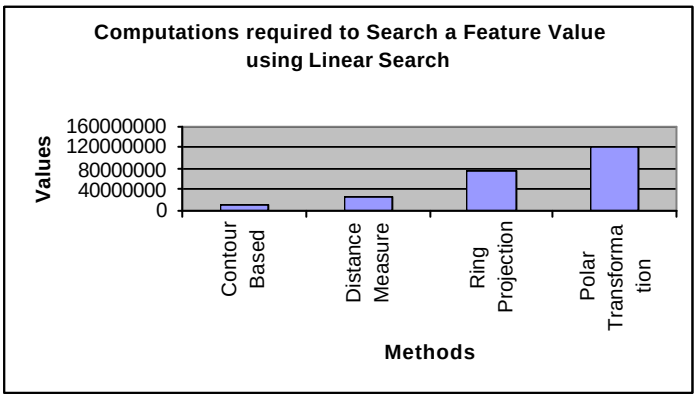


Fig 20 Graph for computations in linear Search tree v/s methods

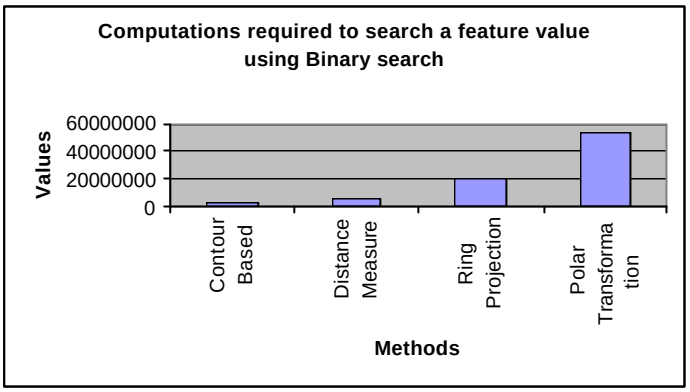


Fig 21 Graph for computations in Binary Search tree v/s methods

Table 5 Overall performances of the proposed and existing methods

Parameters	Name of the Methods
Accuracy	Contour based, Distance Measure, Ring Projection
Computations	Distance Measure
Number of computations involved in linear Search for a Feature Value (LS)	Contour based method
Number of computations involved in Binary Search for a Feature Value (LS)	Contour based method
RST Transformation	Contour based and Polar Transformation method

5. Conclusion

We have presented a new contour based method for the recognition of multi-lingual documents. The Proposed method is compared with the methods based on PTM [Hemantha Kumar .G et al., 2004], DM [Hemantha Kumar G et al., 2003] and RPM [Hemantha Kumar et al., 2004]. We have shown that the proposed method is better compared to other methods in terms of accuracy, computations, time required for searching a character and invariance property (ref. Table 5). However, the performance of the proposed method degrades for low-resolution images. This is an attempt to develop single OCR for all these languages. Further, the method is extended to some other languages also. This would be our future work.

6. Acknowledgment

The authors acknowledge the support extended by Dr. D. S Ramakrishna, Principal, APS College of Engineering, Somanahalli, Bangalore-82.

7. References

1. Anil. K Jain, Oivind Due Trier, and Torfinn Taxt, Feature Extraction Methods for Character Recognition – A Survey, Pattern Recognition, Vol.29, No.4, pp641-662, 1996.
2. Chew Lim Tan, Weihua Huang, Zhaohui Yu and Yi Xu, Imaged Document Text Retrieval Without OCR, IEEE transaction on Pattern Analysis and Machine Intelligence, Vol.24, No.6, 2002.
3. Hemantha Kumar. G, Shivakumara. P, Noushath. S, and Manjunath Aradhya. V.N, A New Invariant Algorithm for Recognition of Alphabets of Multi-Lingual Documents, Proceedings of 6 Th International Conference on Cognitive Systems – ICCS 2004, Centre for Research in Cognitive Systems, New Delhi, India, December 14-15, 2004(Accepted).
4. Hemantha Kumar. G, Shivakumara. P, Noushath. S, and Manjunath Aradhya. V.N, A Novel Feature Extraction Scheme for Malayalam Character Recognition, Journal of the Society of Statistics, Computer and Applications, Vol. 2, No. 1, 2004(New Series), pp 101-113.
5. Hemantha Kumar. G, Shivakumara. P, Noushath. S, and Manjunath Aradhya. V.N, Feature Extraction for Alphanumeric Symbols Recognition: An Approach Based on Distance Measures, Proceedings of I st Indian International Conference on Artificial Intelligence (IICAI-03), Hyderabad, India, December 18-20 2003.
6. Jean. Paul Tremblay and Paul. G. Sorenson, An Introduction to data Structures with Applications, Mc Graw –Hill Book Company, 1988.
7. Masayoshi Okamoto and Kazuhiko Yamamoto, On-line handwriting character recognition using direction-change features that consider imaginary strokes, The Journal of Pattern Recognition Society, Vol.32, pp1115-1128, 1999.
8. Nagabhushan. P and Radhika.M.Pai, Modified region decomposition method and optimal depth decision tree in the recognition of non-uniform sized characters – An experimentation with Kannada characters, The Journal of Pattern Recognition Society, Vol.20, pp1467-1475, 1999.
9. Pal. U and Chaudhuri. B.B, Identification of different script lines from multi-script documents, Image and Vision Computing, Vol.20, pp945-954, 2002.
10. Pal. U and Chaudhuri. B.B, Machine-printed and Hand-written text lines identification, Pattern Recognition Letters, Vol.22, pp431-441, 2001.
11. Pal. U, Belaid. A and Choisy. Ch, Touching numeral segmentation using water reservoir concept, Pattern Recognition Letters, Vol.24, pp261-272, 2003.
12. Pal. U, Chaudhuri .B.B, Indian script character recognition: a survey, Pattern Recognition, Vol.37, pp 1887-1899, 2004.
13. Pal. U, Kundu. P. K, and Chaudhuri B. B, OCR Error Correction of an Inflectional Indian Language using Morphological Parsing, Journal of Information Science and Engineering, Vol.16, pp903-922, 2000.
14. Rejean Plamondon and Sargur N.Srihari, On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey, IEEE transactions on Pattern Analysis and Machine Intelligence, Vol.22, No.1, Jan-2000.

About Authors

Manjunath Aradhya V N, Department of Studies in Computer Science, Manasagangothri, University of Mysore, Mysore-6.
E-mail : mukesh_mysore@rediffmail.com

Hemantha Kumar G, Department of Studies in Computer Science, Manasagangothri, University of Mysore, Mysore-6.
E-mail : mukesh_mysore@rediffmail.com

Shivakumara P, Assistant Professor, Department of Computer Science and Engineering, Acharya Patasala College of Engineering, Kanakapura Road, Somanahalli, Bangalore – 62.
E-mail : hudempsk@yahoo.com

Noushath S, Department of Studies in Computer Science, Manasagangothri, University of Mysore, Mysore - 6.
E-mail : mukesh_mysore@rediffmail.com

Current Status & Process in the Development of Applications Through NLP

V R Rathod S M Shah Nileshkumar K Modi

Abstract

The development of natural language processing systems has resulted in their being increasingly used in support of other computer programs. This trend is particularly noticeable with regard to information management applications. Natural language processing provides a potential means of gaining access to the information inherent in the large amount of text and available through the Internet. In the following survey, we look in further details at the recent trends in research in natural language processing and conclude with a discussion of some applications of this research to the solution of information management problems.

Keywords : Natural Language Processing.

0. Introduction

Work in computational linguistics began very soon after the development of the first computers, yet in the intervening four decades there has been a pervasive feeling that progress in computer understanding of natural language has not been commensurate with progress in other computer applications. Recently, a number of prominent researchers in natural language processing met to assess the state of the discipline and discuss future directions. The consensus of this meeting was that increased attention to large amounts of lexical and domain knowledge was essential for significant progress, and current research efforts in the field reflect this point of view.

1. Passive Voice and Its Usage

The traditional approach in computational linguistics included a prominent concentration on the formal mechanisms available for processing language, especially as these applied to syntactic processing and, somewhat less so, to semantic interpretation. In recent efforts, work in these areas continues, but there has been a marked trend toward enhancing these core resources with statistical knowledge acquisition techniques. There is considerable research aimed at using online resources for assembling large knowledge bases, drawing on both natural language corpora and dictionaries and other structured resources. Recent research in lexical semantics reflects an interest in the proper structuring of this information to support linguistic processing.

Furthermore, the availability of large amounts of machine-readable text naturally supports continued work in analysis of connected discourse. In other trends the use of statistical technique are being used as part of the parsing process, for automatic part of speech assignment, and for word sense disambiguation.

2. The Lexicon

In computational linguistics the lexicon supplies paradigmatic information about words, including part of speech labels, irregular plurals, and sub categorization information for verbs. Traditionally, lexicons were quite small and were constructed largely by hand. There is a growing realization that effective natural language processing requires increased amounts of lexical (especially semantic) information. A recent trend has been the use of automatic techniques applied to large corpora for the purpose of acquiring lexical information from text. Statistical techniques are an important aspect of automatically mining lexical

information. Manning (1993) uses such techniques to gather sub categorization information for verbs. Brent (1993) also discovers Sub categorization information; in addition he attempts to automatically discover verbs in the text. Liu and Soo (1993) describe a method for mining information about thematic roles.

The additional information being added to the lexicon increases the complexity of the lexicon. This added complexity requires that attention be paid to the organization of the lexicon: Zernik 1991 (Part III) and Pustejovsky 1993 (Part III) both contain several papers which address this issue. McCray, Srinivasan and Browne(1993) discuss the structure of a large (more than 60,000 base forms) lexicon designed and implemented to support syntactic processing.

3. Automatic Tagging

Automatically disambiguating part-of-speech labels in text is an important research area since such ambiguity is particularly prevalent in English. Programs resolving part-of-speech labels (often called automatic taggers) typically are around 95% accurate. Taggers can serve as preprocessors for syntactic parsers and contribute significantly to efficiency. There have been two main approaches to automatic tagging: probabilistic and rule-based. Merialdo (1994) and Dermatos and Kokkinakis (1995) review several approaches to probabilistic tagging and then offer new proposals. Typically, probabilistic taggers are trained on disambiguated text and vary as to how much training text is needed and how much human effort is required in the training process. (See 3 Schütze 1993 for a tagger that requires very little human intervention.) Further variation concerns knowing what to do about unknown words and the ability to deal with large numbers of tags.

One drawback to stochastic taggers is that they are very large programs requiring considerable computational resources. Brill (1992) describes a rule-based tagger which is as accurate as stochastic taggers, but with a much smaller program. The program is slower than stochastic taggers, however. Building on Brill's approach, Roche and Schabes (1995) propose a rule-based, finite-state tagger which is much smaller and faster than stochastic implementations. Accuracy and other characteristics remain comparable.

4. Parsing

The traditional approach to natural language processing takes as its basic assumption that a system must assign a complete constituent analysis to every sentence it encounters. The methods used to attempt this are drawn from mathematics, with context-free grammars playing a large role in assigning syntactic constituent structure. Partee, ter Meulen and Wall (1993) provide an accessible introduction to the theoretical constructs underlying this approach, including set theory, logic, formal language theory, and automata theory, along with the application of these mechanisms to the syntax and semantics of natural language.

The program described in Alshawi 1992 is a very good example of a complete system-built on these principles. For syntax, it uses a unification-based implementation of a generalized phrase structure grammar (Gazdar *et al.* 1985) and handles an impressive number of syntactic structures which might be expected to appear in "interactive dialogues with information systems… although of course there is still a large residue even of this variety of English that the system fails to analyze properly." (Alshawi 1992:61).

In continuing research in this tradition, context-free grammars have been extended in various ways. The so-called "mildly context sensitive grammars," such as tree adjoining grammars, have had considerable influence on recent work concerned with the formal aspects of parsing natural language.

Several recent papers pursue nontraditional approaches to syntactic analysis. One such technique is partial, or underspecified, analysis. For many applications such an analysis is entirely sufficient and can often be more reliably produced than a fully specified structure. Chen and Chen (1994), for example, employ statistical methods combined with a finite state mechanism to impose an analysis which consists only of noun phrase boundaries, without specifying their complete internal structure or their exact place in a complete tree structure. Agarwal and Boggess (1992) successfully rely on semantic features in a partially specified syntactic representation for the identification of coordinate structures. In an innovative application of dependency grammar and dynamic programming techniques, Kurohashi and Nagao (1994) address the problem of analyzing very complicated coordinate structures in Japanese.

A recent innovation in syntactic processing has been investigation into the use of statistical techniques. (See Charniak 1993 for an overview of this and other statistical applications.) In probabilistic parsing, probabilities are extracted from a parsed corpus for the purpose of choosing the most likely rule when more than one rule can apply during the course of a parse (Magerman and Weir 1992). In another application of probabilistic parsing the goal is to choose the (semantically) best analysis from a number of syntactically correct analyses for a given input (Briscoe and Carroll 1993, Black, Garside and Leech 1993).

A more ambitious application of statistical methodologies to the parsing process is grammar induction where the rules themselves are automatically inferred from a bracketed text; however, results in the general case are still preliminary. Pereira and Schabes (1992) discuss inferring a grammar from bracketed text relying heavily on statistical techniques, while Brill (1993) uses only modest statistics in his rule-based method.

5. Word-Sense Disambiguation

Automatic word-sense disambiguation depends on the linguistic context encountered during processing. McRoy (1992) appeals to a variety of cues while parsing, including morphology, collocations, semantic context, and discourse. Her approach is not based on statistical methods, but rather is symbolic and knowledge intensive. Statistical methods exploit the distributional characteristics of words in large texts and require training, which can come from several sources, including human intervention. Gale, Church and Yarowsky (1992) give an overview of several statistical techniques they have used for word-sense disambiguation and discuss research on evaluating results for their systems and others. They have used two training techniques, one based on a bilingual corpus, and another on *Roget's Thesaurus.* Justeson and Katz (1995) use both rule based and statistical methods. The attractiveness of their method is that the rules they use provide linguistic motivation.

6. Semantics

Formal semantics is rooted in the philosophy of language and has as its goal a complete and rigorous description of the meaning of sentences in natural language. It concentrates on the structural aspects of meaning. Chierchia and McConnell-Ginet (1990) provide a good introduction to formal semantics. The papers in Rosner and Johnson 1992 discuss various aspects of the use of formal semantics in computational linguistics and focus on Montague grammar (Montague 1974), although Wilks (1992) dissents from the prevailing view. King (1992) provides an overview of the relation between formal semantics and computational linguistics. Several papers in Rosner and Johnson discuss research in the situation semantics paradigm (Barwise and Perry 1983), which has recently had wide influence in computational linguistics, especially in discourse processing. See Alshawi 1992 for a good example of an implemented (and eclectic) approach to semantic interpretation.

Lexical semantics (Cruse 1986) has recently become increasingly important in natural language processing. This approach to semantics is concerned with psychological facts associated with the meaning of words. Levin (1993) analyzes verb classes within this framework, while the papers in Levin and Pinker 1991 explore additional phenomena, including the semantics of events and verb argument structure. A very interesting application of lexical semantics is WordNet 5 (Miller 1990), which is a lexical database that attempts to model cognitive processes. The articles in Saint-Dizier and Viegas 1995 discuss psychological and foundational issues in lexical semantics as well as a number of aspects of using lexical semantics in computational linguistics.

Another approach to language analysis based on psychological considerations is cognitive grammar (Langacker 1988). Olivier and Tsujii (1994) deal with spatial prepositions in this framework, while Davenport and Heinze (1995) discuss more general aspects of semantic processing based on cognitive grammar.

7. Discourse Analysis

Discourse analysis is concerned with coherent processing of text segments larger than the sentence and assumes that this requires something more than just the interpretation of the individual sentences. Grosz, Joshi and Weinstein (1995) provide a broad-based discussion of the nature of discourse, clarifying what is involved beyond the sentence level, and how the syntax and semantics of the sentences support the structure of the discourse. In their analysis, discourse contains linguistic structure (syntax, semantics), attentional structure (focus of attention), and intentional structure (plan of participants) and is structured into coherent segments. During discourse processing one important task for the hearer is to identify the referents of noun phrases. Inferencing is required for this identification. A coherent discourse lessens the amount of inferencing required of the hearer for comprehension. Throughout a discourse the particular way that the speaker maintains “focus of attention” or “centering” through choice of linguistic structures for referring expressions is particularly relevant to discourse coherence.

Other work in computational approaches to discourse analysis has focused on particular aspects of processing coherent text. Hajicova, Skoumalova and Sgall (1995) distinguish topic (old information) from focus (new information) within a sentence. Information of this sort is relevant to tracking focus of attention. Lappin and Leass (1994) are primarily concerned with intrasentential anaphora resolution, which relies on syntactic, rather than discourse, cues. However, they also address intersentential anaphora, and this relies on several discourse cues, such as saliency of an NP, which is straightforwardly determined by such things as grammatical role, frequency of mention, proximity, and sentence recency. Huls, Bos and Claasen (1995) use a similar notion of saliency for anaphora resolution and resolve deictic expressions with the same principles. Passonneau and Litman (1993) study the nature of discourse segments and the linguistic structures which cue them. Sonderland and Lehnert (1994) investigate machine learning techniques for discovering discourse-level semantic structure.

Several recent papers investigate those aspects of discourse processing having to do with the psychological state of the participants in a discourse, including, goals, intentions, and beliefs: Asher and Lascarides (1994) investigate a formal model for representing the intentions of the participants in a discourse and the interaction of such intentions with discourse structure and semantic content. Traum and Allen (1994) appeal to the notion of social obligation to shed light on the behavior of discourse. Wiebe (1994) investigates psychological point of view in third person narrative and provides an insightful algorithm for tracking this phenomenon in text. The point of view of each sentence is either that of the narrator or any one of the characters in the narrative.6 Wiebe discusses the importance of determining point of view for a complete understanding of a text, and discusses how this interacts with other aspects of discourse structure.

8. Applications

As natural language processing technology matures, it is increasingly being used to support other computer applications. Such use naturally falls into two areas, one in which linguistic analysis merely serves as an interface to the primary program, and another in which natural language considerations are central to the application.

Natural language interfaces to data base management systems (e.g. Bates 1989) translate users' input into a request in a formal data base query language, and the program then proceeds as it would without the use of natural language processing techniques. It is normally the case that the domain is constrained and the language of the input consists of comparatively short sentences with a constrained set of syntactic structures.

The design of question answering systems is similar to that for interfaces to data base management systems. One difference, however, is that the knowledge base supporting the question answering system does not have the structure of a data base. See, for example Kupiec 1993, where the underlying knowledge base is an on-line encyclopedia. Processing in this system not only requires a linguistic description for users' requests, but it is also necessary to provide a representation for the encyclopedia itself. As with the interface to a DBMS, the requests are likely to be short and have a constrained syntactic structure. Lauer, Peacock and Graesser (1992) provide some general considerations concerning question answering systems and describe several applications.

In message understanding systems, a fairly complete linguistic analysis may be required, but the messages are relatively short and the domain is often limited. Davenport and Heinze (1995) describe such a system in a military domain. See Chinchor, Hirschman and Lewis 1993 for an overview of some recent message understanding systems.

In three closely related applications (information filtering, text categorization, and automatic abstracting), no constraints on the linguistic structure of the documents being processed can be assumed. One mitigating factor, however, is that effective processing may not require a complete analysis. For all of these applications there are also statistically based systems based on frequency distributions of words. These systems work fairly well, but most people feel that for further improvements, and for extensions, some sort of understanding of the texts, such as that provided by linguistic analysis, is required.

Information filtering and text categorization are concerned with comparing one document to another. In both applications, natural language processing imposes a linguistic representation on each document being considered. In text categorization a collection of documents is inspected and all documents are grouped into several categories based on the characteristics of the linguistic representations of the documents. Blosseville *et al.* (1992) describe an interesting system which combines natural language processing, statistics, and an expert system. In information filtering, 7 documents satisfying some criterion are singled out from a collection. Jacobs and Rau (1990) discuss a program which imposes a quite sophisticated semantic representation for this purpose.

In automatic abstracting, a summary of each document is sought, rather than a classification of a collection. The underlying technology is similar to that used for information filtering and text categorization: the use of some sort of linguistic representation of the documents. Of the two major approaches, one (e.g. McKeown and Radev 1995) puts more emphasis on semantic analysis for this representation and the other (e.g. Paice and Jones 1993), less.

Information retrieval systems typically allow a user to retrieve documents from a large bibliographic database. During the information retrieval process a user expresses an information need through a query. The system then attempts to match this query to those documents in the database which satisfy the user's information need. In systems which use natural language processing, both query and documents are transformed into some sort of a linguistic structure, and this forms the basis of the matching. Several recent information retrieval systems employ varying levels of linguistic representation for this purpose. Sembok and van Rijsbergen (1990) base their experimental system on formal semantic structures, while Myaeng, Khoo and Li (1994) construct lexical semantic structures for document representations. Strzalkowski (1994) combines syntactic processing and statistical techniques to enhance the accuracy of representation of the documents. In an innovative approach to document representation for information retrieval, Liddy *et al (*1995) use several levels of linguistic structure, including lexical, syntactic, semantic, and discourse.

9. References

1. Allen, J. 1987. Natural language understanding. Menlo Park, CA: The Benjamin/Cummings Publishing Company, Inc.
2. Bates, M. and R. M. Weischedel (eds.) 1993. Challenges in natural language processing. Cambridge: Cambridge University Press. 8
3. Rosner, M. and R. Johnson (eds.) 1992. Computational linguistics and formal semantics. Cambridge: Cambridge University Press.
4. Saint-Dizier, P. and E. Viegas (eds.) 1995. Computational lexical semantics. Cambridge: Cambridge University Press.
5. Wiebe, J. M. 1994. Tracking point of view in narrative. Computational Linguistics 20.2.233-287.
6. Agarwal, R. and L. Boggess. 1992. A simple but useful approach to conjunct identification. In Proceedings of the 30th annual meeting of the Association for Computational Linguistics. San Francisco: Morgan Kaufmann Publishers. 15-21.
7. Alshawi, H. (ed.) 1992. The core language engine. Cambridge, MA: The MIT Press. Asher, N. and A. Lascarides. 1994. Intentions and information in discourse. In Proceedings of the
8. 32nd annual meeting of the Association for Computational Linguistics. San Francisco: Morgan Kaufmann Publishers. 34-41.
9. Barwise, J. and J. Perry. 1983. Situations and attitudes. Cambridge, MA: The MIT Press.
10. Bates, M. 1989. Rapid porting of the Parlance Natural Language Interface. In Proceedings of the speech and natural language workshop. San Mateo, CA: Morgan Kaufmann Publishers. 83-88.
11. Black, E., R. Garside and G. Leech (eds.) 1993. Statistically-driven computer grammars of English: The IBM/Lancaster approach. Amsterdam: Editions
12. Rodopi. Blosseville, M.J., et al. 1992. Automatic document classification: Natural language processing, statistical analysis, and expert system techniques used together.
13. N. Belkin, P. Ingwesen and A. M. Pejtersen (eds.) Proceedings of the 15th annual international ACM SIGIR conferenceon research and development in information retrieval. NewYork: Association for Computing Machinery. 51-58.
14. Booth, A. D., L Brandwood and J. P. Cleave. 1958. Mechanical resolution of linguistic problems. London: Butterworths Scientific Publications.

15. Brent, M. R. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. Computational Linguistics 19.2.243-262.

16. Brill, E. 1992. A simple rule-based part of speech tagger. In Proceedings of the third conference on applied natural language processing.

17. Trento, Italy. San Francisco: Morgan Kaufmann Publishers. 152-155. 1993. Automatic grammar induction and parsing free text: A transformation-based approach. In Proceedings of the 31st annual meeting of the Association for Computational Linguistics. San Francisco: Morgan Kaufmann Publishers. 259-265.

18. Briscoe, T. and J. Carroll. 1993. Generalized probabilistic LR parsing of natural language (corpora) with unification-based grammars. Computational Linguistics 19.1.25-59.

19. Charniak, E. 1993. Statistical language learning. Cambridge, MA: The MIT Press.

20. Chierchia, G. and S. McConnell-Ginet. 1990. Meaning and grammar: An introduction to semantics. Cambridge, MA: The MIT Press.

21. Chinchor, N., L. Hirschman and D. D. Lewis. 1993. Evaluating message understanding systems: An analysis of the Third Message Understanding Conference (MUC-3). Computational Linguistics 19.3.409-450.

22. Cruse, D. A. 1986. Lexical semantics. Cambridge: Cambridge University Press.

23. Davenport, D. M. and D. T. Heinze. 1995. Crisis action message analyzer - EDM. Proceedings of the 5th annual dual-use technologies and applications conference. SUNY Institute of Technology at Utica/Rome, NY. 284-289.

24. Dermatas, E. and G. Kokkinakis. 1995. Automatic stochastic tagging of natural language texts. Computational Linguistics 21.2.137-163.

25. Fries, U., G. Tottie and P. Schneider (eds.) 1994. Creating and using English language corpora: Papers from the fourteenth international conference on English language research on computerized corpora, Zurich 1993. Amsterdam: Editions Rodopi.

26. Gale, W., K. W. Church and D. Yarowsky. 1992. Estimating upper and lower bounds on performance of word-sense disambiguation programs. In Proceedings of the 30th annual meeting of the Association for Computational Linguistics. San Francisco: Morgan Kaufmann Publishers. 249-256.

27. Gazdar, G., et al. 1985. Generalized phrase structure grammar. Oxford: Blackwell Publishing and Cambridge, MA: Harvard University Press.

28. Grosz, B. J., A. K. Joshi and S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. Computational Linguistics 21.2.203-225.

29. Hajicova, E., H. Skoumalova and P. Sgall. 1995. An automatic procedure for topic-focus identification. Computational Linguistics 21.1.81-94.

30. Huls, C., E. Bos and W. Claasen. 1995. Automatic referent resolution of deictic and anaphoric expressions. Computational Linguistics 21.1.59-79.

31. Jacobs, P. S. and L. F. Rau. 1990. SCISOR: Extracting information from on-line news. Communications of the ACM 33.11.88-97.

32. Justeson, J. S. and S. M. Katz. 1995. Principled disambiguation: Discriminating adjective senses with modified nouns. Computational Linguistics 21.1.1-27.

33. King, M. 1992. Epilogue: On the relation between computational linguistics and formal semantics.

34. In M. Rosner and R. Johnson (eds.) Computational linguistics and formal semantics. Cambridge: Cambridge University Press. 283-299.

35. Kupiec, J. 1993. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia.

36. In R. Korfhage, E. Rasmussen and P. Willett (eds.) Proceedings of the 16th annual international ACM SIGIR conference on research and development in informationretrieval. New York: Association for Computing Machinery. 181-190.

37. Kurohashi, S. and M. Nagao. 1994. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. Computational Linguistics 20.4.507-534.

38. Langacker, R. W. 1988. An overview of cognitive grammar. In B. Rudzka-Ostyn (ed.) Topics in cognitive linguistics. Amsterdam/Philadelphia: John Benjamins Publishing Company. 3- 48.

About Authors

Dr. V R Rathod is a Professor & Head in Department of Computer Science, Bhavnagar University, Bhavnagar, Gujarat.
E-mail : profvrr@rediffmail.com

Prof. S M Shah is a Director in S. V. Institute of Computer Studies, S. V. Campus, Kadi, Gujarat.
E-mail : prof_smshah@yahoo.com

Mr. Nileshkumar K Modi is a Lecturer in S. V. Institute of Computer Studies, S. V. Campus, Kadi, Gujarat.
E-mail : tonileshmodi@yahoo.com

Two-Tier Performance Based Classification Model for Low Level NLP tasks

S Sameen Fatima R Krishnan

Abstract

An error in classification can occur due to an error of omission, statistically known as a false negative or an error of commission, statistically known as a false positive. In order to build a perfect classifier, the false negatives and false positives have to be zero. With this in mind, we propose a two-tier model for the classifier. The first tier will reduce false negatives to zero and pass the results to the second tier. The second tier will reduce false positives to zero. We demonstrate the working of this model for the task of classifying sentences in Hindi as passive formations. The first tier will consist of a simple pattern matching system for filtering out sentences with likely passive formations without committing errors of omission. This will reduce the size of the corpus considerably. The second tier will work on the reduced corpus and make a complete grammatical analysis of these filtered sentences in order to reduce the false positives to a zero. The Anusaraka System [Bharati 1995] is a very good example of such a system. This paper concentrates on building the first tier. A hill climbing algorithm is proposed, where the start state is a list of patterns commonly found in passive formations. Each step up the hill will update the list of patterns such that the next state will bring down the number of false negatives, thereby reducing errors of omission. The hill climbing algorithm terminates when the false negatives are zero.

Keywords : Natural Language Processing, Automated Language Processing

0. Motivation

In continuation of our effort to establish stylistic variation as a basis for genre-based text classification for English language [Fatima, 2001], research has been extended to the Indian languages. As Hindi is spoken by majority of the one billion Indian population, and as it is the official Indian language as well, the choice was restricted to Hindi language. While for English language, software is readily available to identify various linguistic features, to our knowledge no such readily off the shelf software is available in Hindi. Hence, before extrapolating our studies of English language texts for establishing stylistic variation to Hindi language texts, there was a need to build the necessary software. One of the features that directly and indirectly identifies style is the usage of passive sentences. The current work is a description of a model that can be used to classify sentences based on the voice as passive or active formations.

1. Passive Voice and Its Usage

Voice is that form of a verb which shows whether what is denoted by the subject does something (Active Voice) or has something done to it (Passive Voice). In the active voice, the agent (i.e., doer of the action) is made prominent, whereas in the passive voice, the object (i.e., person or thing acted upon) is made prominent [Wren 1936]. Although passive sentences are harder to comprehend than active sentences, they are sometimes needed.

With reference to automatic abstracting, Borko and Chatman [Borko, 1967] have advanced the view that it seems possible to make stylistic distinctions between informative and indicative abstracts. The informative abstract 'discusses the research' and the indicative abstract 'discusses the article which describes the research'. Distinctions in terms of form, in the usage of voice, tense, and the focus of the abstract have been used to differentiate between informative abstracts and indicative abstracts.

The passive voice in Hindi language is generally preferred [Mohanlal]

to emphasize the object of the sentence; as,

dhan paanii ki taraha bahaayaa jaa rahaa hei.

to avoid naming an unimportant character; as,

patra bhej diyaa gayaa hei.

To assert authority; as,

apraadhii ko pesh kiyaa jaae.

When the doer refers to a sabha, samaj or the government; as,

aaryasamaaja dwaaraa kii antarjaatiiya vivaah karaae jaate heiM.

In official language; as,

binaa TikeT safara karane vaaloM ko daNDit kiyaa jaayegaa.

to express inability to do something

mujha se khaanaa khaayaa nahii jaataa.

1.1 How can we identify passives ?

To identify passive sentences in Hindi, the following observations were made:

Two words in Hindi are popularly found in a passive formation namely ‘dwaara’ and ‘se’. For example:

yaha upanyaas muMshii premacanda dwaaraa likhaa gayaa hei.

mujhse kheera khaayii gayii.

Of the two ‘dwaara’ is more likely an indicative of a passive formation than ‘se’. This is because of the simple reason that ‘se’ can be used both in the sense of ‘by’ and ‘with’, whereas ‘dwaara’ is mostly used in the sense of ‘by’ only. Whenever the usage of ‘dwaara’ or ‘se’ is an indication of the instrumental case, then it is surely not an indication of a passive formation. For example:

caakuu se seba kaaTo.

mei dillii TRena dwaaraa jaa rahaa huuM.

Hence, ‘dwaara‘ and ’se’ by themselves are not an indication of passive voice.

In passive formations along with the head verb, the conjugation of the verb ‘jaana’ is used. For example: ‘jaana’ takes the form: ‘jaa saktaa’ or ‘jaayegaa’, where ‘saktaa’ and ‘gaa’ are auxiliary verbs or modals. The modals give information about gender, number and the tense (‘saktaa’ for future and ‘gaa’ for past). The modals may be

separated from the verb ‘jaa’, like ‘jaa saktaa/jaa saktii/jaa sakte/jaa rahaa/jaa rahii/jaa rahe’ is used in:

khaana khilaayaa jaa saktaa hei.

bacciiyaaM douDaayii jaa rahii thii.

kai khela khilaaye jaa sakte heiM.

or may combine with the verb ‘jaa’ like ‘jaaega/jaaegii/jaaenge’, ‘jaataa/jaatii/jaate’, ‘jaauuMgaa/jaauuMgii/ jaayenge’, is used in

khaana khilaayaa jaayegaa.

paaThshaalaa mein paDaayaa jaataa hei.

mei douDaayaa jaauuMgaa.

Passive formations also use the forms ’gayaa/gaii/gayii/gae/gaye’ of the verb ‘jaana’ as in:

mujhse kheera khaaii gayii /mujhse kheera khaaii gaii

kaii Teliphona kiye gaye./ kaii Teliphona kiye gae.

In addition to information consisting of the verb ending in modals, it can also be noted from the above examples that the vibhakti ‘yaa/yii/ye’ (which is the karma karak) is also used in passive formations. Collectively the vibhakti and the modals for a verb give information about tense, aspect and modality (TAM), and is, therefore, also called the TAM label. TAM labels for passive formations are purely syntactic and are determined from the combination of the verb (‘jaanaa’) with the modals and the vibhakti (‘yaa/yii/ye’).

The above discussion can be extrapolated to include negation. Negation is indicated by the word ‘nahii’ in between the vibhakti and the conjugation of the verb ‘jaanaa’. The following passive formations show the usage of negation:

mujhse khaayaa nahii gayaa

kii khela khilaaye nahii gaye

2. NLP approaches for Identifying Passive Formations

When considering a technology to support high-precision text classification, natural language processing (NLP) is one of the first things that comes to mind. Work done to date on NLP systems since early days have varied widely in their approach to analyzing texts. At one end of the spectrum were systems that processed a text using traditional NLP techniques. At the other extreme lie systems that use keyword/ pattern matching techniques and little or no linguistic analysis of the input text. [Cardie, 97]

The traditional approach to sentence analysis in natural language processing works on the assumption that a system must assign a complete constituent analysis to every sentence it encounters. The methods used to attempt this are drawn from mathematics, with context-free grammars playing a large role in assigning syntactic constituent structure. Using this approach, the following steps are taken for identifying passive formations:

? Tokenization: the input text is divided into sentences and words.

? Tagging: a dictionary or lexicon is looked up, and associated grammatical information (like parts of speech) is retrieved and used for tagging the words.

? Sentence-analysis: Noun groups and verb groups are identified. Verb groups are more important from the point of view of identifying passive formations.

The Anusaraka System described in Bharati, 1995 is a very good example of a complete system built on these principles. A sentence is first processed by a morphological analyzer (morph). The morph considers a word at a time, and for each word it checks whether the word is in the dictionary of indeclinable words. If found, it returns its grammatical features. It also uses the word paradigms to see whether the input word can be derived from a root and its paradigm. The output of morph is given as input to the local word grouper. Its main task is to group function words with the content words based on local information such as postposition markers that follow a noun, or auxiliary verbs following a main verb. This grouping (or case endings in case of inflectional languages), identifies vibhakti of nouns and verbs. The vibhakti of verbs is also called TAM (tense-aspect-modality) label. These TAM labels are used to identify passive formations.

One of the biggest challenges in natural language processing is the need to make NLP systems robust by providing it linguistic sophistication, while at the same time not trading off speed of processing. If a preprocessing step for filtering out likely passive formations precedes the complete analysis of the sentences, we would not have to grammatically process each and every sentence in the text. Hence, it was felt that if passive sentences can be mined for frequently occurring patterns, a simple pattern matching technique can be used for filtering likely passive sentences. This would reduce the corpus size. On the reduced corpus a complete analysis of the sentences can be carried out.

Currently, general pattern matching techniques have become the technique of choice for the extraction phase of an information extraction system [MUC-6 1995]. A number of researchers have investigated the use of corpus-based methods for learning information extraction patterns. The learning methods vary along a number of dimensions: the class of patterns learned, the training corpus required, the amount and type of human feedback required, the degree of preprocessing necessary, the background knowledge required, and the biases inherent in the learning algorithm [Cardie, 1997]

Sentence

"Witnesses conform that the twister occurred without warning at approximately 7:15 pm and destroyed two mobile homes."

Concept-Node Definition

Concept = Damage
Trigger = "destroyed"
Position = direct-object
Constraints = ((physical-object))
Enabling Condition = ((active voice))

Instantiated Concept Node

Damage = "Two mobile homes"

Figure 1: Concept node for extracting damage information

Information extraction is a subfield of natural language processing that is concerned with identifying predefined types of information from text. For example, an information extraction system designed for a terrorism domain might extract the names of perpetrators, victims, physical targets, weapons, dates, and locations of terrorist events. One of the earlier systems for acquiring extraction patterns was AUTOSLOG [Riloff 1996]. AUTOSLOG learns extraction patterns in the form of domain-specific concept nodes. It uses a small set of heuristic rules to decide what expression should activate the case frame, and from which syntactic constituent the slot should be filled. Figure 1 for example shows the concept node for extracting "two mobile homes" as damaged property, from a given sentence, using the keyword "destroyed" as a trigger.

4. Two-Tier Performance based Model for Classifiers

In order to build a true model of a classifier it should be bereft errors. An error is a misclassification: the classifier is presented a case, and it classifies the case incorrectly. Depending on the application, distinctions among different types of errors turn out to be important. For example, the error committed in diagnosing someone as healthy when one has a life-threatening illness (known as a false negative decision) is usually considered far more serious than the opposite type of error-of diagnosing someone as ill when one is in fact healthy (known as a false positive).

In order to distinguish between different types of errors, it is important to build a contingency matrix. The contingency matrix lists the correct classification against the predicted classification for each class. The number of correct predictions for each class falls along the diagonal of the matrix. All off-diagonal are the errors for a particular type of misclassification error. Table 1 is an example of such a matrix for two classes. The label Positive denotes the class of sentences with passive formations and the label Negative denotes the class of sentences with non-passive formations.

Performance of Classifier		**True labels**	
		Positive (Passive)	Negative (¬ Passive)
Predicted labels	Positive (Passive)	True positives (TP)	False Positives (FP)
	Negative (¬ Passive)	False negatives (FN)	True Negatives (TN)

Table 1: Sample contingency matrix for a two-class classification problem

A classic metric for reporting errors of omission and commission are sensitivity and specificity. Sensitivity is a measure of the errors due to commission and specificity is a measure of the errors due to omission. Using the notation in Table 1 sensitivity and specificity can be expressed as:

$$Sensitivity ? \frac{Truepositives}{Allpositives} ? \frac{TP}{TP ? FN} ? \text{Re}\, call$$

$$Specificity ? \frac{Truenegatives}{All\,\text{ln}\, egatives} ? \frac{TN}{TN ? FP}$$

$$\text{Pr}\, ecision ? \frac{Truepositives}{Allpredictedpositives} ? \frac{TP}{TP ? FP}$$

In the evaluation of information retrieval systems, the most widely used performance measures are *recall* and *precision*. However, S*ensitivity* and *specificity* are used to bring out the difference in the error types (errors of commission and errors of omission). *Sensitivity* is the accuracy among positive instances and *specificity* among negative. *Recall and precision* are mostly utilized in situations where *TP* is small when compared with *TN*. A perfect classifier can be described as one in which both false negatives and false positives are zero or in other words in which both the sensitivity and specificity would equal one [Baldi 2000].

Based on the above, a two-tier model is proposed in figure 2. The main goals of each tier are as follows:

Tier 1: This aims at reducing the false negatives to a zero. It achieves this by using a pattern matching filtering system.

Tier 2: This aims at reducing the false positives to a zero. It achieves this by using a complete natural language processing system.

At the end of the first tier, errors of omission are dangerous, however errors of commission can be tolerated. The second tier will take care of errors of commission. As Anusaraka [Bharati 1995] is available for complete grammatical analysis at Tier 2, this paper concentrates on building tier 1.

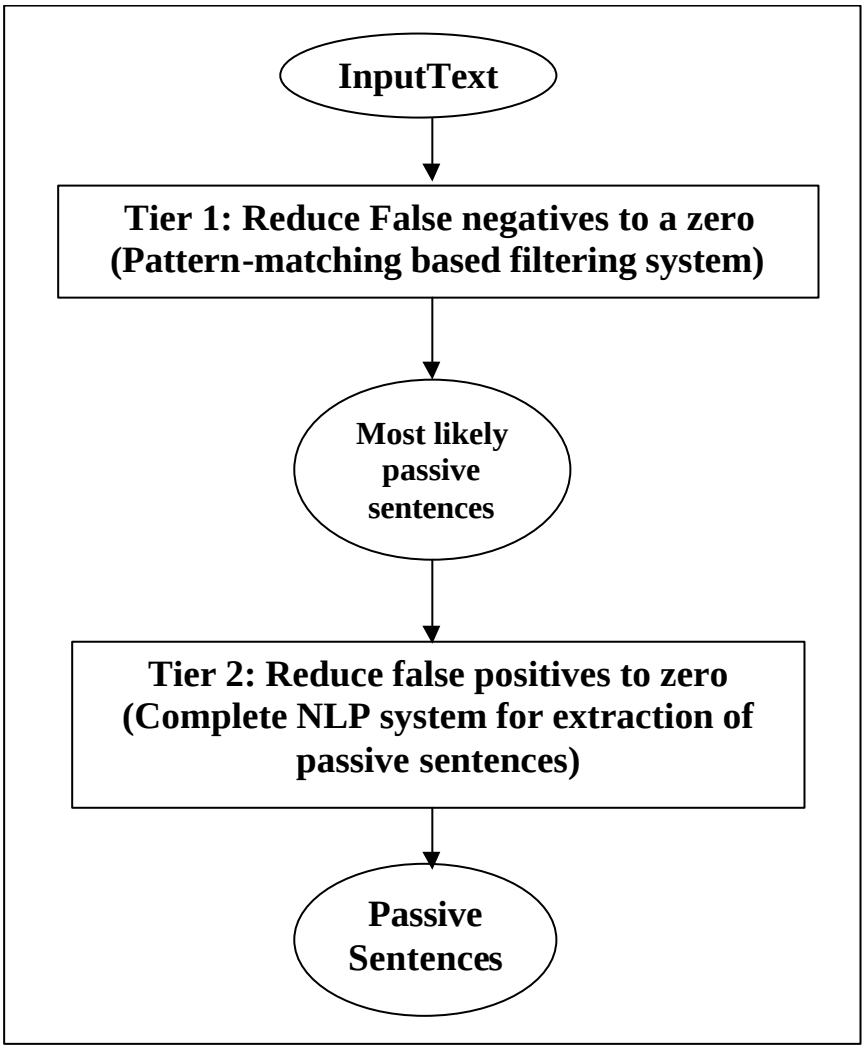


Figure 2: Two-tier performance based model for identifying passive sentences

4. Design of TIER 1

4.1 Hill Climbing Algorithm

In building tier 1, our focus was on the problem of distinguishing sentences as belonging to one class (passive) from another (not passive) – a binary supervised classification problem. An iterative improvement algorithm like hill climbing is used which improves sensitivity by decreasing the false negatives. It works in 2 phases:

Generate Phase: A list of patterns L+, commonly found in the Positive (passive) class of sentences was generated

Test phase: The performance of the list L+ was tested on a corpus of Hindi editorials, C. The number of passive sentences that are not filtered by the classifier were found out, which gave the number of false negatives, FN. If FN was zero we quit, otherwise we return to step 1, that is the Generate Phase

4.2 The State Space

The state space representation used in the hill climbing algorithm is described below:

State : A state corresponds to a list of patterns L+, for identifying sentences with passive formations.

Initial State Generation: Several passive sentences from common usage were taken as input to manually identify string patterns commonly found in passive sentences. This list of patterns in the initial state were generated with the help of a linguist, our background knowledge of Hindi and a corpus. The corpus was chosen as editorials of Hindi newspapers. The initial state was generated manually as a pre-tagged or pre-parsed training corpus for identifying passive formations was not available. Although appealing, the development of annotated corpora is still tedious and labor intensive [Srinivas 2001].

Next State Generation: The passive sentences in a corpus of editorials, C were filtered using list L+. A list of passive sentences not filtered by the help of the existing list L+ were identified. Each pattern in the list L+ was examined and compared against the sentences that were not filtered by the classifier. The list was refined with the help of feedback from linguists, from our background knowledge of Hindi and at times with the help of functions to identify the minimal common pattern between the patterns available in the list L+. As a result the existing patterns in the list were being deleted or updated, and new patterns were identified for inclusion wherever necessary in the list L+. The new list was the next state. The heuristic used was: decrease the number of false negatives FN.

Local Maxima: Each uphill move may involve a number of unsuccessful attempts (i.e., visits to nodes which increase the number of false negatives). A solution is called a local maximum if no uphill moves can be performed starting from the current state. If hill climbing reaches a local maximum, then we backtracked in search of a better solution in other areas of the state space. The cases not accounted for were again incorporated.

Goal State: The list of patterns for which the number of false negatives, FN was zero.

5. Experiments and Results

Our string pattern generation for identifying passive formations was guided by two main objectives: On the one hand, we give preference to the generation of short patterns (simplicity principle), and on the other hand, we attempt to cover every passive sentence by at least one pattern (comprehensive principle).

5.1 Generation of Goal State

After several iterations of the Generate and Test phase, two lists evolved: List L1 and List L2. Both the lists had zero false negatives, which was the aim of tier 1. However, list L2 had lesser false positives than list L1.

List L1 was a shorter list consisting of the string patterns shown in Table 2.

1. aa ga
2. aa jaa
3. ii ga
4. ii jaa
5. e ga
6. e jaa

Table 2: Goal State - List L1

A corpus consisting of 1626 sentences taken from editorials of Daily Hindi Milap and Vaartha were filtered based on the list L1. The results were analyzed and a contingency matrix shown in Table 3 was created

Performance of Filter		**True labels**	
		Passive	¬ Passive
PredictedLabels	Passive	337	61
	¬ Passive	0	1228

Table 3: Contingency matrix for the filter for classifying sentences as passive

$$Sensitivity ? \frac{337}{337 ? 0} ? 1$$

$$Specificity ? \frac{1228}{61 ? 1228} ? 0.95$$

As can be observed from Table 3, the number of false negatives are zero. This means that the list covered all patterns necessary for identifying passive formations, and hence sensitivity is one. Thus the filter at Tier 1 is perfect. However, there are 61 false positives, that is 61 sentences that were not passive were also filtered. This will be taken care of by Tier2. The number of sentences filtered (as likely passive sentences) are 398 (= 337 + 61), which will be fed to Tier 2. Tier 2 will perform a complete grammatical analysis on the 398 sentences filtered by Tier 1. At the end of tier 1 the size of the corpus reduced from 1626 to 398, which was considerable.

The second list was a larger list. The first four string patterns from List 1 were taken as it is. However, string patterns 5 and 6 from List 1 were replaced by string patterns 5 to 10 in List L2, with a view to improve specificity as shown in Table 4.

1. aa ga
2. aa jaa
3. ii ga
4. ii jaa
5. ie ga
6. ie jaa
7. ye ga
8. ye jaa
9. (dwaaraa Ú se) Ù (e ga)
10. (dwaaraa Ú se) Ù (e jaa)

Table 4: Goal State - List L2

The pattern strings numbered 1 to 8 in List L2 look for a single pattern match in the sentence. Example:

If the pattern “aa ga” is found in the sentence

then conclude it is a passive formation

However, the pattern strings 9 and 10 in Table 4 are looking for two patterns in a sentence. If both are available then it qualifies for a passive sentence, otherwise it is discarded. Example:

If “dwaaraa” and “e ga” are found in a sentence

then conclude it is a passive formation.

This was done in order to exclude sentences of the form given below from being classified as passive formations:

raama miThaii khaa*e ga*a.

It was found that the expanded list L2 shown in Table 4 decreases the false positives and thereby improves the specificity while not effecting the sensitivity.

Using the list L2, a contingency matrix shown in Table 5 was created

Performance of Filter		**True labels**	
		Passive	¬ Passive
PredictedLabels	Passive	337	13
	¬ Passive	0	1276

Table 5: Contingency matrix for the filter for classifying sentences as passive

$$Sensitivity ? \frac{337}{337 ? 0} ? 1$$

$$Specificity ? \frac{1276}{13 ? 1276} ? 0.99$$

As can be observed from Table 5, the number of false negatives are zero. This means that the list covered all patterns necessary for identifying passive formations, and hence sensitivity is one. Thus the filter at Tier 1 is perfect. Further, comparing the performance of list L2 (table 4) against list L1 (table 2) the following observations were made:

the number of false positives are reduced to 13 from 61, which increases the specificity from 0.95 to 0.99.

The number of sentences filtered (as likely passive sentences) are reduced to 350 (= 337 + 13) from 398. At the end of tier 1 according to List L2, the size of the corpus reduced from 1626 to 350, which is roughly one-fifth (21.5%) of its original size.

6. Conclusion

In-depth natural language processing (NLP) for high precision text classification is an expensive endeavor that can strain computational resources [Riloff 1994]. As an alternative to full-blown NLP, we have presented a 2-Tier Performance-Based Model for a Classifier. This model represents a compromise between pattern-matching techniques, at Tier 1, and in-depth natural language processing at Tier 2, so that the performance of the classifier is perfect. We evaluate the model for a low-level NLP task of classifying sentences as passive. We built Tier 1 using a Hill Climbing algorithm. At the end of Tier 1 the sensitivity was 1 (perfect) and the specificity was 0.99 (very nearly perfect) and the corpus size was reduced to 21.5%. Tier 2 can use a system like Anusaraka , which does full blown grammatical analysis of the sentences. The load on such a system will be minimal. It has to process just 21.5% of the corpus, and has to increase the specificity from 0.99 to 1. The results suggest that pattern-matching techniques can support high-precision classification for low level NLP tasks without straining computational resources.

7. References

1. P. Baldi, S. Brunak, Y. Chauvin, CAF Anderson, H Nielsen. Assessing the Accuracy of Prediction Algorithms for Classification: An Overview. Bioinformatics, 16 (5), 2000, pp 412 – 424.
2. Herald Borko. Automated Language Processing. John Wiley and Sons. 1967.
3. Claire Cardie. Empirical Methods in Information Extraction. AI Magazine. Winter 1997.
4. Sameen Fatima, R Krishnan. Stylistic Variation as a Basis for Genre-Based Text Classification. IETE Journal of Research. Vol 47, Nos 1&2, Jan – Apr 2001, pp 59 – 63.
5. Mohanlal D, Ashok B. Navyug hindi vyaakraN tathaa racnaa. Lakshmi Publications Pvt Ltd. New Delhi.
6. Proceedings of the Sixth Message understanding Conference (MUC-6). San Francisco, California. Morgan Kaufman, 1999.

7. Riloff E, Lehnert W. Information Extraction as A Basis for High-Precision Text Classification. ACM Transactions on Information Systems, Vol 12, No. 3, July 1994, pp 296 – 333.

8. Riloff E. Automatically Generating Extraction Patterns from Untagged Text. Proceedings of the 13th National Conference on AI. Menlo Park, Calif. AAAI. 1996, pp 1044 – 1049.

9. Bharati A, Chaitanya V, Sangal R. Natural Language Processing – A Paninian Perspective. Prentice Hall of India. 1995.

10. Srinivas B. Annotated and Unannotated Corpora in Natural Language Applications. Proceedings of the Workshop on Lexical Resources for Natural Language Processing, Language Technologies Research Centre, International Institute of Information Technology, Hyderabad, Jan 2001.

11. PC Wren, H Martin, NDV Prasada Rao. English Grammar and Composition. 1936.

About Authors

Dr. S Sameen Fatima is working in Department of Computer Science & Engineering, College of Engineering, Osmania University, Hyderabad, India.
E-mail : sameenf@rediffmail.com

Dr. R Krishnan is working in Advanced Data Processing Research Institute, Department of Space, Govt. of India, Secunderabad, India
E-mail : drrk@hotmail.com

Globalization of Software Applications Using UNICODE Based Multilingual Approach

Sonia Dube Yatrik Patel T A V Murthy

Abstract

Concept of Globalization of Software can be boom to advancement of projects related to Digital libraries and associated software. Much need is felt in the Indian context where support is to be provided for many languages to take care of diversified regional requirements and complexity of INDIC script. In this paper we have presented an approach and Implementation for creating Globalized software using UNICODE based Multilingual approach.

Keywords : Multilingual computing, Unicode, Globalization, Internationalization, Localization.

0. Motivation

Developing software acceptable and adaptable to users across the continents and users with different scripts and languages as in case of India is a challenging job. The process of making software acceptable on global basis among user with different languages is carried out using concepts of Globalization, which can be carried out using concepts of Internationalization and Localization, which focuses on technology development to bring users to follow and adapt standards. Adapting the software to meet the specific requirements of customer for interaction with the software is done using concepts of Localization. Scientific advancements in the field of multi-lingual computing and Software standardization such as Unicode has helped in building user friendly software, acceptable by people from different zones and areas. The National Language Support (NLS) supplied by the Microsoft Win32 application programming interface (API) can be used for making Internationalized Software components, whereas modifying the user interface (UI) elements, translating text, and standardizing terminology are localization steps. Success of Globalize Software can only be ensured when these concepts are incorporated in the software from the design phase and issues of Internationalization and Localizations are properly taken care of.

1. Unicode

Computers during older days were mainly used for number crunching, but with advancements in the processing power and technology associated with multi-media has made them more users friendly and now are readily being used by people in all possible fields, which can be thought of. Initially, computers just dealt with numbers and store letters and different characters by unique numbers, there was no fixed encoding schemes available to assign numbers to different characters. These encoding schemes were not powerful enough to deal with all available letters and characters. The conflict in the encoding schemes created lot of problems in bringing different applications on a single platform. Unicode was invented to overcome the limitations imposed by old incompatible computer encoding standards ASCII, ISCII, JS etc and has now become standard for character encoding. Unicode apart from standardizing the application has also being considered as a major step towards simplifying multilingual computing.

Unicode[1] is a universal encoded character set that enables information from any language to be stored using a single character set. Unicode provides a unique code value for every character, regardless of the platform, program, or language.The role of unicode in maintaining international standard is to define

characters from an internationational perspective, without preventing any culture from defining its own character sets for internal communications.

There are different techniques to represent each one of the Unicode code points in binary format. Each of the techniques uses a different mapping to represent unique Unicode characters.

The Unicode encoding are classified as:

UTF-8 : UTF-8 has been defined by the Unicode Standard to meet the requirements of byte-oriented and ASCII-based systems. Each character is represented in UTF-8 as a sequence of up to 4 bytes, where the first byte indicates the number of bytes to follow in a multi-byte sequence, allowing for efficient string parsing. UTF-8 is highly used in Internet for content exchange and data transfer.

UTF-16 : This is the 16-bit encoding form of the Unicode Standard where characters are assigned a unique 16-bit value, with the exception of characters encoded by surrogate pairs, which consist of a pair of 16-bit values. The Unicode 16-bit encoding form is identical to the International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) transformation format UTF-16. In UTF-16, any characters that are mapped up to the number 65,535 are encoded as a single 16-bit value; characters mapped above the number 65,535 are encoded as pairs of 16-bit values. As more name space is provided in this format and hence there exist provision for incorporation on more languages in Indian context [2]

UTF-32 : Each character is represented as a single 32-bit integer.

2. Multilingual Applications

Multilingual computing [3] is defined as “Use of computer to communicate with people in various languages”. Multilingual computing makes it possible to use the Internet for global communication by supporting more than one language simultaneously and has the ability of handling more than one script of character sets. In multilingual application a single executable can be run in multiple languages and can handle data in multiple languages. At the time of application loading, multilingual enabled application provides the option to the user for language selection. The design of a multilingual application should be such that the language of the application is separate from the language of the data and provision should exist for encoding of data using standards such as Unicode.

In Indian context lot of work is carried out by CDAC for developing Multilingual Technology and Applications such as “Acharya” developed by IITM for providing Multilingual computing environment. Attempts are being made to incorporate Multilingual features in the new version of SOUL.

3. Potential Applications for Multilingual Computing are

- Digital Libraries
- Regional Web Portals
- Education to Home

4. Implementation Scenarios

Unicode based multilingual global applications can be developed using support provided by Visual Studio.Net platform and MS SQL Server, which can be used for backend database storage. Two scenarios can be thought of for Implementation of Multilingual application based on user requirements.

? Homogeneous Framework: In this scenario the Application Language and the Data Language is same and the user selects the language during the start of the application. This is a very crude way of implementing Multilingual applications and can be used where applications require handling of few languages.

? Heterogeneous Framework: In this approach the system supports for application and data to be of different languages, user selects the language when the application starts and data language can be changed as per the user requirement for providing inputs to the applications, however outputs from the application will automatically be displayed in the language in which data was stored. Although there may be very little requirement of changing the application language during the run-time but such requirements can also be taken care of in this framework. This approach can supports handling of different languages in a single session of application and can help in building true Globalized applications.

Comparative Study for different Frameworks

	Homogeneous	**Heterogeneous**
Web Support	Different Web pages for different Languages	Unicode enabled web-pages and browser with same web pages
Database Archival	Different Database for different Languages	Same Database for Different Languages
Run Time Changing of Applications Language	No	Yes
Run time Changing of Data Language	No	Yes

5. Operating System and Database Support for Multilingual Software Development

Windows 2000/NT/XP and above are all completely Unicode enabled. So all provides best support for Unicode and all languages to implement multilingual applications. Windows provides language options to select the required language. Even though the older versions of windows do support but require the specific language edition of the operating system to support.

MS Visual Studio.Net System Globalization namespace can be used for changing the data language in applications being developed on .NET framework.

MS SQL server uses nvarchar and nchar data types for storage of UNICODE data in the database tables.

6. Benefits of Globalization

Globalization of software application offers various benefits to developers and users. From developer point of view the benefits are :

? Developer has to develop application only once, no need to change source code when new language requirement appears.

? Product can compete in global market.

? Users can take benefit of using multiple languages in a single product only.

? Users can maintain global standards in their data by using same software product.

? By creating and using multilingual applications developers and users can save lots of time and money.

7. Conclusion

Globalization of software application by using multilingual approach is the best way to share product among users having multilingual requirements. Unicode has evolved as major standard and has simplified the task of multilingual computing and Globalization of the software. The heterogeneous framework proposed above helps in building true Globalized applications.

8. References

1. The Unicode Standard, Version 3.0, Addison-Wesley, Massachusetts, 2000
2. Devika Madalli, Unicode for Multilingual Representation in Digital Libraries from Indian Perspective, Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital, 2002, Page 398
3. "Introduction to multilingual computing" http://www.riverlion.com/jpfaq/mlcintro.pdf

About Authors



Mrs. Sonia Dube is Project Scientist at INFLIBNET Centre, Ahmedabad, Gujarat and holds MCA. Earlier she was Programmer in Projects at INFLIBNET funded by NISSAT and WHO.
E-mail : sonia@inflibnet.ac.in



Mr. Yatrik Patel is Scientist - B at INFLIBNET Centre, Ahmedabad, Gujarat and holds B.E. in Computer Science. He is engaged in the software R&D activities, SOUL design, development, installation and troubleshooting and having expertise in server configuration & very large databases. He has published several articles at national and international level and recipient of best paper presentation award by IATLIS and second best article prize by RRLF, Kolkatta on public libraries.
E-mail : yatrik@inflibnet.ac.in



Dr. T A V Murthy is Director of INFLIBNET Centre, Ahmedabad, Gujarat and holds B.Sc, M L I Sc, M S L S (USA) and Ph.D. He is President and Fellow of SIS-India, Hon. Director of E.M.R.C., Gujarat University and Member Secretary of ADINET, Ahmedabad. He carries with him a rich experience and expertise of having worked in managerial level at a number of libraries in many prestigious institutions in India including National Library, IGNCA, IARI, Univ of Hyderabad, ASC, CIEFL etc and Catholic Univ and Case western Reserve Univ in USA. He has been associated with number of universities and has guided number of Ph.Ds and actively associated with the national and international professional associations, expert committees and has published good number of research papers. He visited several countries and organized several national and international conferences and programmes.
E-mail : tav@inflibnet.ac.in

Enabling Indic Support in Library Information Systems : An Opensource Localizer's Perspective

Indranil Das Gupta Najmun Nessa

Abstract

This article looks into the unique nature of challenges and opportunities facing the Free & Opensource (F/OSS) based software localizers' community when it comes to enabling support for Unicode-based Indic Scripts in the domain of Library & Information Science (LIS). It describes the early background of Indian language support in LIS domain in terms of technology used, and moves into the present-day scenario of Unicode & Open standard based method of universal archival and access to information repositories that modern libraries represent with their multi-media capabilities. Unicode addresses many of the problems that had plagued earlier systems which had little or no capabilities in terms of universal accessibility, it also brings its own set of problems that demand solutions – e.g. the issue of collation sequences which assume significance when looked at from the perspective of indexed search capabilities in library software. While Opensource provides an open, pro-active, collaborative platform for rapid development, it still has to answer for issues like availability of extensive Opentype fonts, collation sequences, less-than desired quality of rendering by Indic script layout engines, as well as varying levels of maturity of software components that make up the technology stack on which Indic Support enabled Library Information Systems can and are being developed. The authors will try to seek answers to these practical questions by looking into their localization experiences with Koha – the world's first Opensource library software into Bengali (this work is being followed by Hindi localization). Inputs will also include the experiences of the team from ISI, Kolkata which is working on localizing Greenstone Digital Library (GSDL) into Bengali. The article will draw upon the experiences of F/OSS Indic Localizers' community to see whether cross-pollination of ideas can lead us towards the goal of bridging the Digital Divide.

Keywords : Indic Scripts, Unicode, Localization, Library Automation Software.

0. Information Divide & the Emerging Role of F/OSS[1] in ICT4D

In the present day world, information technology (IT) is a key part of the infrastructure development. ICT (Information & Communication technologies) penetration is being measured as part of development indices. Access to information and the information technology has emerged as a key to the development in any sector – be that education, access to health-care, access to markets for rural produce, to overseas trade, entrepreneurial development, public & private investment, and even the governance of a country.

Economic disparities separate the developed nations from the rest. As a result, the developing nations are lagging behind in adoption of IT. This is further aggravated by factors like poor literacy rates, multi-lingual societies with little or no comprehension of English (which is the *de-facto* language of IT). All these factors has given the English dictionary a new word – The Digital Divide.

Today the Digital Divide has emerged as one of the primary obstacles to development. Within countries like India, Brazil & China, it has assumed far greater complicacy because in these countries there has emerged the domestic Digital Divide. People within the country who have access to the latest in information technology while the majority of the population does not have even the most basic mode of access to IT.

The emergence of Free/Opensource software as a global phenomena has resulted in redrawing the ICT4D strategy maps. All across the developing world, the access to the underlying technology (via the programs' source code) and the license to add, modify and improve – either at will or driven by a need to address specific requirements, has opened windows of opportunities of hitherto unparalleled dimensions and importance[2].

1. Relevance of these Developments in Library Information Systems

Through the ages, libraries have been the cornerstone of recorded information collectively available to any society. The emergence of the Internet as an affordable, instant, global, digital communication medium has perhaps brought about the most significant change since Gutenberg's invention of the printing press as to how libraries acquire, disseminate and manage information.

Media changes notwithstanding, libraries have acquired greater significance in the modern knowledge-based societies with their well-defined classification, cataloging services. In developing countries they have a far greater role to play in bridging the Information/Digital Divide.

In India, the demand for delivering timely, multi-lingual, multi-media based digital content has emerged as the need of the day. Special libraries services have sprung up to cater to these needs. The much talked-about Vidyanidhi[3] Project is one such key example.

The reason we chose to address Library Information Systems as against taking up only Integrated Library Management Systems (ILSes, similar to LibSys etc) or Digital Library Software (GSDL and DSpace et. al.) is quite simple. The problems solved, problems pending and the lessons learnt in course of developing Indic-enabled F/OSS solutions are equally applicable across the range.

Being native speakers of the Bangla (BN_IN) language and being actively involved in BN Localization (L10N) efforts as part of AnkurBangla Project, IndLinux Consortium, and L2C2 Initiatives, examples used will heavily depend on our experiences with Bengali. However, most of the examples, except for the very specific ones, apply across most Indian languages.

Our work on localizing Koha or for that matter others' initiatives on F/OSS solutions like the Greenstone Digital Library[4] Software (GSDL) or DSpace[5] couldn't have happened without the Indic Language support being in place on the Free & Opensource platform. We believe that it is essential to share the basic know-how of localization on F/OSS with the Library & Information Science community, now that F/OSS Indic support have become mature enough to provide support to localization of 3rd party software.

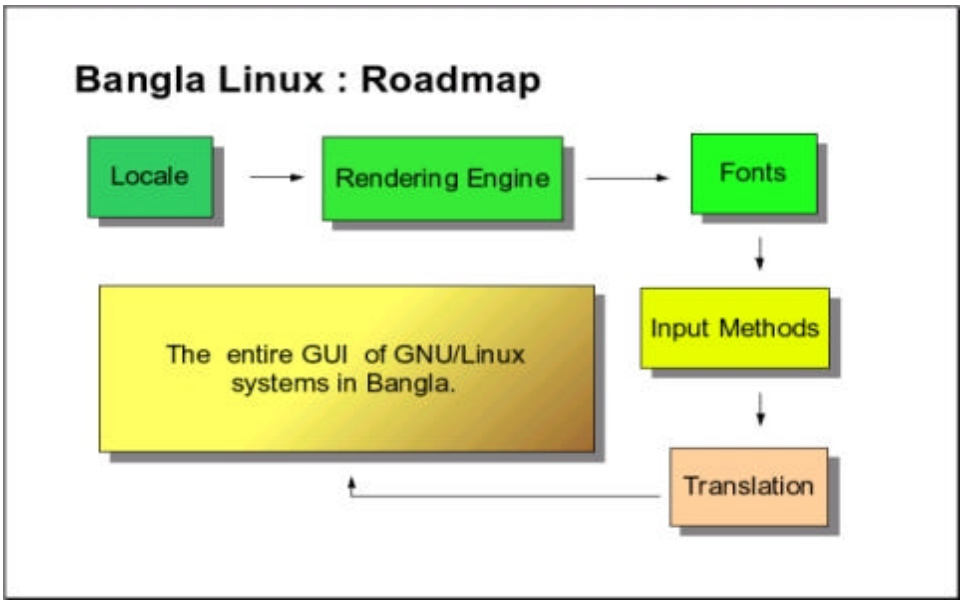
In the ensuing discussion, references to F/OSS will primarily focus of OSS software working on the GNU/Linux[6] Operating System which has the most mature Indic support among other Free OSes.

2. The F/OSS based Indic Localization Roadmap

F/OSS-based development models have always been collaborative from a pluralistic sense. Since its early days it was *assumed* (unlike in the case of a lot of proprietary systems) that the software being developed *would* be used by non-english speaking users. This is not surprising since developers across the globe would collaborate on projects using the Internet as the ultimate project management platform.

So, when F/OSS based Indic Localization (L10N) initiatives got off the ground, the basic software engineering framework was already in place. It goes without saying that this framework did have its shortcomings which has since been addressed. With the framework in place, Indic Localizers could focus on creating the basic artefacts required to deliver a L10N-ised platform to the end-users.

These basic artefacts included – Unicode support, creation/correction of Locale Data[7], correcting/modifying rendering & layout engine programming code, creation of OpenType Fonts, creating Input Methods for text entry and finally User-Interface (UI) translation. Below is a slide from an AnkurBangla presentation depicting the basic components needed to deliver the Bangla GUI Interface on Linux. This applies in case of all Indic Languages.



NB : The above slide omits a major component in the localization technology stack – the *collation sequences*. We shall come back to that in due course.

We shall take a closer look at some of these components, as the issues to be presented applies equally in case of Library Information Systems as in any other application domain based on F/OSS platform as we found in course of our work.

3. Indian Standard Code for Information Interchange (ISCII) – The Past

During the late-80's and early 90's C-DAC[8] (Centre for Development of Advanced Computing), then under the Department of Electronics, Govt of India, created a standard called ISCII (Indian Standard Code for Information Interchange) for use of Indian Languages on Computers.

ISCII uses a 8-bit encoding that uses escape sequences to announce the particular Indic script represented by a following coded character sequence. The ISCII document is IS13194:1991, available from the BIS offices.

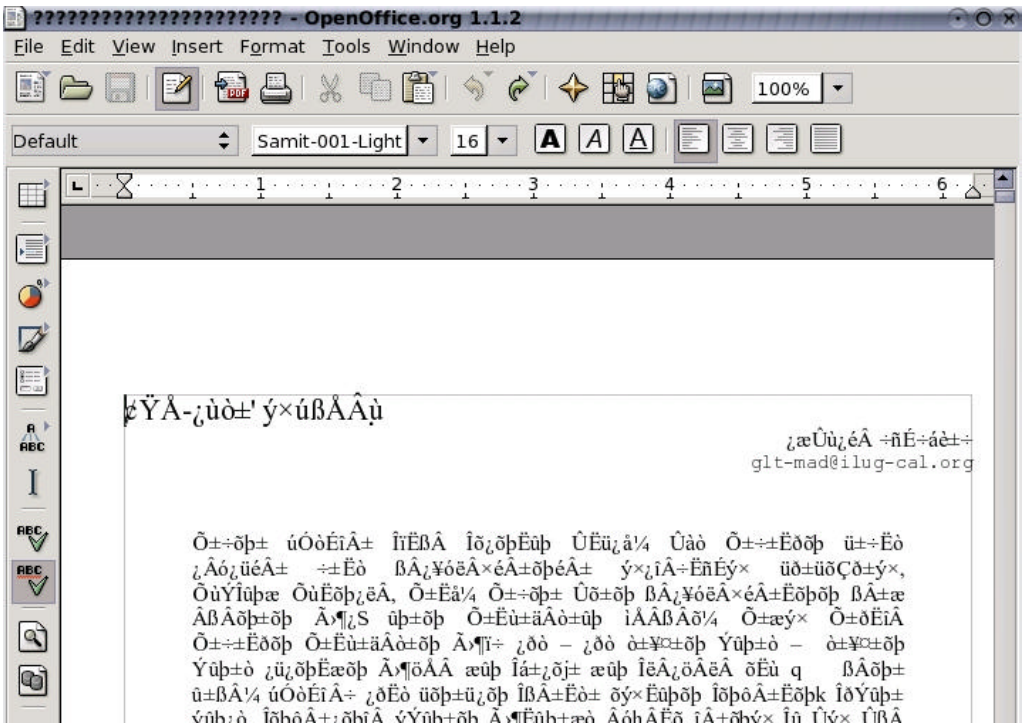
Alongside ISCII, other proprietary Indian language solutions did exist prior to Unicode. The most criticized aspect of these developments was the proliferation of encoded fonts using closed, proprietary formats. As a result, most of these solutions didn't (or rather couldn't) exchange data between themselves or other software. Vendors who created these software did this on purpose to ensure lock-in of the users to their specific software.

4. Unicode – New Challenges, New Possibilities

Unicode – a plain text standard, which is an idea of simplicity, promises to change all that for the future. All major operating systems (Windows, Mac OSX, Linux etc.) today support Unicode as a data format. Most are even beginning to support Unicode at GUI levels. As a result, on Unicode enabled platforms, it now possible to copy a piece of text written in Unicoded Hindi, paste it into a web-page that you are designing or store it AS-IS into a RDBMS like Oracle, Sybase or PostgreSQL.

Fundamentally, computers just deal with numbers. They store letters and other characters by assigning a number for each one. Before Unicode was invented, there were hundreds of different encoding systems for assigning these numbers. These encoding systems also conflict with one another. That is, two encodings can use the same number for two *different* characters, or use different numbers for the *same* character. Any given computer (especially servers) needs to support many different encodings; yet whenever data is passed between different encodings or platforms, that data always runs the risk of corruption.

Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language. The emergence of the Unicode Standard, and the availability of tools supporting it, are among the most significant recent global software technology trends.



A document written in Bangla in old TTF Font using custom encoding

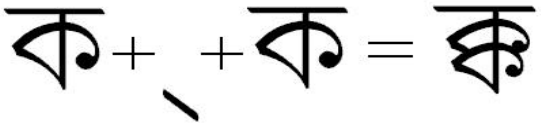
While Unicode seems like the likely answer for long-term adoption of Indic Language software, the immediate questions it throws up are also quite a few. As a standard it is still evolving. One needs to use OpenType fonts with Unicode which aren't that many in number as yet compared to earlier, proprietary solutions. And of course, one needs to look at the exist corpus of data stored in earlier encodings.

ISCII to Unicode converters are fairly well-available in the Opensource. But problems are often faced with converting documents that use TrueType fonts based on closed, proprietary encoding schemes.

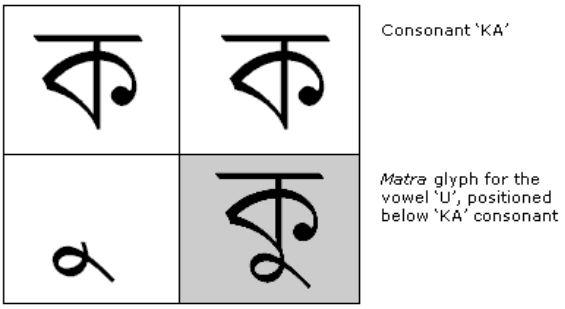
5. Open Type Fonts

The OpenType specification says – "In OpenType all the information controlling the substitution and relative positioning of glyphs during glyph processing is contained within the font itself. This information is defined in OpenType Layout (OTL) features that are, in turn, associated with specific scripts and language systems. Placing control of glyph substitution and positioning directly in the font puts a great deal of responsibility for the success of complicated glyph processing on the shoulders of type designers and font developers, but since the work involves making decisions about the appearance of text, this is the correct place for the responsibility to land. OpenType font developers enjoy a great deal of freedom in defining what features are suitable to a particular typeface design, but they remain dependent on application support to make those features accessible to users."

To understand how they actually work mention needs to be made of "two internal tables need to be introduced now. These are the GSUB and GPOS tables that contain instructions for, respectively, glyph substitution and glyph positioning. Glyph substitution involves replacing one or more glyphs with one or more different glyphs representing the same text string The backing string of Unicode characters is not changed, only the visual representation. These substitutions may be required (as part of script rendering), recommended as default behavior, or activated at the discretion of the user; they may also be contextual, active only when preceded or followed by a certain glyph or sequence of glyphs, or contextually chained so that one substitution affects another."



Example of use of GSUB table for glyph substitution



Example of use of GPOS data for glyph positioning

The list[9] below describes some to the GPLed Bangla OTF fonts. As is evident, the number of glyphs supported by each font varies often by a wide margin.

* Akaash – 409 characters (642 glyphs) in version 0.75
  Ranges: Basic Latin; Latin-1 Supplement; Latin Extended-A; Bengali
  OpenType layout tables: Bengali, Devanagari, Latin
  Family: Serif
  Styles: Normal
  Availability: Free download from The Free Bangla Fonts Project[10]

* Likhan – 286 characters (746 glyphs) in version 001.100
  Ranges: Basic Latin; Bengali
  OpenType layout tables: Bengali
  Family: Sans-serif
  Styles: Medium
  Availability: Free download from The Free Bangla Fonts Project

* Mitra Mono – 250 characters (324 glyphs) in version 0.70
  Ranges: Basic Latin; Latin-1 Supplement; Bengali
  OpenType layout tables: Bengali
  Family: Monospace (but Latin characters are not fixed width)
  Styles: Regular
  Availability: Free download from The Free Bangla Fonts Project

* Mukti – 197 characters (562 glyphs) in version 0.92
  Ranges: Basic Latin; Bengali
  OpenType layout tables: Bengali
  Family: Serif
  Styles: Regular, Bold
  Availability: Free download from The Free Bangla Fonts Project

* Mukti Narrow – 197 characters (562 glyphs) in version 0.92
  Ranges: Basic Latin; Bengali
  OpenType layout tables: Bengali
  Family: Sans-serif
  Styles: Regular, Bold
  Availability: Free download from The Free Bangla Fonts Project

* UniBangla – 184 characters (329 glyphs) in version 1.0

  Ranges: Basic Latin (non-alphanumeric); Bengali

  OpenType layout tables: Bengali, Devanagari, Latin

  Family: Sans-serif

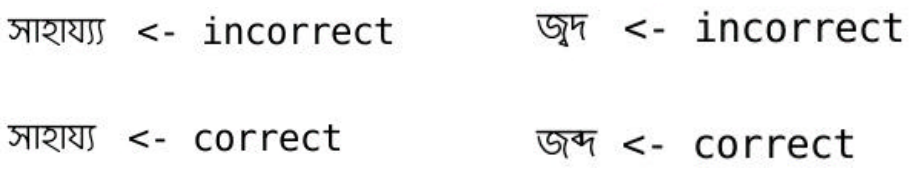
  Styles: Normal

  Availability: Free download from BanglaLinux[11]

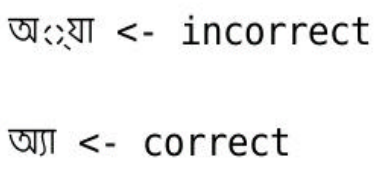
6. Engineering Pango – The GTK/GNOME Rendering Engine

Among the F/OSS Desktop Environments – GNOME[12] has witnessed the maximum work done in Indic L10N domain. The reason was simple. GNOME was the first one to provide support for Indic Scripts through its rendering engine – Pango. It is Pango which takes the Unicoded data, identifies the correct OpenType[13] font (assuming that one is installed), applies text layout processing rules that a language may need (e.g. sanjuktaakshars or conjunct clusters) and render it on-screen.

Pango has seen some major improvements in the last 18 months. Earlier releases had rendering issues affecting the Bengali *ya-phala, ba-phala* marks, handling of ZWJ/ZWNJ control characters etc[14]. These were addressed by members of AnkurBangla Project. The images below describe the ante & post situations.



The ya-phala issue #1 *The ba-phala issue*



The ya-phala issue #2

Later on other issues like the INIT feature[15] which is essential for the Bengali script were taken up and corrected by AnkurBangla developers. The image on the left hand side shows the INIT feature properly, whereas on the right side is the older, incorrect rendering of the script.



7. Collation Sequences

In simple words, collation sequences define the sorting order in any given language locale. A simple, implicit way, is through code-point ordering where the order is based on the numerical ordering of code points e.g. in ASCII A = 65, B = 66, C = 67 etc.

The reason why it assumes such significance in Indic Unicode-based systems is that several Indic language share the same script due to commonality in their source of origin (Hindi, Marathi, Sanksrit, Konkani), and others have scripts that are very similar (Tamil-Malayalam, Kannada-Telugu).

Unicode charts assigned to Indic scripts make no distinction between languages. Therefore, some charts use the same code chart for the following languages:

1. Devanagari: Hindi, Marathi, Sanskrit, Konkani, Nepali
2. Bengali: Bengali, Assamese, Manipuri
3. Arabic: Urdu, Kashmiri, Sindhi

The ISCII-88 standard (Indian language block of Unicode Standard is based on ISCII-88[16]) was based on phonetic commonality rather than correct sorting sequence. This distorted some traditional sorting conventions, and developers should not interpret the character sequence to be the same as their collation sequence. For example, though Hindi and Marathi use the Devanagari Unicode charts, the Hindi sorting sequence is not the same as Marathi. Similar situations exist in case of Assamese and Bengali which share the same script but a different ordering sequence. This requires that sorting be tailored to languages rather than scripts.

In multi-lingual Indic-enabled Library Information systems, the collation data is used/needed by sort/ search routines, and is therefore vital for their efficient operation. Collation data is defined in LC_COLLATE category in locale definition. A default approach (as currently done for Indic locales) is to copy iso14651_t1 table like

```
LC_COLLATE
% Copy the template from ISO/IEC 14651
copy “iso14651_t1”
END LC_COLLATE
```

The above table (stored in /usr/share/i18n/locales/iso14651_t1) doesn’t contain any data for Indic script ranges. So Indic sorting defaults to code points based, as in the Unicode charts. This behavior is defined by the Unicode Collation Algorithm (UCA), providing a default sort order that may be used only when no additional information is available. It can be found in the Unicode Technical Standard #10[17] document.

Code-point based sorting is good enough for simple scripts like Latin, most European scripts where the number of characters is less and do not generally combine with other characters (like *sanjuktaakshars* in Indic Languages). The disadvantage with code point sorting is its fixed forever, and if the encoded script (say Devanagari) were to be used with multiple languages (say Hindi, Nepali & Marathi), having different rules for sorting then it becomes difficult to accommodate them. Since many scripts are common across region/languages its imperative that collation sequence is independent of encoding.

8. Localizing Koha

Having covered the technical background on which the actual work of localization was/is being done, its time we turn towards Koha.[18] About this award-winning software, Joshua Ferraro – a leading Koha developer, says on his website[19]:

"Koha was built using Perl scripting language[20], MySQL Relational Database Management System[21], and Apache[22] Web Server, running on the GNU/Linux Operating System[23]; however, it has been ported to other operating systems (including Windows) and should be compatible with any system running an SQL database, a web server, and Perl."

Both of us had been engaged in the setting up of technical infrastructure of the library at West Bengal University of Technology[24]. During the first phase, it involved implementing a fully F/OSS-based ILS system using Koha which was completely browser-based. And during the second phase, we were engaged in setting up a digital library by extending Koha's capabilities. However, *that* is a different case study altogether.

It was during a discuss with librarians from another organization who were commenting about lack of Indic support in their present Library management system that the idea of localizing Koha came up.

9. The Requirements for Localization

To localize a browser-based software like Koha, a localizer has to address the following potential problem areas:

1. Making sure that the relational database management backend supports Unicode.
2. The server-side scripting engine (PERL in this case) must support regular expressions and strings with Unicode embedded.
3. The web server (Apache 2.0.48) must be capable of handling UTF-8 as a native CHARSET (character set).
4. The browser used on the client systems (Mozilla for us) must support Unicode and have complete rendering support for the script in question (Bengali) into which Koha is to be localized.

10. Creating the Localized Computation Infrastructure

With Koha being deploying on the Fedora Core 2[25] Linux platform, it was essential *not to install* the older version of MySQL database server which came with it. The latest 4.0.x range of MySQL server software which is Unicode compliant, was downloaded, compiled and installed on server.

Using the browser-based MySQL admin tool called phpmyadmin, the database was tested if it was handling Indic language strings properly. INSERT, UPDATE and SELECT SQL statements were all used in different permutations and combinations of Bengali, Hindi & Urdu data strings, to find out the stability of the platform.

Apache webserver was the next in line and we needed to make sure that the list of AddCharset directives in the apache configuration file was listing UTF-8 among others. This was to make sure that the webserver could serve out content (in this case Koha) using UTF-8 character set encoding.

We tested this by uploading a webpage with its charset attribute set to UTF-8 and then trapping the http headers to to see if we were being served the right charset content.

Once these tests proved OK and the system stable enough, it was time to pull in the latest Koha source code from its CVS server on sourceforge[26].

11. Translating Koha

After downloading the latest sources, we tried to follow the instructions provided for people interested in translating Koha into other languages. There are essentially two areas that one has to translate – the administrative intranet interface; and the Online Public Access Catalog (OPAC) interface that is visible to the general public once the Koha server is online. Presently it totals about ~4000 strings in all. Of course, this number doesn't include the online, context sensitive help documentation.

It turned out that the translation help documentation bundled with the sources was out-of-date and didn't give the desired output. Being opensource software, a quick look at the code inside brought about the following conclusion that instead of tmpl_process.pl in the misc/translator directory, one should use the new tmpl_process3.pl script from the same directory.

The tmpl_process3.pl script created the all-important POT (Portable Object Template) file which is the backbone of any translation effort on GNU/Linux. The script internally calls upon the GNU Gettext[27] Internationalization (i18n) library using PERL modules.

The next step was to rename the .pot files into the target language .po files – in this case, it was renamed from default_intranet.pot to default_intranet_bn_IN.po and css_opac.pot to css_opac_bn_IN.po following standard L10N conventions and start the actual translation. Once the translation is done the script is again used to generate the HTML templates in the desired language (i.e. one presently translated to).

12. Translation Methodology

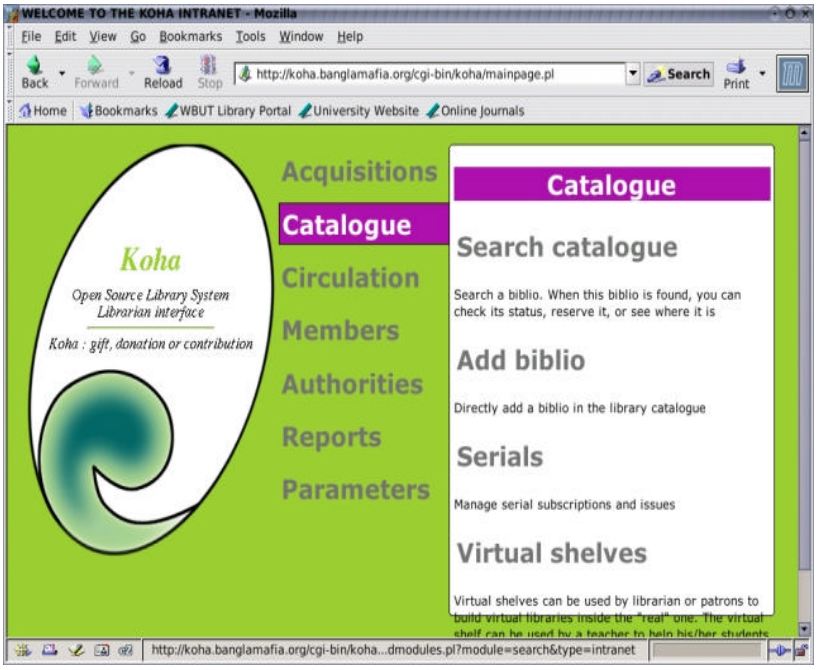
Our experiences during work on the AnkurBangla Project had equipped us fairly well in terms of selecting terminology or creating new ones wherever needed. But in case of Koha, a different route was taken. With the library movement having taken deep roots in West Bengal, there has been numerous efforts to created Library and Information Science Terminology in Bengali.

Using the reference library at Bengal Library Association (Bangiyo Granthagar Parishad) a a cross-reference glossory of terms from the domain was created and this was used extensively during the translation. Where terms from IT were encountered, the glossaries from AnkurBangla Project were used.

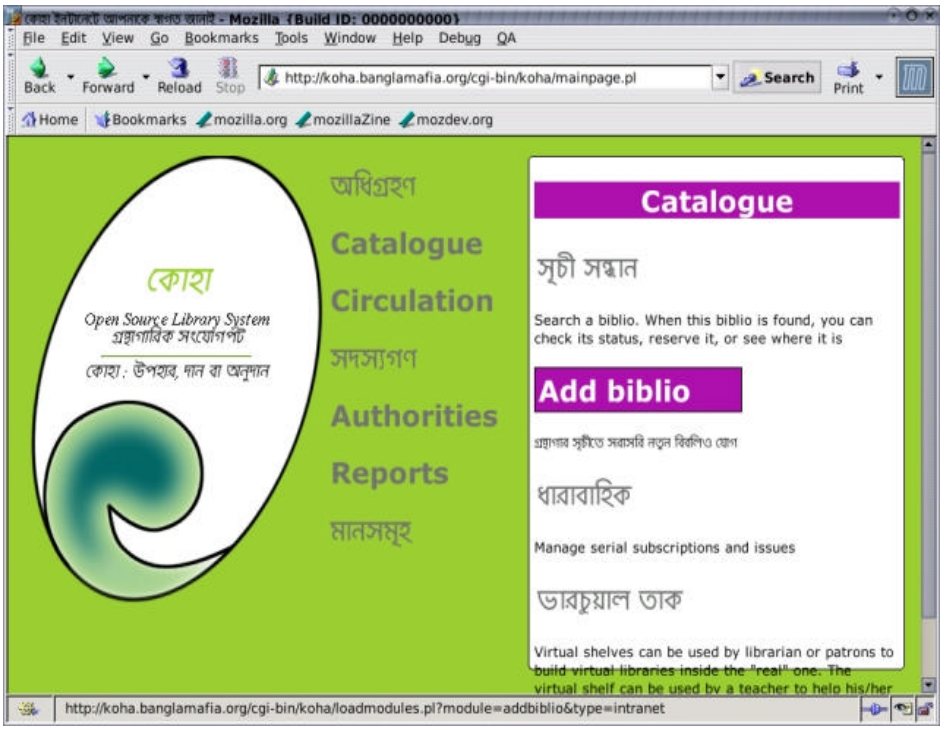
Use of existing terminological referencing was done to ensure that the localized interface of Koha along with it OPAC would be readily acceptable and meaningful to people accustomed to the Bengali terms.

13. Testing the Localized Koha

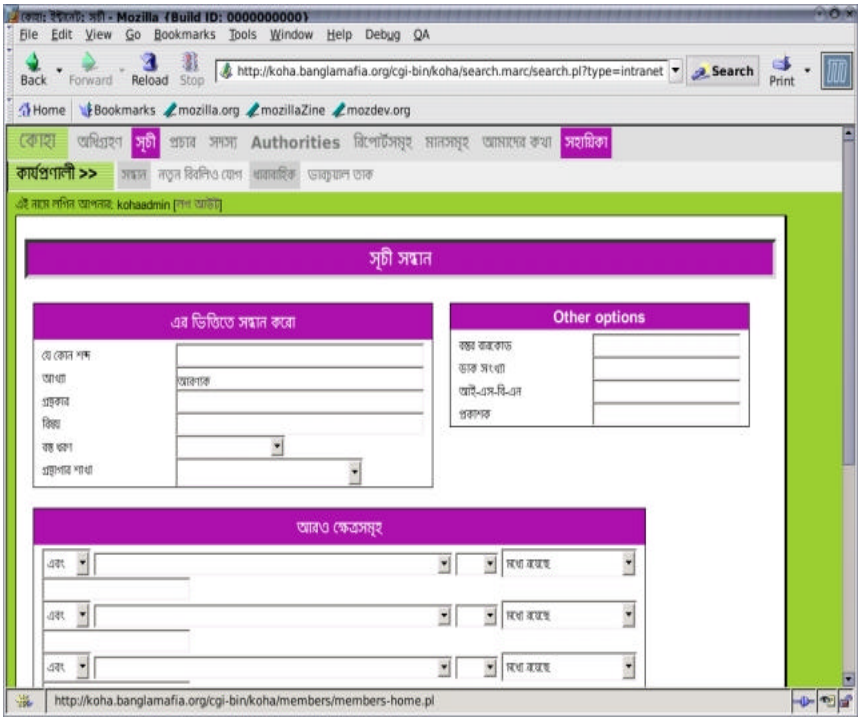
As the saying goes, the proof of the pudding is in eating it, so it was no different for a software which was being localized. Aside from translation, changes to the HTML templating code also had to be made as the existing code was forcing the CHARSET attribute to be iso-8859-1 and not UTF-8. This was causing loss of data in-transit as well as junked on-screen rendering.



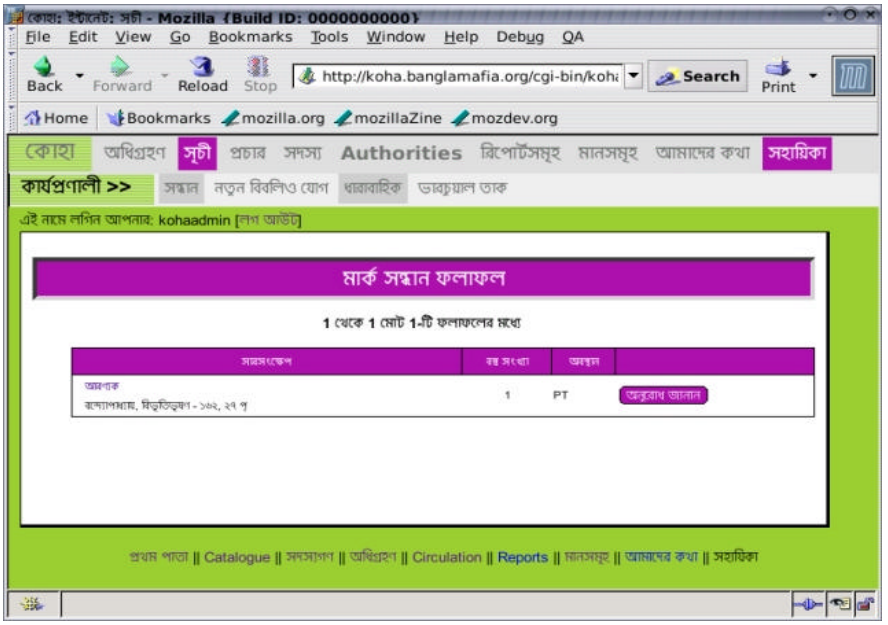
The default english interface of Koha's admin UI



Same interface but in Bengali (partially translated)



The Catalog Search in Bengali for a Title in Bengali



Search Result in Bengali (1 item found)

14. Status of The Project

The localization of Koha is presently in an advanced state. It is expected that it will be finished within the first week of January 2005, which is also when Koha 2.2 (the next version) is scheduled for release. It is expected that with the release of Koha 2.2, Bengali would be the first South Asia language to be supported in Koha. This project is completely non-funded and work done on it is on a volunteer basis by us in our spare time.

During the International Summit Conference held at New Delhi during 8 – 10 December 2004, dialog was initiated with persons from the SARAI[28] unit of CSDS (Center for Studies in Developing Societies), New Delhi, for beginning work in the Hindi & Urdu Localization of Koha from February 2005 onwards.

15. Other Similar Projects

In recent times there has been other efforts in similar directions. The team of Prasenjit Majumdar, Dr. Mandar Mitra & Rajdeep Mukherjee – all from ISI, Kolkata have been working on localizing the opensource UNESCO-funded Greenstone Digital Library (GSDL) Project. They have been using Windows XP to carry on the work of translations. Their work too is expected to be completed by January 2005. The Vidyanidhi project too has been investing in localizing DSpace, in order to acquire support for languages such as Hindi and Kannada.

16. Conclusion

Serious short-comings exist like the lack of collation sequence data on GNU/Linux OS platform. This affects the quality and speed of database searching which is presently based only on code-point ordering sequence. This problem is likely to get addressed with the next release of CLDR data (*See earlier reference to Locale data*).

Also, the sorting & searching algorithms, Natural Language Processing (NLP) which work well on linear Latin based scripts do not work quite the same way when applied to our complex, re-ordered text layouts. Research and development needs to be undertaken in the F/OSS domain to carry forward the task of Indic support in the areas of NLP.

Inspite of all these issues, today the ground for Indic localization of library information systems based on Free / Opensource platforms is well-prepared for the march ahead. It is a march towards a future based on standards-driven, internationally compatible systems offering equitable access to information to all who are in need of information in a language of their own.

17. References

1 Free & Opensource Software

2 For example, the Extremadura province in Spain once had the lowest PC:student ratio in schools in Europe, since the initiation of Project LinEx (mass adoption program of Linux Operating System PCs in education sector) it now among the highest. [Ref : http://www.linuxjournal.com/article/7908]

3 The Vidyanidhi Project [http://www.vidyanidhi.org.in]

4 Greenstone Digital Library Software [http://www.greenstone.org/]

5 DSpace Federation [http://www.dspace.org/]

6 www.kernel.org

7 Common Locale Data Repository (CLDR) Project [http://www.unicode.org/cldr/]
8 http://www.cdacindia.com/index.asp
9 http://www.alanwood.net/unicode/fonts.html#bengali
10 Free Bangla Fonts Project [http://www.nongnu.org/freebangfont/]
11 http://www.sourceforge.net/projects/banglalinux/
12 The GNOME Project [http://www.gnome.org/]
13 Microsoft Typography – OpenType Specification [http://www.microsoft.com/OpenType/OTSpec/]
14 Bugs in the Bengali rendering system of Pango [http://bugzilla.gnome.org/show_bug.cgi?id=113551]
15 Bengali Opentype Specification [http://www.microsoft.com/typography/otfntdev/bengalot/features.htm]
16 http://www.cdacindia.com/html/gist/standard/unicode.asp
17 Unicode Technical Standard #10 : Unicode Collation Algorithm [http://www.unicode.org/reports/tr10/]
18 http://www.koha.org
19 http://kados.org/LibraryScience/koha_at_a_glance.html
20 http://perl.org
21 http://mysql.org
22 http://www.apache.org
23 http://kernel.org
24 West Bengal University of Technology [http://www.wbut.net]
25 The Fedora Project [http://fedora.redhat.com]
26 http://sourceforge.net
27 The GNU gettext project [http://www.gnu.org/software/gettext/]
28 SARAI [http://sarai.net]

About Authors

Mr. Indranil Das Gupta has been an active user and evangelist for Free & Open Source software for the past several years. Aside from his vocation as consultant helping in managing the adoption and migration to Free & Opensource technologies, he has been active in the area of Localization of Free/Opensource Software in Indian Languages. As part of the IndLinux Group (www.indlinux.org), he is currently trying to create a model for productization of Indic F/OSS initiatives for mass-scale adoption using the L2C2 framework. Heis presently based in Kolkata.
E-mail : indradg@l2c2.org

Ms. Najmun Nessa is presently working at the West Bengal University of Technology, Kolkata (www.wbut.net) as an Asst. Librarian. She holds a Masters Degree in Library and Information Science from Jadavpur University, Kolkata. Along with Das Gupta she has been setting up a completely Opensource based Integrated Library Management System using Koha alongside establishing a Digital Library at this fledgling University. A founder-member of Indian Koha Interest Group, and she is keenly interested in Indic Language Cataloguing in digital formats. As a step towards that direction she has worked on localizing Koha version 2.2 to Bengali in collaboration with Das Gupta.
E-mail : najmunnessa@yahoo.com

Multilingual Computing in Malayalam : Embedding the Original Script of Malayalam in Linux and Development of KDE Applications

Rajeev J S Chitrajakumar R Hussain K H Gangadharan N

Abstract

Indic Language Computing can be fully realized only through embedding vernacular scripts in operating systems. With the advent of OTF (Open Type Font) embedding local scripts in OS compliant with Unicode has become a reality taking computing beyond word processing. Microsoft has already come to this field strongly by embedding Devanagari in MS Windows. Compared to the closedness of Microsoft OS, free and open environment of Linux is ideal for the early accomplishment of multilingual computing. This paper describes initiatives of Rachana team in embedding Malayalam script in GNU/Linux operating system. Modules are added for KDE with its rendering engine QT so that the original exhaustive character set of Malayalam developed by Rachana is embedded fully in compliance with Unicode. For the first time, prospects are open to create DBMS and information systems using Malayalam script. Computing in Malayalam language is being initiated in the true sense only now. The procedures set up by Rachana-GNU/Linux is highly beneficial to the goals of INFLIBNET in fulfilling a total integrated bibliographic control of Indian literature in their native scripts.

Keywords : Multilingual Computing, Localization, Unicode, Desk Top Publishing.

0. Introduction

Language is the foundation of all information systems. Language being the medium of information, there can be no information technology without language. Though IT has successfully assimilated voice and visuals in building up multimedia applications, secondary data indispensable for describing audio-video elements are coded using text. Later, data or information is retrieved and processed using the same text. Words and text are formed using the basic unit of written language called alphabet, character or *lipi*. Lipi in a language is the most systematized and standardized signs used to describe concrete or abstract concepts/ sounds. Without lipi there can be no information systems or information technology.

The computer system to input, render and process text has traditionally been Latin (Roman) based. Support for Indic languages would be implemented using custom rendering engines/shaping engines or using special cases such as Latin font encoding and custom keyboard input systems on top of the Latin based system. This however had several problems – either the custom keyboard input systems wouldn't be applicable to all application programs, or the font encoding would interfere with the correct rendering.

This led to the realization that in order to implement Indic Language solutions it would be necessary to embed the processing code into the Operating System itself, i.e., as first class citizens of the text world just like Latin based languages. Embedding means to allow input, rendering and processing of a language script in the traditional GUI widgets such as Textboxes, Labels and Buttons. Language computing in its truest sense, extending the capability of computing to all spheres of digital application, can only be achieved through this embedding to make the script of the language a 'live' part of the operating system as well as applications.

For the past 15 years word processing and DTP have been smoothly going on in all Indian languages. At the same time none of these languages has achieved a perfect DBMS in local script. We should admit the truth that information technology in India has not yet accomplished information system development in any Indian language! By embedding Indian languages in OS our languages will become as natural as English to the computer and we can make use of our scripts in all the conceivable fields of digital applications. Application programs could utilize operating system facilities for input, rendering and processing of the text and developers need only to provide the text in a suitable form known as encoding. Embedding would also allow more complex programs such as spreadsheets and database management systems to provide support for these scripts, in a uniform manner.

The work done by the authors in embedding Malayalam language falls into following categories:

? Fixing the character set of Malayalam

? Designing fonts

? Choosing an Operating System and GUI

? Coding for Embedding the script

? Adapting applications like text editors, word processors, spread sheets, Graphic utilities, DBMS and DTP to the embedded system.

Accordingly, the paper discusses the following topics:

? Malayalam Lipi and Rachana Language Campaign (Fixing the character set)

? Unicode and Open Type Font (Specifying the character rendering according to an international standard and developing Malayalam OTF fonts)

? Development of Rachana-GNU/Linux Distribution (KDE, OpenOffice, Scribus, etc.)

1. Malayalam Lipi and Rachana Language Campaign

It is from Tamil that Malayalam was born. Tamil is the most important among Dravidian languages. However, it is from the traditions of Sanskrit, the Indo-Aryan language, that Malayalam draws its rich diversity of words and compound alphabets (conjuncts).

It was in 1821 that Benjamin Bailey, a Jesuit priest, designed the first Malayalam metal types for the printing machine. From the basic 56 characters, he forged around 600 conjuncts in beautiful metal type. These letters adopted by Benjamin Bailey were in use for hundreds of years in Malayalam script. Later Herman Gundert designed and added several more conjuncts, and the Malayalam language came to possess 1000+ unique and rich type characters. These two pioneers were also authorities on comparative linguistics of Indian languages, thereby the design of Malayalam characters and types naturally encompassed pan Indian and local specificities. The people of Kerala recognize their language and have become the most literate of communities by learning and using this script. That this character set developed by them have survived and spread extensively during the past one and a half centuries shows their wide acceptance and faithfulness to the original script.

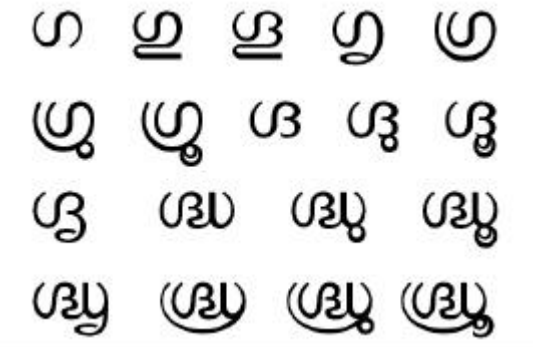
During early 1970s this sophisticated and systematized script language suffered a serious setback. This was the time typewriters started appearing on office tables. The demand for adopting Malayalam as the official language also became strong during this time. Considering the need for typing office files and

correspondence, the nearly 900 characters of Malayalam language was reduced to just 90 to fit into the keyboard of a typewriter. Even some of the fundamental vowel signs were excised. The most aesthetic and functionally superior Malayalam script was trashed without any logic or sensitivity to history. The stable structure attained by Malayalam script suffered cracks and several incongruities developed even in semantic level. This fatal programme was led by a government agency, the Kerala Language Institute and they even succeeded in implementing the truncated alphabets for producing the textbooks of primary standards in 1973.

When computerized typesetting (DTP) became popular in 1980s several software packages and fonts emerged. Several font designers, working in institutions outside Kerala and ignorant of Malayalam language, designed conjuncts casually generating contradictory character mapping which is not found in any other Indian languages. Integrated and stable character set of Malayalam language that survived for centuries became disarrayed and incoherent, and this non-systemization raised the greatest hurdle to attempt areas of digital computing other than word processing.

It was in response to this non-systematization of Malayalam that a language campaign under the banner 'Rachana' (which means 'Graceful Writing') was launched with the following objectives.

- The unique character set developed by a people over centuries transcending class divisions is not just a geometrical sign but the symbol of a culture.
- A language should be revised and modernized when deficiencies are observed in use and communication. And not based on the limitations of a transient historical phenomenon of a typewriter machine.
- The return to the original script is the only way to surmount the disintegration of Malayalam language in learning, comprehension, writing and printing.
- Modern information technology has made it possible to include and manage the exhaustive character set of Malayalam in any application. Rather than cut the alphabets to fit a machine, technology should be tamed to serve the language.
- The original Malayalam alphabets should be made ready for use in the modern language technology. The current information technology is advanced enough to embed the original exhaustive character set of Malayalam in all fields of digital computing.



Conjuncts formed by GA, DHA, DHHA, REPHAM and Consonant-Vowels, showing the exhaustiveness of Rachana character set

With the declaration of Rachana font comprising the exhaustive character set under GNU-GPL (General Public License) in February 2004, the efforts to embed the original Malayalam script in GNU/Linux platform has started.

2. Unicode

The Unicode is a universal encoding format designed to represent the symbols and script elements of the world in a uniform manner. The Unicode is a minimalistic encoding which includes currently all major scripts in use. The basic principle “Encode the characters, not the glyphs” denotes the minimalism of the Unicode encoding. By encoding only abstract characters to code points, the encoding would be able to reflect the semantics of the script rather than represent a mere number. This simplifies higher level processing such as EASCII to Unicode conversions and text stream to visual rendering.

In short the advantages of Unicode are listed below:

? It is a minimalistic encoding designed to represent all other encodings.

? Along with the OTF (Open Type Font) it allows development of languages with complex visual rendering requirements.

? It allows easy migration from an existing encoding scheme to the Unicode.

? The determination of script/code page can be done automatically in the Unicode, since each script is allocated a unique code block.

2.1 Emergence of OTF (Open Type font)

Fonts are the means by which characters in a language can be rendered visually on the screen or in print. It is one of the basic subsystems of text processing in the computer. Initially fonts were bitmap fonts. Soon, for the purposes of digital typography, fonts were designed with Bezier curves, which allowed arbitrary scaling of the font without loss in quality. The abstract curve representation of a character is also known as glyph.

For new languages that entered the computing arena, like Indian languages, the availability of only 256 slots in ASCII based systems made several constraints in the number of glyphs that could be designed in any given font. Combinations of basic characters known as ligatures or conjuncts could be designed and used by allocating a code-point to it. But the space available would remain as low as 256. This forces incomplete and disintegrated implementation of various languages (or families) like Indic, which need a lot more than 256 code-points to represent the entire repertoire. This is what happened in the case of Malayalam language when the attempts were made to accommodate its 1000+ original/ traditional characters.

OpenType Font (OTF) is the new technology with a variety of features that allow complete implementation of Indic languages satisfying all their peculiar characteristics. Microsoft and Adobe introduced it jointly in 1997 to meet the requirements of complex scripts and multi-lingual documents, as well as new techniques in rendering. Although OTF can be used with a variety of encoding, it is best implemented with the Unicode.

For each Unicode encoded character, the font designer can design glyph shapes for that character. Total number of shapes in the encoded and unencoded slots may come around 65,000 (i.e. 2^{16}). The unencoded set contains glyphs for combinations of encoded characters. In this way, an Indic text that contains mostly conjuncts can easily be represented and accordingly a font can be designed accommodating any number of glyphs.

An OTF can only be used advantageously in conjunction with a shaping engine (rendering engine) that is usually implemented in the Operating System (OS) or the windowing toolkit. The shaping engine provides text layout services, which transform a piece of text into glyphs from both the encoded and unencoded sets. This achieves the complex shaping requirements for conjuncts in Indic languages with the use of basic abstract characters of the Unicode.

This process of Shaping/ Rendering occurs as follows:

- The text is analyzed and broken down into segments. Segmentation is done at cluster boundaries.
- Then the string is mapped to a set of glyphs representing the basic characters.
- The basic glyphs are then further transformed on the basis of OTF features. OTF features are special tags associated with the unencoded-glyph shapes, which establishes correspondence between complex shapes and their basic components. The application of features can happen either by substituting the complex shape with a sequence of basic shapes, or by positioning the shapes over the existing shape.
- The final sequence of glyph shapes is laid out according to the requirements of the software.

For speed, interoperability and simplicity of laying out fonts, the system applies several optimization techniques. In general, the position of the OTF layout process is in the OS. For highly specialized applications that require fine use of text visuals, it may also be re-implemented for the particular system. Needless to say, the process should be standard on all systems so that there should not have different input methods and different visuals for the same text.

2.2 OTF in Malayalam

Indic language computing will be greatly benefited from the development of OTF technology. That the limitation of 256 slots is left behind by the 65000 code points alone makes OTF receptive to conjunct-rich Indic scripts.

An example of the application of OTF in Malayalam script will explain conjunct formation and its Unicode representation clearly.



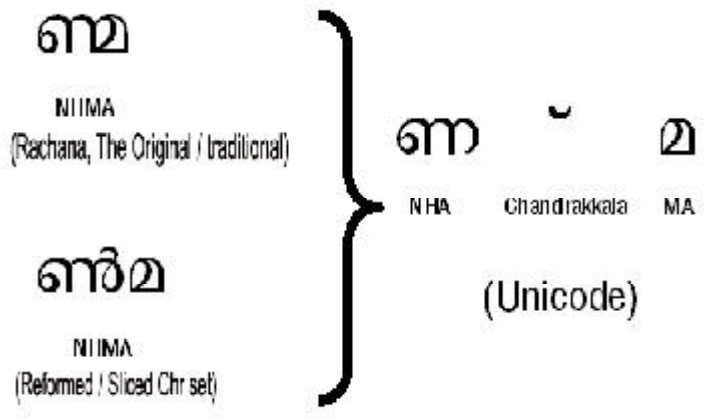
Consider the *NA Chillaksharam* which is represented by a single basic character in the Unicode block. As per the philosophy of the Unicode, glyphs that may be represented by sequences of basic characters will not be allocated a specific code point. So, the *NA Chillaksharam* is not given such a code point. In order to allow rendering of the chill form, the OTF provides a “Halant” feature. According to the Unicode standards, the *Na Chillaksharam* form is represented in the encoding by the sequence

NA + Chandrakkala + ZWJ

All the three characters that constitute this sequence are basic characters in the Unicode. The Halant feature is tagged on the *NA Chillaksharam* glyph that is placed in an unencoded slot and this is applied to *NA + Chandrakkala + ZWJ* glyph sequence. Such a sequence is indeed transformed into the *NA Chillaksharam* form when rendered on screen.

2.3 Unicode and legacy encoding

Considering the different character sets of Malayalam (Original and Reformed), an understanding of the Unicode encoding model can be derived easily from the following example:



In Rachana encoding, there exists a single code-point for the representation of *NHMA*, where *NHMA* is

NHA + Chandrakala + MA

In other encoding for reformed Malayalam such as ISM-Gist and SreeLipi, there does not exist a code-point for this character. Instead the character is rendered as the combination of two characters

(NHA Chillaksharam) (MA)

Unicode solves this problem by encoding only the basic minimal characters from which the compound characters are generated. *NHA, Chandrakalla* and *MA* have individual code-points and the complex shape *NHMA* is encoded as the basic code-point sequence such as

NHA (followed by) *Chandrakkala* (followed by) *MA*

Rachana as well as other encoding can easily be encoded using the Unicode. Along with appropriate OTF fonts and layout tables, it is possible to provide the exact rendering for both encoding schemes of Traditional and Reformed. OTF can reuse glyphs developed for older systems and allow migration to the Unicode system. In time, all documents using either encoding can be converted and the users can be fully migrated from legacy encoding to the Unicode.

3. Development of Rachana-GNU/Linux Distribution

Free operating systems like GNU/Linux and FreeBSD provide source codes and allow modification by independent developers. Due to this, it is possible to implement very good embedding solutions. GNU/Linux provides a plethora of GUI systems and libraries all of which can be modified to provide script embedding. In general, the two most common GUI platforms on GNU/Linux are KDE and GNOME. Embedding the script solution in, say, KDE would allow all KDE based applications to reuse the rendering system. Some of the well-known applications are distributed in either platform. Also, some systems like OpenOffice.org (a replacement for MS Office) have created their own infrastructure for rendering.

3.1 GNU/Linux

GNU/Linux is a total operating system available for a variety of hardware platforms providing a complete spectrum of functionality for users. GNU/Linux runs on platforms ranging from space satellites to home desktops and handheld devices.

A highly simplified view of the architecture of a GNU/Linux computer is shown below:

<table>
<tr><td colspan="2">Environments</td><td>Servers</td></tr>
<tr><td>Command Prompt
Shell (provides command prompt)</td><td>GUI
Desktop Environment (KDE, GNOME, etc)
Graphics and Windowing Toolkits (Qt, GTK, etc)
X (provides basic graphics)</td><td>Various Servers such as RDBMS, Web server, email server, etc</td></tr>
<tr><td colspan="3">System Libraries (glibc and lots more) + system utilities</td></tr>
<tr><td colspan="3">File system</td></tr>
<tr><td colspan="2">Non-hardware specific Kernel parts</td><td>Hardware-specific kernel parts
Computer hardware and peripherals</td></tr>
</table>

Layers of GNU/Linux

The figure shows the layered concept of OS design, where each layer uses the facilities available in the lower layer to provide facilities to the higher layers.

At the very bottom lies the computer hardware and peripherals such as CPU, RAM, CDROM drives, hard disk drives, printers, mouse and display systems, along with the kernel. The popular kernels include Linux and Hurd. It is important to note that Linux is just a kernel and it is useless to an end-user without the rest of the GNU system, which lies in the higher layers.

Above this layer lies the file system that provides facilities for handling files and directories. The Linux files system can be of several types, each with its own features. Some file systems provide an automatic backup capability called journal, which can prevent all kinds of data loss (except for disk drive failure).

System libraries are the functional units, which provide a variety of services (access to the file system, networking capability, password authentication, etc). System utilities are common programs used for listing directories, performing common system tasks, etc.

Above this layer is where the actual end users do their works. There are two environments: the command line environment (Shell or CLI) allows user to type commands and see results. GUI (Graphical User Interface) is the commonly used interface by users, where there are windows, buttons (widgets), the start menu, graphical clock and other applications like office suites and web browser. It is in the GUI that Indic language embedding is most useful to the end user.

Servers are special programs that continually run in the background and provide specific services such as the web server (useful for making websites) and the email server (for storing and transporting email). Developers make applications using servers and end users make use of GUI tools to access these applications.

GUI is split into three layers: the basic X system allows drawing of lines, circles, etc. and takes care of various hardware events. For example, when the mouse is clicked the X system transfers this event to the application along with the position where the mouse is clicked. X is also in charge of display and input/output hardware when running GUI applications.

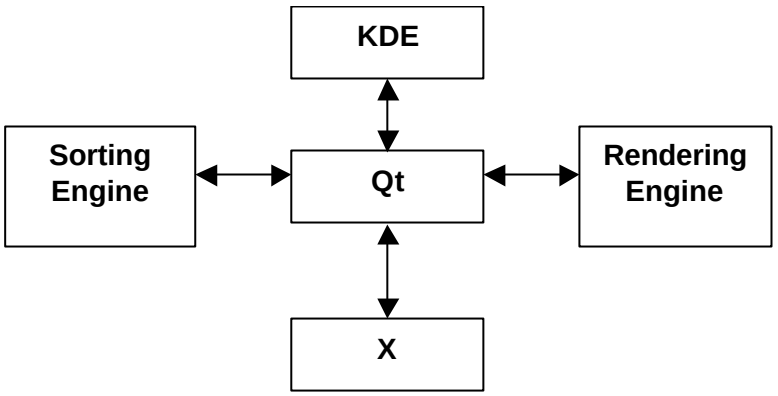
Above the X system lies the graphic toolkits: they provide an easy access to build GUI applications. Unlike the X system, they are more fully featured for developers and provide a unique look and feel to the applications. The graphics and windowing toolkits utilise the services of X to perform its work.

At the top lies the Desktop environment which provides the common concepts such as file manager, desktop, task bar, graphical clock, start menu, common interface for configuration and easy access to the various applications on the system. The Desktop unifies the applications and provides an area with common facilities that can be used by the applications to provide a unique experience to the user. There are several Desktops available on Linux, most popular being KDE and GNOME.

3.1 Development of KDE Applications

Authors of this paper had decided to develop Malayalam embedded system on the KDE platform. It was chosen not just because of the high quality of the existing system, but future development of domain-specific systems such as Library Management systems would be much easier using KDE and its underlying Qt GUI library.

Authors also decided to adopt some non-KDE applications for implementing Malayalam such as OpenOffice.org and Scribus (a professional DTP tool). Although Scribus is a Qt application, it is considered separately, because it has its own text rendering infrastructure due to its need for highly accurate placement and rendering of text to achieve tight typographical control.



The above diagram shows the layered relationship between various parts of the GNU/Linux GUI system. At the bottom is the X system, which takes care of handling the monitor screen and hardware for display. In the middle, Qt provides general services for drawing widgets like Textboxes, Labels and Buttons. At the top is KDE, which provides the common desktop.

Qt is the rendering engine for performing the task of taking an encoded text string and rendering it on the screen. Obviously, Qt is the place for inserting solutions for embedding. The authors proceeded to analyze the requirements of the script rendering engine, along with the script properties, font requirements and encoding formalism. Based on this analysis, the rendering engine component was completed.

Some of the applications that were tested with the new rendering engine are given below:

? Konqueror : Web browser, which allows reading of Malayalam websites

? Kmail : email client like Outlook Express and Eudora

? Kword : Word processor (part of the Koffice suite)

? Kspread: Spreadsheet program (part of the Koffice suite)

? Kpresenter : Presentation software (part of the Koffice suite)

? Kchart: Statistical chart program (part of the Koffice suite)

? Kformula: Formula editing program (part of the Koffice suite)

? Kivio: for flowcharts and other charts (part of the Koffice suite)

? Kolourpaint: a simple yet sophisticated paint program

? Kedit: KDE text editor like Note Pad

? Kgpg: Message encryption program

? Kopete: Instant messenger (with Yahoo, MSN and Jabber protocols)

? Konversation: an IRC client

? Korganizer: Calendaring, scheduling and journal software

? KAB: Address book for storing contact information

? Knoda: Database development system

3.1.1 Knoda

Knoda is particularly useful to build database applications. In the simple usage, it can be used to make forms and connect the forms to databases for data entry and retrieval. In the advanced usage mode, one can write full programs in Python and connect them with the forms. Python is a fully featured language, easy to learn as well as having a very powerful and fast interpreter. As the backend database Knoda may use PostgreSQL server or MySQL server as well as the embedded database SQLite. SQLite allows the development of read-only and embedded databases for CDROM publishing, or for standalone software without the need for servers.

3.2 Office Suites and DTP

The independent software such as OpenOffice.org and Scribus also follow a similar approach to rendering. OpenOffice.org is a full office suite very similar to MS office suite. Scribus is a DTP package in Linux similar to QuarkExpress. KDE architecture can be reused for the implementation of a script under these systems as well. In Scribus such an infrastructure did not exist at the time of the development, and it was up to the Rachana team to provide this system, thus allowing all languages (including non-Indic scripts) to be implemented in the acclaimed DTP tool. When compared to OpenOffice, Koffice (the native office suite of KDE) requires lesser computer resources such as memory and CPU power. Combined with the ease of use and power of the Qt toolkit, Koffice is a much better candidate for further developments.

3.3 Future Directions

Future direction of the Rachana-GNU/Linux distribution with respect to the Library Management community can take place in several simultaneous but interconnected directions. Various tasks to be completed do not allow a linear view, yet we present it as a list.

OS :

- Extension of embedding to GNOME/Pango
- Research on innovative methods for text entry specific to the complexities of Indic languages
- Extending search mechanisms of KDE and GNOME to the Digital Knowledge Archive

Domain-specific Software :

- Support tools for Indicnet
- Content Management system customized for libraries (including support for Multimedia)
- Development of OPAC, either by extending existing ones or developing new software taking into account the diversity of Indic scripts and local requirements
- Storage of data on a Grid
- Support for intelligent queries on structured data in the Digital Knowledge Archives (e.g., structured query on patents and legal codices)
- Development of computer aided teaching learning system for teacher-student community, especially in K-12 environments

Knowledge Initiatives :

? Development of Indicnet which is a Wordnet like knowledge-base of Indic words, meanings, synonyms, antonyms, homonyms etc

? Phased implementation of a national-level Digital Knowledge Archives available to all over the Internet/national network

? Construction of metadata repository

? Construction of Grid infrastructure

3.4 Bibliographic control of Indian Literature and INFLIBNET

When using a comprehensive library automation package, the present practice is transliterating local bibliographic data in to English, which is very often inconsistent and misleading. Implementation of localized KDE/Qt allows the development of local language bibliographic systems that are highly integrated with the existing infrastructure. Library automation and creation of bibliographic databases at the national level can be successfully carried out only if DBMS supports all the Indian languages. Rachana-GNU/ Linux is a step towards this direction.

Some of the avenues for future developments in the Library Management systems on GNU/Linux, as envisaged by INFLIBNET, could include the long wished-for national library integration system. If the SOUL package of the INFLIBNET is enabled to accept and process data in all Indian languages it will open up immense possibilities. This will facilitate inputting data and searching information using native scripts. Such software will be ideal for creating online as well as CD-based Indian National Bibliography.

Other schemes may include development of a national digital library resource with multimedia, which would allow final integration of knowledge resources from all the languages all over the country through a single window. The authors hope that the proposals in the future directions (section 4.3) will be carried out as part of the INFLIBNET programmes to realize the ultimate goal of total bibliographic control of Indian literature.

4. Conclusion

Microsoft has already come to the field strongly by embedding Devanagari in MS Windows in 2004. They have declared their plans to do similar embedding for all other Indian languages. Since the scripts and syntax of each Indian language has got its own peculiarities and complexities Microsoft's task will not be as easy as they expect. On the other hand linguists, typographers and IT experts can be easily assembled in every Indian state and can be effectively mobilized for embedding their scripts in GNU/Linux that offer open source code supported by an international fraternity of developers. Compared to the closedness of Microsoft OS, free environment of Linux is ideal for the early accomplishment of multilingual computing. This is in addition to the excellent facilities already well established in GNU/Linux such as its exceptional networkability, security and robust operations.

5. Acknowledgements

We would like to thank Mr Joseph Sebastian (BTC Engineering, Kuwait) and Dr. Varghese Paul (Cochin University of Science and Technology, Kerala). Sincere gratitude to Dr. Mammen Chundamannil (Kerala Forest Reseach Institute) and Mr. K. Raveendran Asari (former Librarian, Mahatma Gandhi University, Kottayam) in preparing and editing the paper.

6. References

1. K. Desktop Environment : htt://www.kde.org, http://www.koffice.org
2. Qt Toolket : http://www.trolltech.com
3. Unicode Consortium : http://www.unicode.org
4. OpenType Specification 1.4 : http://www.microsoft.com/typography/

About Authors

Mr. Rajeev J. Sebastian is completed his B.Tech from CUSAT.
Email : rajeev_jsv@yahoo.com

Mr. Hussain K H is working as a Documentation Officer in Kerala Forest Research Institute, Peechi, Thrissur, Kerala
Email : hussain@kfri.org

Mr. M. Chitraja Kumar is working in a Kerala University. Thiruvananthapuram, Kerala
Email : chitrajakumar_rachana@yahoo.com

Digital Mapping of Area Studies: A Dynamic Tool for Cultural Exchange

Chitra Rekha Kuffalikar D. Rajyalakshmi

Abstracts

The paper stresses the importance of Area Studies by visualizing the digital technology as a strong effective tool for Digital mapping. It emphasis the importance of rare documentary sources related to Area studies with special emphasis on the 'Area research on Nagpur'. The article highlights the Digital mapping projects at various levels and gives a bird's eye view of the work being carried out at the local level.

Key words : Area Study, Archives, Digital Mapping, Digitisation, Preservation

0. Introduction

Information networking in the recent times is passing through an important phase, when the digital technologies are offering a stunningly efficient new means to store, and retrieve information from networks, which in turn generate enormous benefits to higher education, and research. The digital revolution in telecommunication and Computing has unified and integrated number of hetrogenous services, evolving digital systems which permit easy signal reconstruction, thereby enabling better quality services. The move to an image based communication environment implies mixing of varied styles of communication with a steady synchronised flow of information. Hence, in this new technological era Digitisation has changed the entire concept of the ways scholars, Students, and general population find, use and disseminate scholarly information.

1. Area Studies : Their Importance

Every region, and their precincts pass through a series of events, which are embodied and recorded in various documentary sources. It is only through these repositories, that every society is able to visualize its past cultural heritage and contributions in various fields of human endevour. Thus, the past acts as stimulus for present as well as the future advancements in the society. Prabhakar keshav sardeshmukh Maharaj very rightly says that "Money does not last, Empires disappear. Nothing else, but knowledge lasts, eternally."

Area Studies present the varied aspects of the growth and development of a particular geographical area, and documents relating to these, act as the primary sources of information. Mapping Area gateways, classify the local occurences, and events, and in turn, serve as a key source of reference and citation for researchers, They also deepen one's interest in the Art, Architecture, Culture and Heritage studies. The Area gateways serve as the road maps of cultural exchange for specialists, scholars and general masses alike.

1.1 Documents for Area Mapping : Their value

Documents relating to a particular area exhibits a wide gamut and variety ranging from Oral tradition, Archeological sites and remains,

Artefacts, Cultural material, Manuscripts of Individuals and Institution, to films and CD-ROM'S. Each document has a definite value assigned to it, and plays a major role in deciphering the past by linking the same to the present, and future to pass on this cultural knowledge, from one generation to the other.

Among the key multilingual documentary sources, mainly classified as the Manuscripts, printed non-conventional and neo-conventional forms, are the Maps, Photographs, Dairies, chronicles, Letters, Memoirs, Reminiscences, Speeches, Travelogues, News papers, and Archival records from Corporate Bodies, Government, and Educational Institutions, Records of Temples, Churches, Mosques etc., Business Houses, Radio programmes and films released on different occasions, videos, CD-ROM'S etc.

Among the primary sources, and those widely used documents for any 'Area research' are the Archives, Individual records, Manuscripts in its various forms and languages etc., and hence, need the utmost care and preservation.

1.2 Archival Materials : Their preservation

According to Fr. Vijay kiran and Ramesh Babu "The mission of preserving historically valuable items encompasses three archival functions, Arrangement, preservation and security. Arrangement, is the organization of archives or manuscripts in accordance with accepted professional principles; preservation is both the protection of records from physical deterioration, and damage; and the restoration of previously damaged items; Security enhances the protection of the documents from misuse and corruption."(1)

1.3 Preservation Reformatting

According to "Heather Brown" Preservation reformatting, the copying of information from one form to another is a key international Preservation Strategy, it saves the original from further damage. It also provides a copy, that can be used long after the original has crumbled to dust."(2) Among the widely used preservation reformatting techniques include transcription, photocopying, conventional photography, microfilming and digitization.

1.4 Reformatting using Digital Technology

In the words of 'Abby smith' "Digitisation is an excellent medium for access to information. Digital surrogates can make the remote accessible, and the hard to see visible. They can bring together research materials that are widely scattered about the globe, allowing viewers to conflate collections, and compare items that can be examined side by side solely by virtue of digital representation. Through image processing, one can even transcend the limits of human eye."(3)

Hence, one can view digital reformatting as an excellent tool for providing increased access to information by a global audience. The user is usually not restricted to a straight linear Sequence of a microfilm format. Digitisation is a superior option for reproducing half tones, photographs, and coloured originals. Enhancement of images by proper cleaning etc., make the faint marks on the original legible.

2. Digital Area Mapping Projects : Efforts at the International level

A close 'review of the literature' on the above projects give a clear indication, that the Western countries have taken a lead in starting digital initiatives and according to S. M. Shafi(1) Digital scriptorium, MASTER, Oxford, Lund, Bodleian, N.L.M. and EMSI etc are some of important projects which have been initiated to preserve, and make the manuscripts accessible to the masses." Jean Mare-comment(2) has spoken of two key digitisation projects related to the Area studies namely "Comintern Archives for Russia" and 'Albania National archive Project.'

The comintern archives contain finding aids with reference about 2,40,000 files, half of these are personal files. The Principal Language of the files is Russian, and a great proportion is German. The archives

representing 60 countries in 90 languages have around 10 to 12 million documents. The project has been completed, and a joint venture between a Russian IT company, and a Dutch publisher is in offing, to provide an internet website in Moscow with free access to database, and fees for the images. The 'Albanian Central Archives' in Tirana have finding aids with references of about 6,60,000 files and 8,40,000 documents. The language being predominently Albanian. The project makes accessible all its resources in the reading room of Albanian Central Archives. Researchers are able to locate instantly the reference number of all files and material relating to their topic of interest.

Among the other country initiatives are the Chinese Academy of Sciences, Cuban Heritage online, Unesco's, Memory of the world, Indiana University's Indiana Studies programme, The Thames Pilot Project etc. Harsha Parekh(3) mentions about the American Memory project of LOC(http://memgry.loc.gov/), British Libraries treasure collections (Turning the page technology), Gutenberg Bible or Magnacarta (http://www.bl.uk/collections/treasures/digitisation.html) projects such as these have increased understanding of the past, strengthened national pride and identity, and informed both the far flung diaspora and generations to come.

2.1 Area Study projects : Efforts at National level

Though, the initiatives of developing digital Area Study projects in India, are either lacking when compared to their western counterparts, or in their primitive stage of development, there have posotively been some concrete efforts to work on such projects, specially in areas which have had a rich cultural heritage, and their documentary collection is distributed in private and public institutions in variety of media, languages, scripts, collection, sizes, in different conditions.

Notable few are the 'Makhtootat multilingual multiscript Digital Library project' of the University of kashmir which explores a variety of research and development problems related to multilingual multiscript medieval manuscripts, and is working on to build a system, that supports research and developmental activities. Digital Library initiatives in Nepal is also worth mentioning here.

Several University Libraries and Institutes of repute like the Punjabi University, Patiala, Mysore University, Pune University and C-DAC, INTAC, Indira Gandhi National Centre for Arts etc. are examples of National program that aims to preserve the knowledge held in millions of Indian Manuscripts for the benefit of future generations.

3. Projects at the Regional and Local Levels

Of late, a number of agencies at state, and Local levels are understanding the growing importance of such area studies, and have expressed deep concern over the loss and decay of the documentary evidences. Institutions, and Individuals both are trying their best to make first hand efforts in preserving these historical evidences using modern technologies.

The vidarbha Heritage Society, Vidarbha Nature and Human Science Centre, Vidarbha Research Society are surging ahead by taking a leading role in applying digital technologies for the documentation of Heritage sites and the Archives related to these. Noted Historians, Nature lovers and several Corporate agencies are also working in this direction.

3.1 Efforts at Individual level

A Ph. D. level project was undertaken by the author as a mark of tribute to the City (Nagpur) on completion of its Three Centuries in the year 2002. The 'Bibliographic Area study' presents a study based on all the

scattered documentary sources on varied facets and brings them together in a standard format. The data collection complete, and the project in its writing stage would be submitted early next year. The city has passed through several reigns in the past three centuries, and has generated documents in various languages and forms scattered in almost every part of city and region with Individuals and Institutions.

3.2 Database on the Area Study

Three kinds of separate databases have been generated. The first being the *Bibliographic data base* which runs into thousands, are on Ten major aspects of the city, presented in the Standard MLA format. The Textual data base presents a study based on cited works and bibliographies. The pictorial data base presents the pictorial orange city of Nagpur in various times, their rulers, monuments, etc with descriptive tags and index. These have been photographed, and scanned on CD'S.

3.3 Documentary evidences : Their specialities

3.3.1 Dr. Bhausaheb Kolte alias Nagpur University Library

Started in the year 1923, it has a very rich collection of 3,43,903 books; 32,500 back volumes of periodicals; 11,000 thesis, 15,000 manuscripts etc. in the library. It is a fine blend of a rare collection of manuscripts, books, maps, reports gazetteers, Thesis etc in various languages predominantly Hindi, Marathi, Modi, Urdu Arabic, Persian, and English. A sizeable collection is excluisevely on Area Studies, predominantly the manuscripts. The Nagpur University with its own publication unit has published more than Hundred titles and a number of them are on different aspects of this area.

3.3.2 Departments of the University

The University presently has around 46 Departments with a centralized PG Campus and Mahatma phule campus Library. Each Department with its seperate collection adds to the documentary sources. Quite a few Departments which have existed for more than 65 years now have a good collection only on this Area (Nagpur).

3.3.3 University affiliated colleges

There are around 120 colleges in the city, and a few them have already completed their centenary. These have rare documentary sources only on the Area studies, and a major proportion has already been lost and decayed for want of proper preservation measures and techniques.

3.3.4 Public Libraries, Corporate bodies, and Government Institutions

A number of public libraries having a rich history of more than Hundred years, have also their major share to contribute in building a strong Bibliographic database on this Area, but once again, due to lack of proper preservation techniques, the collection is in shambles and needs urgent restoration. Corporate Bodies and Government Institutions, have generated their own records which add rare Bibliographical value in these studies.

3.3.5 Central Museum and its Library

The third largest museum in the Country, it exhibits rare archeological findings, artefacts and rich cultural Heritage of this region. Quite a few galleries are devoted exclusively to highlight the various facets of this region in different times. The library has a rare collection of 3983 documents.

It has one of the finest collection on the Area Studies and with the modern technologies showing its lasting impact, the museum is all set towards digitizing its records, and has launched its independent web site “central museum nagpur.com.” recently.

3.3.6 Individual collections and Records

Renowned families, and their generations who have witnessed the fall, growth and development of this area have collected their own records in the form of chronicles, sanads, diaries, etc which have a very significant value in the Bibliographic data base. Collections of Artefacts, and rare material of this region in the individual collection adds authenticity to the data.

3.3.7 Vidarbha Archives : Key multilingual sources of the Area study

Set up in the year 1971 at Nagpur, the record office has various documents related to vidarbha, and a major chunk of its collection are exclusively on Nagpur Area in mixed script of Modi, Persian, English and Marathi. The bundle of correspondences with several files and documents are today extremely old and worn out. The collection dominates thirty one major departments with records from 1906-1956, and files ranging from 6 to 9907. Several Archives have been transferred to Record office, Bhopal after the reorganization of states in 1956.

Among the private collection, related to the Nagpur Area, the following are notable few

Table 1 private Archival collections in vidarbha Archives

S.No.	Collection	Span (years)	Details
1.	Subedar Collection	1750-1900	4 Bks and bundle of papers in Modi Script.
2.	Chatte Collection	1750-1800	Land records, Rewards, 400 documents etc. in Modi script relating to Bhonsles of Nagpur, Mehekar government, gifts of Holkar, History of Bhausaheb etc and 61 Books of Historical importance
3.	Adgaonkar DeshpandeCollection		1798-1897Eleven bundles in modi Script. Records are related to Income sources, Land records, and accounts of these families.
4.	Khaparde Collection	N. G.	27 budles in mixed Script of Eglish, Marathi and modi. Life sketch and khaparde’s contribution in different fields.
5.	Rambhau Danaji Patil	1905-55	18 Diaries in Marathi and modi script written by shri. R.D. Patil
6.	Pantawane Collection	N.G.	Religious prayer books written by Sridhar belonging to the ancestors of Shri. P.V. pantawane

Source : unpub. Catalogues and records of Vidarbha Archives

N.G. – Not Given

Apart from the above, several private collections of the region (Vidarbha) are also included in the Archives along with 13Rolls of Microfilms in these private collections.

3.3.8 Akashwani, Doordarshan Kendra and South Central Zone Cultural Centre.

Among the Non-Conventional documentary sources, all the centres mentioned above generate special programmes, and materials on different occasions to boost up the art and culture of this region. The video cassettes on the Nagpur Tercentenary, conventional photographs, films, audio and video cassettes, CD'S etc. form a major chunk of this Collection. Several sports agencies have added new dimension to these and quite a few Health care Centres have also contributed a sizeable share in building documentary sources for this region.

3.3.9 Web Sites on Nagpur Area : A Modern tool of Information retrieval

With the growing impact of technologies, Web Sites launched by Individuals, Organisations and Institutions have become a handy, reliable tool for the area study. Among the popular web sites on 'Nagpur' are the 'nagpurkhoj.com.' geocities.com., etc.

4. Limitations of the Study

The scope of this area study on Nagpur being sufficiently large, the coverage of the data bases cannot be comprehensive. Looking to the limitations of time, manpower money and other constraints, digital reformatting of the documentary sources for the present Study has been restricted to few selective areas (for ex. The pictorial representation of Archives, Monuments etc of this region). Conventional photography, scanning of photographs from the documentary sources on CDS microfilming etc. have been used on a large scale.

5. Directions for Digital Area Mapping : Some recommendations

There is an urgent need to take up such large scale projects with a proper funding base, either in the form of Minor or Major Research Projects preferably by younger professionals in coordination with experts (Individuals and Institutions) from different fields.

5.1 In-house Computerised Bibliographic Databases

The Scatter of collection in different parts of the city in different libraries, urgently need creation of Bibliographic Databases. The In-house Computerised databases exclusively on the Area study of Nagpur developed by all the agencies mentioned above would be extremely helpful in bringing all the scattered bibliographical sources together.

A recent survey has also revealed, that almost all the important Government agencies like (AnSI, ASI, IBM, GSI etc) and a number of educational institutes, and other organizations, and centres of repute have computerized their In-house operations. The main challenge would be to create an on-line Database on Website. Since all the agencies may not have the internet connection, the alternative means to access resources in digital form would be CD-ROM. Establishment of Locally generated Digital libraries, on-line news services etc would be extremely helpful in giving a meaningful direction to such Area Studies.

6. Issues of Concern in digital Preservation

Digital access, though dynamic and far reaching have several issues of Concern. 'Harsha Parekh'(1) has voiced her concern about the managerial matters like misuse of technology (piracy, distortion and plagiarism, IPR etc), Costs of digitization (financial viability from commercial point of view), resource mobilization, Skilled manpower etc. 'Masoom Raza' and 'R.L.Arora'(2) while speaking of digital preservation issues have opined that "Adopting appropriate selection guidelines, securing archived items from intentional or unintentional alteration, recognising the creator's responsibility etc" would be issues of serious concern in the digital environment."

7. Conclusion

Digital technology has opened new doors for an effective collaboration between libraries. It has also boosted the energies of individual libraries by bringing them together to offer powerful services based on technological expertise.

Understanding Area Studies in the real sense, and giving them due importance has almost become mandatory in the recent times. The importance of manuscripts and other documentary sources relating to the history and development of any area are invaluable sources for the authentic mapping in any Area study, but due to lack of effective organizational support, these studies have not caught up with recent times. Preservation of documentary sources has also not been understood in proper sense. We need to pay great attention to long term preservation, and the use of digital data as an effective tool for data exchange. It is only then, that the Digital Mapping of the Area Studies could be projected as a dynamic tool for cultural exchange and carry the same for the benefit of future generations.

8. References

1. Aiyepeku, W.U. Geographical literature on Nigeria, 1901-1970 : An Annotated Bibliography. New York:G.K Hall and Co., 1974. 214.

2. Barry, Jeff. "Digitising Cuba: Bringing the cultural Heritage of an Island Nation Online".Proc. of the intl. Conf . on Digital Libraries, V1, Feb.24-27, 2004: New Delhi : TERI, 2004. 167-71.

3. Handa, C. Nagpur Guide. Nagpur: Handa Travels, 2002.

4. Jeevan, V. K. J, and Dhawan, S. M. "Problems in Transition to a Digital Library". DBIT 22 (2002): 13-19.

5. Kaur, Trishnajit. "From Manuscripts to Digital documents'.Proc. of the intl. Conf . on Digital Libraries, V1, Feb.24- 27, 2004: New Delhi : TERI, 2004. 437-441

6. Kuffalikar, C. R. "Preservation for Posterity Need to Digitise the Local History Collection': A Nagpur Tercentenary special Proc. of the natl. Seminar on Impact of Digitisation on Development on Development of Information Professionals, Feb.28-Mar.1;2003:Ed Ashwini A. Vaishnav, and Shashank Sonawane. Aurangabad : BAMU, 2003.CD

7. Kuffalikar, C. R."Rejuvenating Local Histories : Preserve or perish Proc. of the intl. Conf . on Digital Libraries, V1, Feb.24-27, 2004: New Delhi : TERI, 2004.V.II

8. Kumar, P.S.G. Archive as a source of Information Indian Encyclopedia of Library and Inf. Sc. VII AF-AR. New Delhi: S.Chand and Co., 2002. 390-396.

9. Kumbargoudar, P.K., and Mestry, Mamta." Ideals and Illusions of Digital Libraries". University News 40(2002):5-8.

10. Mahajan, S. G. Pune City :Its History, Growth and Development. : 78 to 1998. Pune :Mansanman Prakashan, 2002.223

11. Mahajan, S. G. Pune Sharacha Dnyankosh. Pune : Dnuankosh Pratishthan, 2004.

12. Nagpur University, Nagpur : A profile. University News 40 (2003): 9-10.

13. Pradhan, Mohan Raj. "Digital Library Initiatives in Nepal". ".Proc. of the intl. Conf . on Digital Libraries, V1, Feb.24-27, 2004: New Delhi : TERI, 2004. 179-184.

14. Raju, A. N. "Need for Development of Local History Collections at the District Central libraries in Andhra Pradesh, Herald of Library Science 24 (1985): 273-280.

15. Raju, A. A. N. whither our Documentary? University News 40(2003):9-10.

16. Rajyalakshmi, D. "Digital Libraries". New Horizons 11(2001): 20-29.

17. Rath, Prabhash Narayan, Murlibhara, N, and Shinde, Swati. "Digitisation as a method of Preservation of cultural Heritage." Proc. of the intl. Conf . on Digital Libraries, V1, Feb.24-27, 2004: New Delhi : TERI, 2004. 392 -405.

18. Ratwani, M. R. And Ali, Amjad. "Electronic Preservation of Oriental Manuscripts. University News 40(2003):5-6

19. Skimmer, G. W. Heich, W., and Tunita, S. Modern Chinese Society : An Analytical Bibliography. V.I Publications in western Languages 1644-1972. VII Publications in Chinese 1644-1969. V.III Publications in Japanise. Califorania : Stanford Univ. Press, 1973.531.

20. Tourville, E.A. Alaska : A Bibliography. 1570 -1970. Boston : G. K. Hall and co., 1934. 738.

21. Word, R. E., and Shulman, F.J. The Allied occupation of Japan : An Annotated Bibliography of western language Materials. 1945-52. Chicago : ALA, 1974. 839.

22. Vijay Kiran, and Ramesh Babu. "Digitization of Archives." Proc. of the natl. Seminar on Impact of Digitisation on Development on Development of Information Professionals, Feb.28-Mar.1;2003:Ed

23. Ashwini A. Vaishnav, and Shashank Sonawane. Aurangabad : BAMU, 2003.CE2)

24. Brown, Heather. "Preserving cultural Heritage for Future Generations : A Hybrid solution". Proc. of the intl. Conf .on Digital Libraries, V1, Feb.24-27,2004: New Delhi : TERI, 2004. 406-417.3)

25. Smith, Abbey. "Why Digitise?" CLIR Issues. Council of Library and Information Resources (23 Dec. 2003) http://www.clir.org/pubs/issues/issue08.html

26. Shafi, S.M. "Makhtootat : Multilingual and Multiscript Digital Library of Medieval Manuscripts Initiative". Proc.of the intl. Conf .on Digital Libraries, V1, Feb.24-27,2004: New Delhi : TERI, 2004. 442-448.

27. Comment, Jean-Marc. "Archiving cultural heritage and history through Digitization : Case Studies from Russia (Comintern archives) and albania (national archives) Proc. of the intl. Conf .on Digital Libraries, V1, Feb.24-27, 2004: New Delhi : TERI, 2004. 387-391.)

28. Parekh, Harsha."Digitization in India : Developing and Implementing a National Policy." Proc. of the intl. Conf .on Digital Libraries, V1, Feb.24-27, 2004: New Delhi : TERI, 2004. 202-207.

29. Parekh, Harsha."Digitization in India : Developing and Implementing a National Policy." Proc. of the intl. Conf .on Digital Libraries, V1, Feb.24-27, 2004: New Delhi : TERI, 2004. 202-207.

30. Raza, Masoom, and Arora, R. L. Digitization on, Preservation and Management of Rare Materials in Modern Library system". IASLIC Bull. 49 (June 2004): 89-92.

About Authors



Ms. Chitra Rekha Kuffalikar presently she is working as Head, Learning Resources Centre, Mahila Mahavidyalaya, Nagpur, Maharashtra. She holds MA (Ling); MLISc, from Nagpur University. She has over 30 years of work experience in the LIS profession. She has contributed more than 50 papers in regional, national and international level seminar, conferences and journals. She is a Life Member of ILA, IASLIC, LIS Study Circle, Amravati, NUCLA etc.
E-mail : crekhak@yahoo.co.in



Dr.(Mrs.) Desaraju Rajyalaksmi is Reader and Head of the Department of Library & Information Science, Nagpur University, Nagpur, Maharashtra. Prior to this position she has worked in NEERI, Nagpur and University of Qatar, Doha also. She has over 20 years of work experience in the LIS profession. She holds M.Sc.(Zoology); MLISc; PhD in Library & Information Science. She has contributed more than 25 papers in national and international level seminar, conferences and journals. She is a Life Member of ILA, IASLIC, IATLIS, FISC, APLIBA, IWSA, Member of ALA(1985-87).
E-mail : desarajurl@rediffmail.com

Technology Enablers for Building Content Management Systems

Vasudeva Varma

Abstract

Managing content in a reusable and effective manner is becoming increasingly important in knowledge centric organizations as the amount of content generated, both text based and rich media, is growing exponentially. Creating content is expensive and unless it is staged, deployed and reused effectively, these costs cannot be justified. An important aspect of content management technologies is that they are far from being mature and they are getting better with time and this will continue to stay this way for some more time. In this paper we discuss important features of next generation content management systems and key technology enablers that make it possible to achieve next generation functionality. We describe two important technology enablers that are implemented with state of the art research besides identifying evolving and futuristic research areas that will help push the content and information systems to the next level. We describe a technology framework for managing a unified taxonomy and ontology network and a common messaging platform.

Keywords : Content Management Systems, Content Management, Common Message Platform

0. Introduction

We have entered an era of information overload. This is in contrast to what we have experienced in the past century where the main challenge was to find enough information. Most organizations are transforming into knowledge organizations where the key assets of organization are turning out to be people and knowledge. Every knowledge centric organization is producing more and more content and information. The new generation information and content management challenges can be classified into two major groups: information staging and information retrieval. Information staging include activities such as finding the information sources, building content crawlers, building content indexers, metadata creation and building tools to characterize the content. Information retrieval phase includes content query processing, content delivery using push and pull technologies, content monitoring and feedback engines.

As the information, documents of structured and unstructured nature are growing exponentially; we have a challenge of finding most relevant document(s) in the least possible time. Hence, obtaining very high precision and recall in information retrieval systems is very important. In addition, mergers and acquisitions are major hurdles faced by the twenty first century content management system architects. As number of organizations are being merged or acquired, making sure that the content of organizations can also be merged seamlessly is also very important. We need to plan for open architectures while building content and information systems to enable communication between completely different systems.

Content management is a discipline that manages timely, accurate, collaborative, iterative, and reproducible development of web and inter-organizational digital assets. It combines a mechanism to store a collection of digital assets with processes that seamlessly mesh the activities of people and machines within an organization. Content management responds to the unique combination of problems posed by digital asset development, typically web related.

1. Content Management System Features

Content management has to support various activities – creation of content, content publishing, storage and efficient retrieval. The structure of the content, content types and content aggregation must be addressed in an effective and unified platform.

- Structure of the Content: The structure of the document itself contains a lot of useful information and provides useful semantic information. This information will be used in the content analysis stage. For example, knowledge about a particular section like "summary" in a document solves half the problem – because we know what to expect. The retrieval and extraction efficiency are dramatically improved if we have the structural knowledge of the document.
- Content Types and Content Sources: The right technology should enable parsing, extracting and indexing various content types and content sources. Content Types include – text, voice, video and graphics and non-trivial or unstructured textual information. Content sources can be configured by the client application – for example, email documents, company information databases, websites, HTML, Word, PDF documents, digital video libraries etc. Each of these content types and content sources are configurable by the client application.
- Content Aggregation: We need to deal with multiple rich media content types. What is stored electronically is only a document or a database record. When presented finally to the user, the information becomes and turns into Content. For example, a single search may result in various rich media objects that are interlinked or completely independent from each other. We must provide powerful presentation and content aggregation schemes so that the retrieval of information is unified and can be presented through multiple devices. The application program can combine the search results in a meaningful manner and present the rich-media objects using various devices.

Enterprise Content Management (ECM) represents various technologies for web content management, document and record management and digital asset management [PWC, 2003]. Web content management deals with the process of content creation, revision and approval, and a version control system. A workflow system will enable all these processes in an appropriate sequence. Document management systems provide role based access to individuals, collection, meta-tagging, and coordination for creation, modification, use and storage of electronic and paper documents through their life cycle. Document management systems support versioning as well as methods of storing essential metadata so that the information can be classified, searched and reused more effectively. Content integration technologies provide opportunities to feed relevant and approved content into key business applications.

Enterprise Content management can also be seen as an amalgamation of related product categories. Various software vendors and service providers were catering the needs of web content management, documents and records management, search engines, portals, Knowledge management tools, syndication of content independently and later realized that all these individual product and service markets can be brought under a single umbrella of content management. Digital asset management and digital rights management are two other functional solutions that are related to the ECM.

1.1 Content Management Systems - Future Trends

IDC analyst, Susan Feldman [Feldman, 2001] observed that the next major stop in content management, Search and information retrieval domain is "conversational systems" where the user and the application are engaged in a dialogue to arrive at the right content within a shortest time possible. This area is making use of various technologies including advanced natural language processing, semantic analysis,

machine learning, and user modeling. She also felt that the growth rate of rich media content will be more than text based content.

A very important trend one can notice in the market today is about personalization and localization of search and content. Search results become more meaningful if they are customized to the user who is making the query and address his or her information need in a given context.

Given the changing landscape of technology trends, demands of knowledge centric organization and the criticality of building effective content and information management system, we need to build an open ended system that can accommodate the changes in technology, user requirements as well as platform dependent specifications. In this paper we discuss important technology drivers that will shape up future content management systems and give a component based framework as a common messaging platform for building content management systems. We believe this framework can help in coming up with large scale, flexible and scalable content and information management systems.

2. Key Features of Next Generation Content Management Systems

In this section we describe some of the key features and functionality of an effective content management system. These features can guide in identifying the technology drivers that will enable us build the next generation content management systems.

2.1 Enhanced Content Crawling

It is important to capture the digital assets from various sources and systems including that of file storage, web, databases and storage networks. The multi-media content can originate from any sources. The content crawlers and document filters can automatically 'grab' content from content databases, video sources, voice sources, XML Sources, emails etc. The content capture is shown in the following diagram:

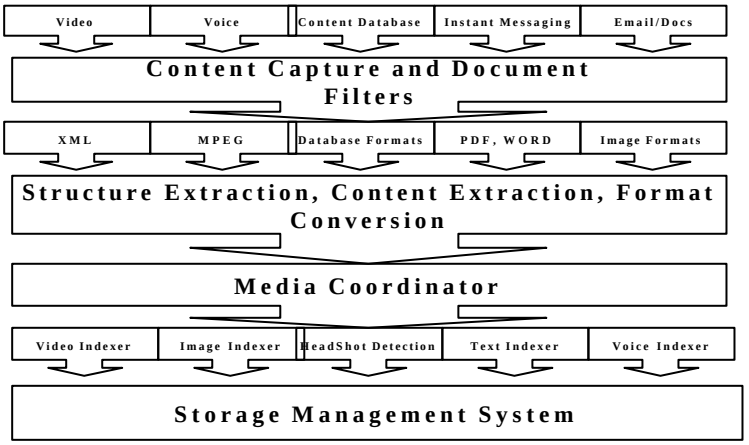


Figure 1: Content Crawling and Object/Document Management in Content Management Systems

The content from various sources is being captured by crawlers. Depending on the document type and document format, various filters can be used to separate the media specific content.

2.2 Digital Document and Object Management

Digital document and object management functionality deals with storage, filtering, extracting, structural analysis of rich media documents and objects. Consider for example, a multi media document that consists of a lot of structural information, textual information and images. We need to be able to parse and extract the structural information and different media objects from the document. Once this is done, structural analysis of the document provides us a lot of useful semantic information about the document. A format conversion might take place (for example, converting word file to ASCI text file) and is handed over to the media coordinator. The media coordinator will in turn pass the content to specific media indexing routines. We need store the structure of the document, pass the different media objects to their respective indexers and maintain the document in its original form in centralized storage servers. We need to create unique document identifiers at the same time. After indexing the content, storage management system will take over to capture both the document and the index of the document in appropriate database structures.

2.3 Storage Management

The back-end support issues for a multimedia content publishing system are very complex and need careful planning and implementation. This typically involves not only the storage management but also storage tracking, as we cannot expect entire content to reside on just one server. We need to provide tools to manage the digital assets in terms of their storage and manipulation that include re-packaging, re-creation and re-structuring

2.4 Media Specific Categorization and Indexing

Once the document is parsed and various media objects are extracted from the document, we need to index the sub-documents (or media objects) based on the media type. For example, video objects need to be parsed and indexed by the video indexers, similarly, textual objects need to be indexed by text indexers, image objects need to be indexed by the image indexers. There may be specialized and more than one indexer for any specific media type. The content management system architecture needs to be extensible to add new indexing engines.

2.5 Document Retrieval

It is important to find the right digital content in the shortest interaction time and in a very intuitive manner. We need to employ techniques such as “pearl growing” (improving and building upon the initial query). Ability to combine keyword or text based approach with sample image or image parameter based approach. An example query would look some thing like: “show me all the Toyotas which are shaped like this [insert or select an image] and are in black color and registered in Delhi”.

The system should be able to navigate through vast digital content with ease and efficiency. Users should be able to interact with the digital content servers in a very meaningful manner using the latest technologies such as conversational systems to arrive at right content in fastest possible manner.

2.6 Summarization

Summarization is important for textual documents as well as rich media documents. For example, video summarization is possible using the techniques that include video skimming or fast flipping of select frames and video information collages. For audio and text media we will use the summarization techniques developed in natural language processing and linguistics research. For images we can create thumb nails.

2.7 Personalization

Customization of the content for an individual requires mixing and matching content elements. Personalization has become very important in web content management systems and this area has proven to be highly specialized and complex. A new trend of personalization of results obtained by search engines is gaining popularity within search community. Personalization takes into account various parameters such as past and present user behavior, the context in which the search is being made, and predicted information need of the user.

3. Technology Enablers for Content Management Systems

It is not easy to build a high end content management system without certain technology enablers also known as technology drivers. To achieve the functionality described in the previous section, we need to build the foundation in the form of technology enablers. These technology enablers include good linguistic analysis, rich media processing, a common messaging platform where various components of the content management systems will be able to communicate in an optimal and uniform fashion and lastly a network of ontologies and taxonomies that form backbone of content processing. Linguistic analysis and rich media processing are key research areas that hold the key to the success of future content and information management systems. These research areas are growing very fast but are still far from being mature. In this section we deal with the other two technology enablers, namely common messaging platforms and taxonomy and ontology networks. We share our experience of creating these two fundamental building blocks in the context of architecting a framework for content management and knowledge management systems.

3.1 Universal Taxonomy and Ontology Network - UTON

The main purpose of any ontology is to enable communication between computer systems in a way that is independent of the individual system technologies, information architectures and application domain. The key ingredients that make up ontology are a vocabulary of basic terms and a precise specification of what those terms mean.

The term ‘ontology’ has been used in this way for a number of years by the artificial intelligence and knowledge representation community, but is now becoming part of the standard terminology of a much wider community including object modeling and XML.

Adoptable, high performing, large scale ontologies that can be extended to support multi-media play a crucial role in building effective content and information management systems and applications. This section describes the architecture of Unified Taxonomy and Ontology Network (UTON).

The ontology or taxonomy defines the central semantic network – in other words, it is a repository (industry specific, customizable or universal) that servers as basis for all the indexers. The content management system should be able to operate with multiple taxonomies and ontologies at the same time. It should be possible to switch between taxonomies or ontologies depending on the context and the input document. Hence it is important to come up with a framework where multiple taxonomies or ontologies can co-exist and accessed using unified protocols.

The content management systems and information management systems make use of ontologies at several functional points that include: document categorization, indexing, document (parts or entire document) retrieval, user query expansion, query matching, and result verification. Since rich media documents are also becoming pervasive and important (perhaps more important than the textual

documents) there is an emphasis on extending the ontologies work for multimedia documents as well. For this purpose, we need to build ontologies that support rich media document processing.

Ontologies can be used to provide semantic annotations or meta-tagging for collections of images, audio, or other non-textual objects. These annotations can support both indexing and search. Since different people can describe these non-textual objects in different ways, it is important that the search facilities go beyond simple keyword matching. Ideally, the ontologies would capture additional knowledge about the domain that can be used to improve retrieval of images.

UTON stores multi media concepts, relations among these concepts, cross linkages, language dependencies in its repository and provides interfaces to storage and retrieval functionality and the administrative functionality (including user and version management). The knowledge and semantic information is stored within the network in the form of a DAG (Directed Acyclic Graph). The storage and retrieval interfaces provided by ontology network are being used by various media indexing and categorization components. Ontology developers, editors and administrators will have different interfaces.

All these interfaces interact with higher level UTON objects such as Ontology, Concept, term and relation. If ontology consists of concepts belonging to more than one domain or sub domains, then another higher level object called context will come into play to help disambiguate concepts belonging to more than one domain. The following paragraphs describe each of these higher level objects:

- Ontology: the ontology is the topmost entity, necessary because the intention of UTON is to contain a network of taxonomies and ontologies, likely to be contributed by different sources. Depending on the number of domains the ontology contains a set of contexts will form the ontology itself. As attributes, the ontology has a name (mandatory and unique), a contributor, an owner, a status ("under development", "finished" ...) and documentation (an arbitrary string in which the contributor or the owner can specify relevant information).
- Context: a context is actually a grouping entity; it is used to group terms and relations in the ontology. Within a given ontology, every context should have a unique name. The context object comes into picture when there is a possible existence of ambiguous concepts (see below for the description of concept), terms and relations among them when a given ontology covers more than one domain or sub domain, which is typically the case.
- Concept: a concept is an entity representing some "thing", the actual entity in the real world and can be thought as a node within the ontology structure. Every concept has a unique id. A concept also has a triple "source-key-value", which is the description for that concept. The "source" identifies the source from which the description originates, the "key" is a string which gives a hint to the user on how he should interpret the value, and finally the "value" is the description of the concept. One concept can have more than one source-key-value triple, and thus have its meaning described in different ways. As an example, let's consider WordNet [Fellbaum, 1999]. In WordNet synsets denote a set of terms (with their "senses") which are equivalent. Every term also has a glossary, which is an informal description of the meaning for that (particular sense of the) term. In this respect, from WordNet, we can extract two different descriptions for a concept, two different source-key-value triples, namely the glossary (Source: WordNet - Key: Glossary - Value: "<informal description denoted as a glossary in WordNet>") and the synset (Source: WordNet - Key: Glossary - Value: <enumeration of synonyms forming the synset>). As a different example, when a concept exists in various media (text, video, audio and image), a concept represented using source-key-value triple will give the appropriate media value, when retrieved using appropriate key.

- Term: a term is an entity representing a lexical (textual) representation of a concept. Within one context, a term is unambiguous and, consequently, it can only be associated with one concept and of course, several different terms within one context can refer to the same concept, implicitly defining these terms as synonyms for this context. Terms in different context can also refer to the same concept, and in this way implicitly establish a connection between these two contexts.

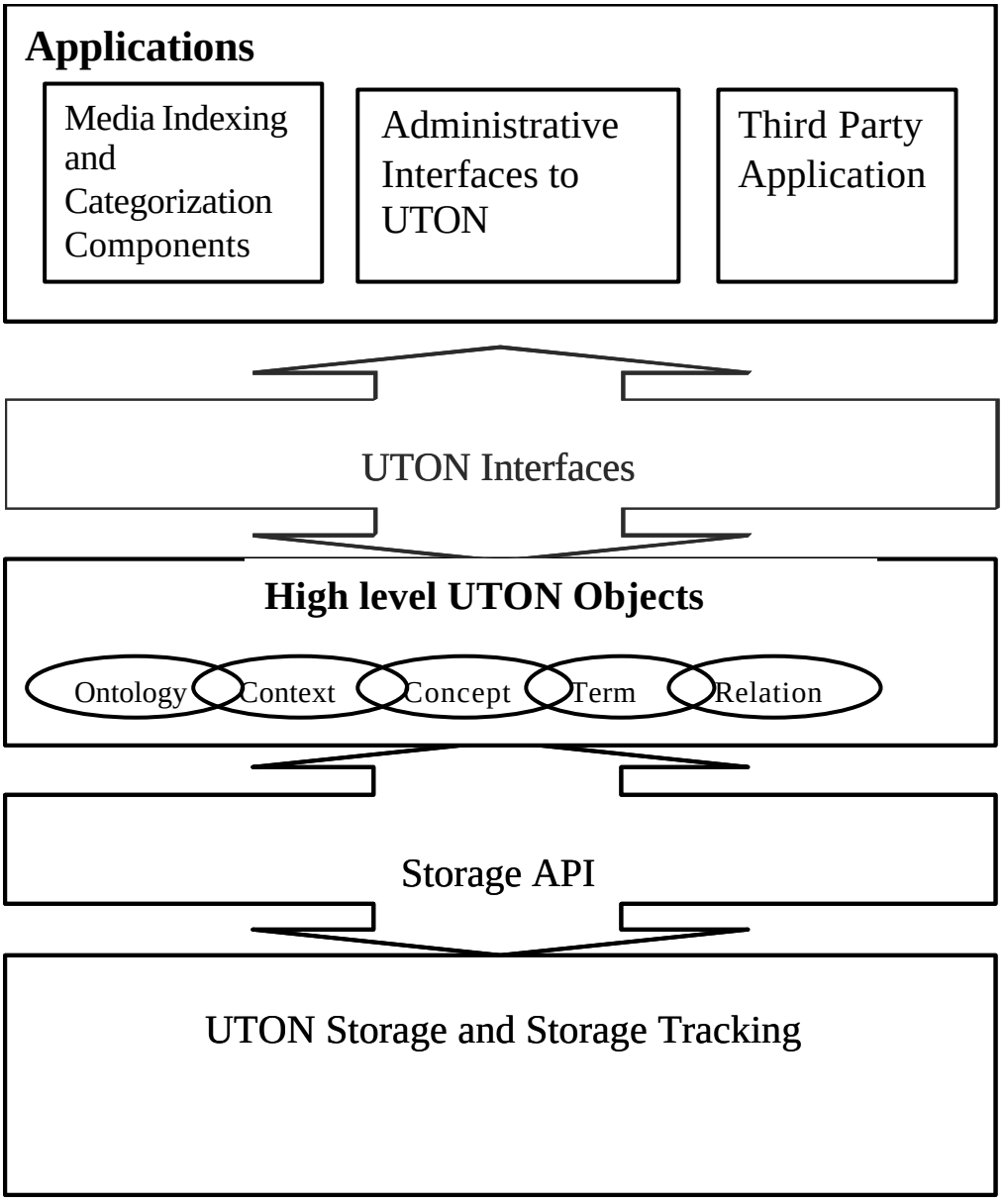


Figure 2: Architecture of UTON

- Relation: a relation is a grouping element; it can be interpreted as a set of triples consisting of a starting term (also called the "headword" of the relation), a role (relation name) and a second term (also called the "tail" of the relation).

As we can see in the above figure, the general architecture components are:

- UTON Storage: the storage system is the place where the UTON data is stored – typically a Relational Database Management System (RDBMS).
- Storage API: Provides a unified access to the basic structures of UTON. The API should be accessible from any high level programming language.
- Higher level UTON objects: UTON objects are expressed in a data description language format, or as objects in any high level programming language. They are retrieved and stored using the storage API.
- Applications: applications can use the UTON by integrating the ontology objects returned from the storage API in their program code.

This architecture and design of UTON [Varma, 2002] will enable multiple ontologies and taxonomies to co-exist and make it possible to access them in a unified manner. Our major focus is to build a network of large scale ontologies and taxonomies that are highly scalable and with high performance and guaranteed quality of service. All the components can be distributed and can be running on a set of server forms to obtain the required scalability and performance.

We have developed UTON in the context of developing information extraction, indexing and categorization engines for a content management system that is heavily rich media oriented. WordNet played a major role in coming up with an initial ontology.

3.2 An Approach to Developing Common Messaging Platform

A major challenge in building content management system is in scaling up the system. In real world scenario, we have huge amount of content flowing in from various sources and each content document may be separated into sub documents depending on the type of media. These subdocuments need to be indexed and characterized by appropriate indexer. In addition, metadata like named entities, author information is extracted and summaries are generated and documents are categorized into predefined categories. This new metadata has to be seamlessly merged into the characterization of existing document repository. All this has to be done in real time to make sure that users have access to latest information and content.

Each of the tasks mentioned above may be executed by a farm of severs where the appropriate components are deployed. These components need to be supported by an efficient communication channel that does not become a bottleneck in achieving high performance. If each component is allowed to talk to every other component then network traffic can become very high and congestion can occur causing the content management system to breakdown. To address this important performance engineering issue, we have designed a common messaging platform where every message from every component will be sent through this platform and will be received by the appropriate target component. The solution is building content based common messaging platform where every message is very thin (textual – XML based) but will enable the target component to take act on the data. We were scouting for commercial off the shelf components and selected "Elvin" which we later customized for our need.

Our approach to building the common messaging platform is based on component architecture. All modules are components and these components are connected by a messaging platform. Each of these components can have multiple instances and these instances can be running on multiple machines.

Elvin is a messaging platform that keeps track of the state of the components that are registered with it and receives and passes the messages appropriately to the corresponding components. We used the messaging platform in the following manner.

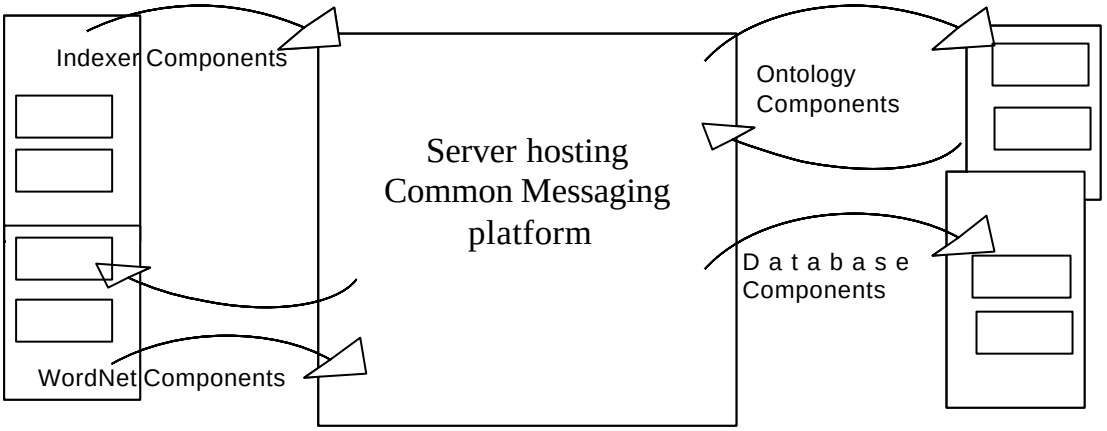


Figure 3: Common Messaging Platform

There can be one or more instances of various components such as Indexers, WordNet servers, UTON components and Database components. A component such as an indexer may in turn have more than one instance. For example, a video indexer may be running three of its instances and an image indexer will have two instances.

We have implemented a communication protocol in architecting the common messaging platform that no component communicates directly with any other component. All communication between common messaging server and client components are in the form of messages. For example, the video indexer sends a message to common messaging server requesting synsets for each of the terms it obtained. The common messaging server sends a message to the available instance of WordNet component giving the particular term as part of the message and the WordNet component gives back the synset back to the common messaging server and so on.

Multimedia indexers would parse the corresponding media objects from the documents and then return a set of terms. This is the key point because the entire complexity of the media object analysis process is inside the indexer and whatever may be the media object, the indexer returns only the terms (basically some textual strings) which will be easier to pass between the components. If we had to pass the actual multimedia objects it would result in an unusable system because we will not be able to achieve reasonable response and processing times. If actual objects need to be shared by different components then only the location addresses will be passed and those component will directly interact with the storage tracking system. In this manner, we achieve high performance and high scalability.

4. Conclusions

The importance of building highly scalable and feature rich content management and information management systems is becoming very clear as we face the challenges of information overload. We have discussed common features of current generation content management systems and the trends of new generation systems that demand rich media processing, conversation based interfaces and personalization aspects. We have tried to identify the key functionality of future content management system and the technology enablers to achieve this functionality. Language processing and rich media analysis will play a major role in making future content management systems more effective but, they are far from being mature disciplines. With the existing technology know-how, we have tried to present two enablers that form a part of the foundation for building next generation content management systems.

One technology enabler is to build a Uniform Taxonomy and Ontology Network (UTON) where multiple taxonomy or ontology belonging to different domains can co-exist with uniform interfaces. This will help us in building enterprise wide content management systems that can be used across departments, locations and service or product offerings. UTON can aid in creating metadata of the content, content navigation and in scaling the content and information management systems. Second technology enabler is creating common messaging platform using component architecture so that very complex and high scale content management systems can be built with high levels of performance and with minimum communication overheads. The first technology enabler will help in improving the functional aspects of content management system as it addresses quality of content characterization and the second enabler will help non-functional aspects such as scalability and performance.

5. References

1. Demoz, http://demoz.org
2. Fellbaum, Christiane (Ed). WordNet: An electronic lexical database, MIT Press, 1999.
3. Feldman, Susan “Content Management” in eInform Volume 3, Issue 7, IDC News letter, 2001
4. Lenat, D. B. and R. V. Guha. *Building Large Knowledge Based Systems*. Reading, Massachusetts: Addison Wesley, 1990.
5. Price Waterhouse Coopers, Technology forecast 2003-2005. 2003
6. Harabagiu Sanda M, Moldovan Dan I, Knowledge processing on an extended WordNet appeared in [Fellbaum, 1999]
7. Semantic web: http://www.semanticweb.org

About Author



Dr. Vasudeva Varma is a faculty member at International Institute of Information Technology, Hyderabad Since 2002. Prior to joining IIIT-H, he was the president of MediaCognition India Pvt. Ltd and Chief Architect at MediaCognition Inc. (Cupertino, CA). Earlier he was the director of Engineering and research at InfoDream Corporation, Santa Clara, CA. He also worked for Citicorp and Muze Inc. in New York as senior consultant. He obtained his Ph.D. from the Department of Computer and Information Sciences, University of Hyderabad in 1996. He has five patent applications and several publications in journals and conferences. He recently obtained young scientist award and grant from Department of Science and Technology, Government of India, for his proposal on personalized search engines. His areas of interests include search and information extraction, knowledge management and software engineering. He is heading the Search and Information Extraction Lab at Language Technologies Research Center (LTRC). His team is developing search engines for Indian languages, working on named entity extraction and personalized search engines. He is also interested in experimenting with non-conventional methods for teaching software engineering in general and case study based approach In particular.

Email : vv@iiit.net

Searching Patent and Patent Related Information on Internet

Sumati Sharma Mohinder Singh

Abstract

The paper gives a brief account of worldwide patent and patent related information available through internet. It enlists the important web resources of various international/national bodies, commercial vendors and others providing either information/literature on the subject or patent document search itself. Each entry indicates its web addresses along with a brief description of site and type of information/literature hosted on it. It also highlights the importance of patent literature for R&D work and lists a number of key features of a patent document which makes it a unique source of information

Keywords : Patents, Information Retrieval, Internet Resources

0. Introduction

*21*st century is marked with the fast changing information technology competitive world. In any industry, Research and Development activities are high priority areas and information is playing a vital role by providing the base for any R&D activity. In this context patent information is a vital source of information for any R&D work. Patents reveal solutions to technical problems. More than 80 percent of all technical knowledge is described in patents. To have a better understanding of this highly valuable source of information let us first understand what exactly a patent is. A patent is defined as a grant by the sovereign or state to an inventor or to his/ her assignee giving exclusive rights to make use, exercise and vend an invention for a limited period in exchange for disclosing it in a patent specification. The disclosure should be such that a person trained in the art (i.e. in that field or discipline /subject) should be able to reproduce the invention. When the patent is granted, the owner gets the right to exclude others from using the invention. In more simplified terms “a patent is a declaration from a government that an invention or process is new or innovative enough to be granted the exclusive right to manufacture or otherwise use the invention for a set period of time”.

1. Significance of Patent Literature

To highlight the fact, that for any researcher, patent literature is a primary source of information, some salient features of this source of information are described below which make patent a unique source of information.

- A large percentage of technology disclosed in patents is never published in any other document.
- Patents contain complete details of the invention, including its method of working.
- Patents are easily accessible through use of International Patent Classification.
- Patents form one of the earliest publications of patented invention.
- Patent documentation forms a single storehouse of technological information, covering widest range of technical fields irrespective of level of sophistication of technology.
- Patents are presented in a standardized format. Once familiarity is gained, access to relevant information is easy.
- Provide state of the art on a specific micro field/subject/technology
- Indicator of advancement and direction of R&D in a specific subject field/technology

2. Importance of Patent Literature Searching

In recent years, there is a growing awareness about the importance of patent literature and patent literature searching has gained very high momentum due to the fact that

- R & D sector is facing a very tough competition on global basis. To survive in such a competitive world, the companies/organizations are required to spend a good amount of money on their R&D projects. As R&D itself is a very cost intensive, investor will certainly expect high returns also for his/her growth and survival. Patent searching on global basis becomes the first and foremost step while taking up any new R&D project to ascertain that the same product or process is not patented elsewhere. This saves considerable efforts, money and time in today's competitive world.
- Secondly, researcher/company/organization must get their invention/innovation patented in their names other wise someone else may take the benefits of their research findings by getting it patented in their names. Since the patent is granted only for novice, non-obvious and useful inventions, one cannot lose in investment just because of the ignorance or unawareness. One has to be up-to-date about what is the R&D status in a particular area.
- Patenting a process or product is also necessary, if one wants to use his/her invention research breakthroughs for future commercial exploitation and gain achievable profits out of it. Further applying for a patent also involves some expenditure which will go waste if one is denied grant of the patent by patent examiner/patent granting authorities if the patent on the same process/ product is already filed.

3. Patent and Related Information Resources on Internet

Thanks to internet technology, now days it is not a difficult job to search worldwide patent databases sitting at the internet terminal and get the desired information quickly.

Listed below are some of the important websites on patents and related issues. One can visit the appropriate site as per his/her information requirements.

3.1 World Trade organization (WTO) (http://www.wto.org)

This is the official site of the World Trade Organisation, which lists all important announcements/happenings in WTO, hosts awareness material on GATT and WTO and related topics. Site also provides access to statistics database and full text of publications like WTO Annual Report (published in the first half of each year), Dispute Settlement, Discussion papers etc. In addition to several subjects, it has a specific section on Intellectual Property Rights, where several relevant official documents on TRIPs agreement are also available. [http://www.wto.org/wto/intellec/intellec.html].

3.2 World Intellectual Property Organisation (WIPO). (http://www.wipo.int)

This is an international body taking care of all forms of IPR documents and IPR related issues. Presently 181 countries are members of this international organization. Site gives detailed information regarding member states, budget, various treaties, conventions etc. The database of International Patent Applications contains all published PCT applications in all disciplines starting from 1978. Complete bibliographic, administrative and legal information, during international phase is provided. Patent drawings are also available. Approximately 600,000 images are available.

3.3 Patent Cooperation Treaty (PCT) Database (http://www.wipo.int/pct/en)

The database covers full-text of PCT(Patent Cooperation Treaty) published applications issued under the World Intellectual Property Organisation (WIPO). At present 181 member states participate in the PCT system. One PCT application may be valid in any of all 181 Designated States. India is also a member of this system. Database contains the first page data which includes bibliographic information, titles and abstracts (searchable in English or French), descriptions (specifications), claims (the majority of which are in English, but some may be in French, German or Spanish) and drawings of published PCT applications. The first page data of applications published each week in section I of the Gazette is added weekly to the database. The database currently contains data relating to applications published from January 1, 1997.

3.4 Intellectual Property Digital Library (IDPL) (http://www.wipo.int/ipdl/en/)

The Intellectual Property Digital Library (IPDL) web site provides access to various intellectual property data collections hosted by the World Intellectual Property Organisation (WIPO). These collections include PCT (Patents), Madrid (Trademarks), Hague (Industrial Designs), Article 6ter (State Emblems, Official Hallmarks, Emblems of Intergovernmental Organisations) and access to some other data collections like Health Heritage (Traditional Knowledge Test Database), JOPAL (Journal of Patent Associated Literature) also.

3.5 PCT Electronic Gazette (http://www.wipo.int/pct/en/gazette/index.jsp)

The PCT Electronic Gazette contains data relating to PCT international applications published as PCT pamphlets from January 1997, and where applicable, republished from April 1998. Bibliographic data, abstracts, drawings and images of PCT pamphlets are provided for all published and republished international application in the collection. For international applications published and/or republished since April 1998, the searchable text of claims and descriptions is also provided.

The bibliographic data, abstracts, drawing and images of the international applications published and republished each week are available from the collection on the international publication data. The searchable text of claims and descriptions of published international applications is available from the collection as soon as possible after international publication (generally 2 to 3 days after the publication of data.).

3.6 Madrid System for the International Registration of Marks (Madrid Express Database) (http://www.wipo.int/)madrid/en/)

This system gives a trademark owner the possibility to have his/ her mark protected in several countries by simply filing one application with a single office, in one language, with one set of fees in one currency (Swiss francs). The Madrid System for the International Registration of Marks is applicable among the countries party to the Madrid Agreement. The Madrid Express database includes all international registrations that are currently in force or have expired within the past six months.

3.7 Hague System for the International Registration of Industrial Designs (Hague Express Database) (http://wipo.int/hague/en/)

The Hague System for the International Registration of Industrial Designs is applicable among the countries party to the Hague Agreement. This system gives the owner of an industrial design the possibility to have his/ her design protected in several countries by simply filing one application with the International

Bureau of WIPO, in one language, with one set of fees in one currency (Swiss francs). The Hague Express Database includes bibliographical data and, as far as international deposits governed exclusively or partly by the 1960 Act of the Hague Agreement are concerned, reproductions of industrial designs relating to international deposits that have been recorded in the International Register and published in the International Designs Bulletin as of issue no. 1/1999.

3.8 Article 6terof the Paris Convention (http://www.wipo.int/article6ter/en)

Article 6terof the Paris Convention for the Protection of Industrial Property (Paris Convention) is applicable to the States party to the Paris Convention as well as to all the Members of the World Trade Organisation (WTO), whether or not party to the said Convention through the Agreement on Trade-Related Aspects of Intellectual property Rights (TRIPS Agreement).

The purpose of Article6ter is to protect armorial bearings, flags, other State emblems, abbreviations and names of international intergovernmental organizations and of those States and Members identified above.

3.9 JOPAL (Journal of Patent Associated Literature) (http://www.wipo.int/scit/en/jopal/jopal.htm)

This database contains bibliographic details of articles published in leading scientific and technical periodicals from 1981 to May 2003.

3.10 Health Heritage (Traditional Knowledge Test Database) (http://www.wipo.int/ipdl/en/search/tkdl/search-struct.jsp)

This compilation provides a test database of public domain traditional knowledge, which is made available on the WIPO web site at the request of the Indian Government. The test database is based on the “Health Heritage CD-ROM’, which was compiled by the Council of Scientific research (CSIR) of India. It contains documentation data of codified traditional knowledge, all of which is already in the public domain. It may be used as a trial product to test the perceived potential of traditional knowledge databases for improving the availability of disclosed traditional knowledge as searchable prior art. All the documentation data in the database were collected and compiled by the Indian CSIR and were provided to WIPO with a request to make the data available online.

3.11 US Patent and Trademark Office (USPTO) : Published Applications (http://www.uspto.gov/appft/index.html)

It is a Published Applications Database, which is covering U.S. applications published by the USPTO since March 2001. Coverage includes: (1) A1, publication of the patent application; (2) A2, second publication of the patent application; (3) A9, corrected patent application; (4) P1, publication of plant patent application

3.12 US Patent and Trademark Office (USPTO) : Issued Patents (http://www.uspto.gov/patft/index.html)

The USPTO Full-Text Database provides access to the full-text and full-page images of patents published by the USPTO since 1976. Database covers U.S. Utility Patents, Reissue Patents, Defensive Publications, Statutory Invention Registrations, Design and Plant Patents.

3.13 European Patent Office (EPO) (http://www.european-patent-office.org/espacenet/info/access.htm)

The EPO grants European patents for the contracting states to the European Patent Convention (EPC). EPO provides access to European published applications and grated patents full-text. It gives full-text coverage of published applications from1987, complete bibliographic information from 1978 to the present , and full-text and bibliographic information coverage of granted patents from 1991.

3.14 epoline (http://www.epoline.org)

epoline is the name given to the range of electronic products and services produced by the European Patent Office (EPO) for the intellectual property community. epoline provides secure and integrated means of electronic communication between patent applicants, their representatives, the EPO and the patent offices of the EPO's member states. It also provides for online filing, fee payment, file inspection and Register enquires etc.

3.15 espacenet Network: Europe's Network of Patent Databases (http://gb.espacenet.com/espacenet/gb/en)

esp@net is a free service on the internet provided by the European Patent Organisation through the EPO and the national offices of its members states. It enable users to search for published patent applications in their original language from Great Britain, other European countries, the European Patent Office and WIPO(PCT). Network also provides access to published patent applications with an English abstract and title from worldwide-30 million documents and Japan.

3.16 EPIDOC-INPADOC Databases (http://www.european-patent-office.org/inpadoc/general.htm)

Ten different services (databases) are produced by EPO covering various facets of patent literature. Out of various services produced by EPO, Patent Family Service (PFS) and Patent Register Service (PRS) , are the largest patent databases in the world in terms of both the countries and time-span covered . PFS deals with all patent documents applied in 65 patent offices worldwide, and the PRS deals with the legal status of patents (are they in force or not) in 22 patent offices. Respectively approximately 25,000 and 40,000 documents are added to the PFS and PRS databases each week. Various INPADOC databases listed below also indicate the type of treatment they give to subject.

3.17 EPIDOS-INPADOC Patent Family and Numerical List (PFS/INL)

A database which brings together patent publications with similar claims from a wide range of countries. The publications are sorted into "families, so that the user can find out in what countries a patent for a given invention has been applied of or granted. This makes it easier for companies to monitor the import and export strategies of their competitors and to determine the countries in which the invention is not protected and can therefore be freely used.

3.18 EPIDOS-INPADOC Patent Register Service (PRS)

The PRS is legal status database, which shows whether a particular patent is still valid or has expired. This facilitates the exploitation of inventions which are no longer protected and ensures that user do not pay unnecessary license fees for lapsed patents. All legal status changes, before and after grant, are listed.

3.19 EPIDOS-INPADOC Numerical Database (NDB)

This service is used to find the number of patents documents in a list arranged by date. The bibliographic data relating to each stage of publication can be reviewed at a glance.

3.20 EPIDOS-INPADOC Patent Classification Service (PCS)

The International Patent Classification (IPC) comprises around 60,000 technical subdivisions. The PCS lists patent documents according to their IPC classification, so that information about the state of the art in a specific area of technology is readily obtainable at any time.

3.21 EPIDOS-INPADOC Patent Application Service (PAS)

This enables users to find out about the activities of a specific applicant and thereby keep track of trends in research and development and see how the market is moving.

3.22 EPIDOS-INPADOC Patent Applicant Priorities (PAP)

A service listing patent documents by priority date under the name of each applicant, enabling users to monitor development in patent families.

3.23 EPIDOS-INPADOC Patent Inventor Service (PIS)

This service lists patents by the name of the inventor, whose research activities can thus be monitored. It offers an easy way of collecting specialist technical literature by a particular author.

3.24 EPIDOS-INPADOC Patent Gazette (IPG)

This is an international gazette enabling users to monitor the patent publications of over 50 countries and organization. Documents are listed by patent number, applicant and inventor. Published on a weekly basis, the IPG contains details of all the documents processed in the previous seven days.

3.25 EPIDOS-INPADOC Watch

Through the WATCH system users can monitor the updates to the EPIDOS-INPADOC database and PRS- carried out once a week. Lists showing changes in patent families and legal status are dispatched to clients automatically.

3.26 EPIDOS-INPADOC CAPRI system

This database contains over 14.4 million patent documents reclassified according to the IPC, together with 4.4 million documents from Japan and 625,000 from the former Soviet Union. The collection extends to 1920 and beyond.

3.27 The UK Patent Office (http://www.patent.gov.uk/)

The office is responsible for intellectual property (Copyright, Designs, Patents and Trade Marks) in the UK. Site also provides information on other IPR related issues.

3.28 Japan Patent Office (JPO) (http://www.jpo.go.jp/torikumi_e/head.htm)

Provides access to Patent Abstract of Japan (JAPIO). Also gives information & other news regarding patents.

3.29 Thomson Derwent- Derwent World Patents Index(DWPI) (http://thomsonderwent.com/products/patentresearch/dwpi)

Derwent World Patents Index is the most comprehensive database of patent documents published in the world. The database currently contains 13 million patent records.

3.30 Thomson Derwent- Derwent World Patents Index First View (DWPIFirstView) (http://thomsonderwent.com/products/patentresearch/dwpifv/)

This is a new, fast-alerting companion file to Derwent World Patents Index(DWPI). DWPI First View contains previews of the latest published documents in advance of their inclusion in DWPI. This file contains bibliographic data for all new patent documents, along with original titles, abstracts, technical drawing images, and English-language abstracts for patents from China, Japan, Korea, Taiwan and Russia.

3.31 Delphion Database (http://www2.delphion.com)

Provides access to world's top patent collections, like the USPTO, EPO, WIPO PCT and INPADOC. Search results may be downloaded using a variety of services(options) given.

3.32 Questel.Orbit (http://questel.orbit.com)

This service lets you conduct highly sophisticated searches on the most extensive collection of patent databases, each from the world's major patenting authorities and information providers. This is a fee based commercial service. Questel-orbit also offer a offline patent search known as Questel PATService.

3.33 National Informatics Centre(NIC), New Delhi (http://patinfo.nic.in/new2.html)

Intellectual Property & Know How Informatics (patent) Division of National Informatics Centre, Department of Information Technology, CGO Complex New Delhi, India provides patent search service through a number of international patent databases.

3.34 Patent Information System(PIS) Nagpur (http://www.patentoffice.nic.in/ipr/pis/pis.htm)

Government of India, Ministry of Commerce and Industry, Department of Industrial Policy established Patent Information system [PIS] at Nagpur in the year 1980 in order to obtain and maintain a comprehensive collection of patent and patent related information on world wide basis and provide access to this collection through services. Full text of Patent Documentation available at PIS includes patents from almost all major patenting authorities.

3.35 Indian Patent Searchable Database: Patent Facilitating Centre (http://www.indianpatents.org.in/db/db.htm)

Patent Facilitating Centre (PFC) of Technology Information, Forecasting And Assessment Council (TIFAC), in addition to its CD version, is now providing access to Indian Patent Database on internet also.

To sum up, patent is an excellent source of primary information for researchers but is still an under utilized source which needs to be appreciated more. Above listing of web sites is not an exhaustive one, this is only a small portion of enormous wealth of information available on net on the topic. There are many more web sites providing information on the subject. To start with, one can visit any appropriate site listed above and can go in deep to explore further useful links or new links be explored as per one's information needs.

3. References

1. Search for Patents. www.questel.orbit.com
2. WIPO. www.wipo.int
3. Indian Patent Searchable Database. www.indiapantents.org
4. World Trade Organization. http://www.wto.org
5. European Patent Office. http://www.european-patent-office.org
6. US Patent and Trademark Office (USPTO). http://www.uspto.gov
7. Carvalho, Nuno Pires De (2002). TRIPS regime of ***patent*** rights. London : Kluwer Law International.
8. Beresford, Keith (2000). Patenting software under the European Patent convention. London : Sweet and Maxwell.

About Authors



Mrs. Sumati Sharma, Scientist 'D' is working in DESIDOC and looking after the Defence Science Library. Her area of expertise is Documentation and Information Services. She has been awarded Associateship in Information Sciences from INSDOC, Delhi. She has about 10 papers to her credit.
Email : sumati_s@yahoo.com



Dr. Mohinder Singh, Scientist 'G' is working as Director, DESIDOC, DRDO. He has played a significant role in developing DESIDOC as a Digital Documentation and Information Centre. With the core expertise in the field of documentation, he has also got expertise in developing IT based information and network services. Dr. Singh was awarded Certificate of Merit for his significant contribution in the field of documentation and project support activities by DLRL, DRDO. He has about 35 papers and 12 books and other publications to his credit.

XFML, Standard for Distributed Information Architecture

Aparajita Suman

Abstract

The unplanned dumping of information on WWW is the primary reason for the chaotic web of today. The way out of this WEB (in literal sense) seems to be only through planned organization of the burgeoning sources of information using faceted techniques. The role of XFML i.e. eXchangeable faceted metadata language gets defined in this context. The metadata language allows one to relate the topics of concern and to provide a meaningful way of accessing information. This paper tries to explore its nuances and suitability for defining distributed information architecture.

Keywords : Faceted Metadata Language, Metadata Information Architecture, XFML

0. Introduction

XFML, i.e. eXchangeable faceted metadata language is no doubt another addition to the series of XML based languages rather but has a different and distinct flavor. XFML is an XML format for exchanging metadata in the form of faceted hierarchies, which are also called 'taxonomies'. It helps in solving one aspect of metadata problem, i.e., interchange of faceted classification and indexing data.

Facets can be said to be clearly defined, mutually exclusive and collective exhaustive aspects or properties or characteristics of a specific subject or class [1]. It is quite similar to human way of thinking, when we use multiple facets to describe a subject rather than attempt to fit it into some type of taxonomy or hierarchy. That's why faceted classification (i.e., describing things by their characteristics rather than assigning them a universal category) always has advantage over standardized ontologies.

Whenever indexing and classification of facets is done for a particular subject area, then, it is always advantageous and easy for others to use the same data rather than going for duplication of effort. For example, let us imagine a situation where we have a digital library in a particular subject area. We post the entries in our tailor made categories, now, some other library may also have digital library service in the same subject area and it may post entries in its own categories. Here, possibility of categories being similar is greater and the difference might only be in the way they are addressed.

The role of faceted metadata language gets defined here very well, as both the libraries can publish their metadata in XFML format. It allows them to relate their categories, where they continue to have their respective names but the two library systems know that they are really the same and related. Actually, we import the XFML file of the other library and relate the categories, then, the DL software is configured to import the file daily to check for changes. So, the XFML compatible software can automatically generate links to new features about the same topic as the other digital library.

1. Need of XFML

Library science research has proved that the best way of organizing knowledge is to put topics in various facets because categorizing, browsing and searching becomes easy in this case. The content management systems serve the purpose to some extent but they are deficient in content mapping. Most CMs implement a certain level of metadata but the standard way of publishing this metadata remains missing.

XFML came up as a solution to the problems associated with imperfect and ever changing taxonomies. In content management systems, metadata is intertwined with content definition; so, adjusting the string of metadata involves a lot of work. But in XFML map, metadata and content is separate. Here, we create a map that exists on its own, regardless of whether there is any content that relates to the map or not. So, it provides a lot of power to the user by allowing her to work with the map as an entity of its own, import facets of other maps, merge topics etc. **[2]**

2. Definition

XFML is an XML format for publishing and sharing hierarchical faceted metadata and indexing efforts. It provides ways to build connections between topics, information that helps in designing tools to automate the sharing of indexing efforts **[3]**. Its primary goal is to allow us to publish our metadata categories on web in a standard format and connect our metadata with the metadata, other people might have published. So, finally it enables automatic link generation to related content on our own and other websites.

Here, faceted classification is helpful because instead of building one huge tree of topics, multiple smaller trees are used that can be combined by the users to find things more easily.

3. History and Evolution

On May 30, 2002, XFML 0.1 was published, and then came XFML 0.2 on July 6, 2002. It was introduced with a number of refinements compared with the version 0.1 i.e. elements which allow software to easily display singular or plural versions of topics, facets and occurrence types. The main outstanding issues were whether topics should be allowed to have multiple parents. The other additions till date are XFML core (Oct 08, 2002), revisions to XFML core specification (Dec 13, 2002) and classification to XFML core specification **[1].**

4. Features of XFML [2]

- XFML takes the distributed approach and at the same time, allows individual authors to connect their metadata schemes by merging topics. It is impractical even to think of creating a centralized metadata store for the web, so, we always need shared metadata to make the web easier to get around it. This distributed connected metadata network can be made using XFML and which will function very much like a centralized metadata store.
- Taxonomies can be easily shared and published using XFML. Publishing the taxonomy means that using it one can get relevant incoming links and there is no need to do the labor-intensive task of creating taxonomies.
- XFML is well equipped to deal with changes in taxonomies as metadata in XFML map can evolve more easily than in current content management systems.
- Faceted taxonomies are generally more powerful for websites than classic hierarchical taxonomies. So, use of XFML is more suitable as it is based on facets.
- XFML allows us to index anything on the web as no write-access is needed to a web page to index it in the topic map
- Finally, allowing meaningful metadata connections between separate systems is what makes XFML so powerful

5. Difference from other Standards (Metadata Related Technologies)

XFML is a very specific and focused format, as opposed to eXtensible Topic Map or Resource Description Format, which are generic metadata formats [4]. It is optimized for a specific goal i.e. to enable sharing and connecting faceted metadata between websites; moreover, it is easy to write code for XFML at the same time.

5.1 XFML and Dublin core

Dublin Core is a specification that is additional to XFML: these two specifications can work together. XFML indicates relationships between topics, but not what the topics mean. Dublin Core can be incorporated into XFML to do that.

5.2 XFML and RDF

RDF is all about adding meaning to the web documents while XFML explains the manner in which the facets are related on web. RDF also lets web pages publish their metadata but it is too complex to code while XFML has been designed for easier code implementation.

5.3 XFML and TopicMaps:

XFML is a subset of topicmaps. The topicmap specification was designed so that subsets of XTM could easily be created. It is very easy to turn an XFML document into an XTM document (a topicmap) but it doesn't work the other way round. So, one can work with XFML and later in the future when one needs to do things XFML doesn't provide, but topicmaps do, one can easily convert the XFML map to a topicmap.

6. XFML, Standard for Distributed Information Architecture

In due course of time, people felt the need of standards that will allow web sites to share data with respect to their categorization, organization, and labeling. Further, creating standards for distributed information architecture would have allowed for easier and more effective combination of content, resources, and metadata across sites [5].

To understand the problem better, let us imagine a scenario where we have five sets of photographs in five different places, and may be also labeled under five different headings, but actually they are not five different things rather five different ways of looking at the same thing. The photographs may be of the same person who is father at one place, friend at another, husband at the third place, brother at the fourth place and boss at the fifth place. So, the photograph of the same person will be addressed in five different ways at five different places but any change in the address or status of the person will affect all of them equally.

By utilizing a standard for distributed information architecture, one can store aggregated information about any topic irrespective of their different locations and descriptions. The basic philosophy of XFML is to make real a distributed, loosely connected web of metadata. It gives us the freedom to choose topics for our XFML map and how do we want to organize the information.

XFML can be considered as a standard because it has many features that facilitate the development of distributed information architectures. It creates a loosely coupled net of published taxonomies where authors themselves can create and share taxonomies, merge XFML documents, mutually define metadata and facets, and import XFML from other authors. It came up as a format specifically for publishing and connecting faceted metadata between websites.

7. How Does XFML Work

Before going to the detailed specification of XFML, let us understand its working in a broad manner [6]:

- First we find another site with similar topics which may be of our interest, this should also publish its data in XFML format to allow linking.
- Then we configure the XFML software to trace their map once, per specified-time-period.
- In our XFML map, we use the connect element to link topics on our map to identical / very similar topics on the map of the othe website. For example, we may have a topic called “Classification” while they have one called “Knowledge Organization “. The meaning is the same, so a link can be created between them.
- Now, whenever the other site publishes something on the topic on their site, then the software at our end automatically links to it as “related reading” on our site. This is how XFML works and facilitates metadata sharing across web sites.

8. XFML Specifications

It can be divided to three categories **[7]**:

- A set of concepts, i.e. a conceptual model.
- An XML format for expressing these concepts, and
- A set of processing instructions that explain how applications should work with XFML data.

The availability of one or more XFML documents on a website can be indicated by a particular logo indicating link to the XFML document. In case of multiple XFML documents, one can have multiple buttons, although a page explaining the differences is always a good idea.

An XFML document is a valid, well-formed XML document, and conforms to the XFML DTD and XFML Core specifications. Example of an XFML document:

```
- <xfml version="1.0" url="http://xfml.org/spec/example.xml" language="en-us">
<!—This document conforms to XFML Core. See http://purl.oclc.org/NET/xfml/core/—>
- <mapInfo>

- <publisher>
  <name>DRTC</name>
  <url>http://drtc.isibang.ac.in</url>
</publisher>

- <generator>
  <name>word processor document</name>
</generator>
</mapInfo>

  <facet id="Subject categories">subject categories</facet>
  <facet id="author">author</facet>

- <topic id="Knowledge management" facetid="subject categories">
  <name>Knowledge management</name>
  <connect>http://othersite.com/xfml.xml#18753</connect>
```

<psi>**http://www.cia.gov/cia/publications/KMresources.html**</psi>
<description>**Knowledge Management tools and techniques**</description>
</topic>

-<topic id="**Knowledge management tools**" facetid="**subject categories**" parentTopicid="**Knowledge management** ">
<name>**Knowledge management tools**</name>
</topic>

- <topic id="**ADIS student**" facetid="**Authors**">
<name>**ADIS student**</name>
</topic>

- <page url="**https://drtc.isibang.ac.in**">
<title>**Seminar volumes of DRTC**</title>
<description>**All the seminar volumes of DRTC can be accessed from here**</description>
<occurrence topicid="**authors**" />
<occurrence topicid="**Knowledge management**" strength="**1**" />
</page>
</xfml>

9. Compatible Softwares and Formats

Cardinal XFML Parser: is an XFML Core compatible XFML processor implemented in Visual Basic 6.0 and built upon the MSXML 4 DOM implementation. Cardinal provides an XFML abstraction to simplify the development of tools to create and consume XFML documents.

FacetMap : Facetmap is a system for managing faceted hierarchies. It was the first fully XFML compatible application, and lets you import XFML documents.

Drupal : the popular Drupal CMs supports XFML export.

Compatible formats : XFML documents can be converted to other formats. With some formats, some of the information gets lost; with some formats, conversion can only go one way; with others, full, two-way conversion is possible.

XTM : It is the XML expression of topicmaps. XFML is a subset of XTM, any XFML document can be expressed in XTM, but not the other way around. Work is going on to develop a stylesheet that transforms XFML into XTM.

RDF : Research is going on for a RDF serialization of XFML. Presently any XFML document can be expressed as RDF, but not the other way round.

XFML libraries are existing libraries with facets and topics one can easily copy when creating a new XFML file. Example is IPTC library, which is categorized on the basis of subject, genre, media and news item types.

10. Limitations of XFML

- Only parent-child relationships between pieces of content are possible; it limits the types of associations that could be made.
- It is not possible to have multiple languages (real languages like English, French, Spanish—not programming or markup languages) within an XFML document as in the case of XHTML.
- It is still being developed, so, XFML has to be created manually and is not supported in any editing software, or on any existing Web sites, only demos or examples available.
- Moreover it is in very nascent stage, so, there is no guarantee that it will develop as full-fledged standard.

11. Conclusion

XFML is focused on sharing indexing efforts with faceted metadata, this is important because creating metadata vocabularies is really hard, and indexing lots of pages is even harder, but, sharing these efforts can be made possible by using <connect> and <psi> elements of XFML. As XFML Core is a frozen standard, so we can safely implement it. Work on XFML 2.0 is going on, but that is a long way off, and it will be a language with a different purpose than XFML Core. In addition to this, tool support is also taking off, so, it will be very easy to use it in near future and the websites will be able to talk to each other in real sense about their metadata.

12. References

1. Official publication of XFML core. http://purl.oclc.org/NET/xfml/core/ (accessed on 20/9/04)
2. Exchangeable Faceted Metadata Language (XFML) as Fuzzy-Lightweight XTM and RDF. http://xml.coverpages.org/ni2002-06-04-a.html (accessed on 1/10/04)
3. eXchangeable faceted metadata language. Home Page. http://xfml.org (accessed on 2/10/04)
4. Software that supports XFML. http://xfml.org/software.html#compatibleformats. (accessed on 25/10/04)
5. Lash, Jeff. Standards for distributed information architecture.http://www.digital-web.com/columns/ianythinggoes/ianythinggoes2002- 09.shtml. (accessed on 20/10/04)
6. Pilgrim, Mark. Mark goes XFML http://simon.incutio.com/archive/2002/12/05/markGOesxfml. (accessed on 22/10/04)
7. Introduction to XFML. http://xfml.org/spec/1.0html#introductiontoxfml. (Accessed on 27/9/04)
8. Xfmllib, A python library for reading and writing XFML documents. http://diveintomark.org/projects/xfmllib/. (accessed on 20/11/04)
9. Dive into mark. http://diveintoaccessibility.org/ (accessed on 12/11/03)

About Author



Aparajita Suman has done ADIS from DRTC, Indian Statistical Institute, Bangalore and is presently working as Junior Research Fellow at Defense Research and Development Laboratory, Hyderabad
E-mail : s_aparajitha@yahoo.co.in, aparajita@drtc.isibang.ac.in

Content and Information Management with Special Reference to India

J C Sharma

Abstract

This paper describes the basic concepts and meaning of Content and Information Management. It elaborates what and why Information Management. Its advantages and challenges along with effects on Indian Scenario have also been discussed.

Keywords : Content Management, Content Management System

0. Introduction

It is very important to define content before discussing the content management itself. Word "content" is dependent on its context. Content is made up of terms, elements and things that have no meaning without a well-understood context. The simple meaning of content is information put to use. Information put to use when it is packaged, presented and published for a specific purpose. Content is not a single piece of information but a collection of pieces of information put together to form a cohesive whole. Books, newspapers etc. all have content. The web is no different because the sites are also made of articles, indices, graphics etc. properly organized and presented precisely.

The traditional goal of managing content was to get it published in papers. For a long time, technology did not matter for the publishing industry but in recent years it has become a key for its survival. Technology has brought about a significant change. It changed the publishing industry, business of content and the importance of content management.

1. What is Content ?

Precisely content is in essence, any type or unit of digital information that is used to populate a page. It can be text, images, graphics, video, sound etc.

2. What is Content Management ?

Content management is effectively the management of the content by combining rules, process or workflow such as centralized web masters and decentralized web authors/editors who can create, edit, manage and publish all the content of a web page in accordance with a given frame work or requirement.

Management is the process of organization, planning, command, coordination and control to achieve defined objective. Content management is effectively collecting, managing and making information available in targeted publications. In other words, it is a discipline that involves the collection, management, and publication of content with clearly defined rules, methods, documented workflows, and applicable tools and techniques with effective publishing system.

3. Content and Information Management

Management of experts database includes collection of interrelated data, set of programmes for the data base access, search mechanism, content security and data validation. Content management involves

both the definition of structure, which forms the core of database, storage policy and provision of mechanisms for content manipulation.

4. Concepts of Content Management

It includes the understanding content domain, from which all of the structural decisions flow. The notice on contents components allows content processes (collection, management & publication) to be automated. Target publications are the end result of any content system. A framework unites all of content into a single system of Meta information.

5. Need for Content Management

A content management system helps organize and automate collection, management, and publishing processes. CMS is needed because:

- Of need to process large amount of information effectively.
- Since information is changing too quickly, only systematic process can provide solution.
- Some times there is requirement to publish more than one publication from a single base of content.
- Of need to make the design of a publication static to maintain uniformity and time saving
- Of need to make content dynamic, versatile, and powerful.

6. Content Management System

The system itself is definable as a tool or combination of tools that facilitates the efficient and effective production of the desirable web pages using the managed content.

To combine all three, we can say:

"A CMS is a tool that enables a variety of technical (centralized) and non-technical staff (de-centralized) to create, edit, manage, and finally publish a variety of content. These include text, graphics, video, sound, etc. and are being constrained by a centralized set of rules, process and, workflow that ensure a coherent, validated web site appearance."

A content management system helps organize and automate your collection, management, and publishing process and is needed when this process becomes too complex to manage manually. The need for a content management system can be assessed by the amount of content, the amount of change in the content, and the number of publications intended for creation. A library, archive, or museum management or cataloguing system, a picture library system, a word processing or other text file containing lists of digital resources; a presentation file, PowerPoint file, or a Multimedia application is not a content management system.

7. Knowledge Management and Information Management

Knowledge Management is the process of transforming Information and Intellectual assets into enduring value. It connects people with the knowledge that they need to take action. In corporate sector, managing knowledge is considered as a key to achieve competitive advantage breakthrough.

What's difference between Information Management and Knowledge Management? Both concepts refer to managing (handling, directing, governing, controlling, coordinating, planning, organizing) processes and the products of these processes. In addition, since knowledge is a form of information, it follows that Knowledge Management is a most robust form of Information Management that provides management of activities not generally available in Information Management.

One difference between basic Information Management and Knowledge Management is that basic information management focuses in managing how information is produced and integrated into the enterprises; while Knowledge Management does the same with respect to knowledge. A second difference between basic Information Management and Knowledge Management is that basic Information Management focuses on managing more narrow set of activities than Knowledge Management. The two information processes managed by an organization are information Production and Information integration whereas the two basic knowledge processes are Knowledge Production and Knowledge Integration.

Some of those who have tried to define Knowledge Management in relation to librarianship, Information Management and /or Information Resources Management concede that there is much about Knowledge Management that may arouse a sense of deja-vu among many information professional.

8. Advantages of Content Management

- Advantages of content management are many. Depending on the contributor, creator, website administrator, the advantages are enumerated as follows:
- Anytime, anywhere web publishing: content management helps contributors change content whenever and wherever necessary.
- Faster updating : updating content is faster.
- Efficient workflow management : with content management organization have mechanism to control authoring, workflow publishing and document management functions.
- Process flow web content : new content comes from both content contributors and existing corporate databases.
- Eliminate content bottlenecks : A good content management solution extends the responsibility for updating website content to business users. As a result, new content is no longer piled up, for web site content has become more valuable since it addresses their special needs.
- More valuable content : use of a content management solution encourages closer relationships with customers, partners, vendors and especially employees. These groups find that their web site content has become more valuable since it addresses their special needs.
- Increased savings by empowering non –technical, lesser -paid business users to self publish.
- Site consistency: enforces compliance with corporate publishing standards.
- Gives visitors access to more timely and valuable content: because it can be changed quickly and easily.
- Encourages longer site visits because of more in- depth and useful content.

9. Challenges of Content Management

Developing a centrally controlled, distributed content management system is a challenging task. Integration of internal and external information is needed. Organizing the content for efficient information access is

required task. Content should provide context for searching and search results. Bringing uniformity and consistency in content authoring, publishing and presentation is need of the hour. Content management should provide personalized services.

10. India Scenario

In India there is a lot of scope for the librarians to become content managers by taking advantage offered by the technology for publishing and distributing content. Today the community at large wants web based information sources due the location, time, accuracy and speed advantage. Gradually the users are turning away from the conventional library systems. The need is to provide information on their desktops.

Publishing industry, personal publishing, research organizations, and government are engaging themselves in publishing e-content.

11. Future Scope and Conclusion

As in any content management system, the success and level of utilization of the database is highly dependent on the quality, quantity and precise access to the database. Hence, initiatives have been taken to create profiles of experts in various R & D organizations. New search techniques are to be incorporated for speedy access to the database. Web enabled access to users have been provided over Internet to extend the availability of this information both for content creation and retrieval. In order to make this web – site more interactive, discussion for chat rooms are planned, where the users can post their queries or can chat with the experts on line.

CMS products are available, as well as there are developers who can build a customized CMS. This discipline is gradually growing, and with its growth easy-to-use solutions are likely to emerge. Comprehensiveness and flexibility required to deliver the appropriate publications are very important conditions for choosing the system. Electronic publications –the web has given tremendous boost to the growth of content management. In the competitive world and age of knowledge economy, it is critical for communicating on a large scale. Content management system has enabled to control information. Now information can be delivered in a manner that produces a richer, more timely, more targeted experience for audiences, and a rational, cost-effective process for the publisher.

Content management solutions can help companies achieve their corporate and organizational objectives for three types of networks viz. Internet, intranets and corporate portals. By offering timely and therefore more valuable content, companies increase the numbers of repeat visitors and ultimately increase revenues. Affordable content management solutions that offer web authoring, editing and publishing capabilities to non-technical staff are now available. These solutions are designed to help with web creation and to assist with planning, coordinating and tracking site changes.

12. References

1. Bratton (John) and Gold (Jeffery) Human Resources Management: Theory and Practice – Palgarve. 2nd Ed. 1999. 336,337.

2. Goyal SL, Rajneesh (Shalini) Management Techniques: Principles and Practices. – Deep and Deep Pub.200. 130,133.

3. Gangathran (M) Information Resources in the 21st Century. SRELS Journal of Information Management. Vol. 41 No. 1. March 2004.

4. Gopinath (MA) Knowledge Management Policies Options. SRELS Journal of Information Management. Vol. 41.No. 2. June 2004.

5. Jamabhekar (Ashok) Content Management an Overview. Internet Engineering for Library and Information Centers. CALIBER-2002

6. Lowe (Pqul) The Management of Technology : Perception and Opportunities- Chapman & Hall. 1995. 79 to 81.

7. Rue (Leslie W) and Byars (Lloyd L) Management: Skills and Applications-Irwin.1997. 437 to 439.

8. Sreekumar (MG) and Gopinath (Saji) supply Chain Management of Information. Information Services in a Networked Environment in India. CALIBER-2000

About Author

J C Sharma is working as Assistant Librarian at National Institute of Technology, Kurukshetra, Haryana**.** He has presented number of papers in seminar, conferences and journals. He is also a member of many professional bodies.

Email : jcsharmain@yahoo.com

DLIST: Distributed Digital Management of the Scholarly Publication

Kamalendu Majumdar U N Singh

Abstract

In present scenario publishers of freely available electronic journals have the potential for positively transforming the highly problematic economics of scholarly publishing. But actual use of electronic scholarly publications shows that they present serious utilization barriers to would-be readers. Practical awareness, indexing, and archiving in this new literature can overcome these barriers and is essential if such promising publications are to transform scholarly communication. DLIST (Digital Library for Information Science and Technology) can meet these needs by integrating electronic publications into existing information systems. Thus disintermediation of the scholarly publishing process will help the clientele of the low budget library in all respects.

Keywords : Digital Libraries, D-Space

0. Introduction

The vision of the digital library is not new. This is a field in which progress has been achieved by the incremental efforts of numerous people over a long period of time. However, a few authors stand out because their writings have inspired future generations. Two of them are Vannevar Bush and J.C.R.Licklider. In July of 1945, Bush, then director of the U.S.Office of Scientific Research and development, published an article titled “As We May Think” in the Atlantic Monthly This article is an elegantly written exposition of the potential that technology offers the scientist to gather, store, find and retrieve information. Much of his analysis rings as true today as it did 50 years ago. The Atlantic Monthly has placed a copy of “As We May Think” on its web site. Anyone interested in libraries or in scientific information should read it.

In the 1960s, Licklider was one of the several people at Massachusetts Institute of Technology who studied how digital computing could transform libraries. Like Bush, Licklider was most interested in the literature of science; however, he foresaw may developments that have occurred in modern computing.

1. New trend

Library gurus invariably have long lists of difficult issues to confront. These days, high on my list is the future of our university libraries. Although libraries form the basic infrastructure of the academic endeavor, I have come face to face with an unhappy fact: University librarians are now being forced to work with faculty members to choose more of the publications they can do without. The ballooning costs of academic publications are preventing faculty members and researchers from gaining access to the world’s scholarship and knowledge.

2. Hurdles

Institutions are facing unprecedented budget crises just as expanding faculties and student bodies are increasing the demand for scholarly information. Even in the best of economic times, university libraries cannot hope to keep pace with the 6- to 12-percent annual inflation rate in the price of scholarly journals. And the fiscal environment today is particularly difficult. Without proper user of the huge bandwidth and the classical internet architecture, neither university librarians nor faculty members can deal with the challenges of preserving access to scholarly resources.

3. Worldwide efforts

In this regard, the librarians should consider several strategies, including development and support to new models of scholarly publishing that can cut the costs of distributing and retrieving information. Here it can be mentioned that several organizations are experimenting with less-expensive ways to disseminate faculty research. Some of them are already well known, like

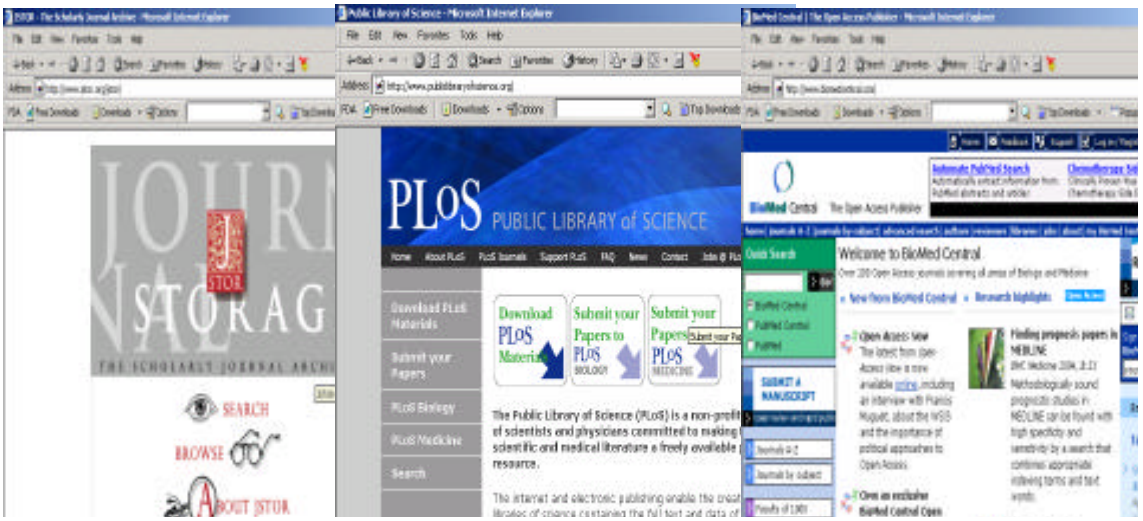


Fig: I Some online portal for fair use.

JSTOR, which digitally archives more than 300 journals in various disciplines, and Stanford University's HighWire, which stores online several hundred journals in biology, physics, and other sciences. Others, like BioMed Central and the Public Library of Science in both biology and medicine, are only just emerging. The same strategies should be applicable to our country.

4. DLIST: Present Status

IIT was established with the objective of taking the country to technological excellence that will catapult the economy to the big league. Half a century has passed and we can proudly say that it is a goal fully accomplished. However, we have miles to go. Online publication is perhaps such an area**.** We anticipated to have a digital library which is able to deal with the e-collections e-persons and e-groups. At the same time it should enable with OAI (Open Archives Initiative) along with the support of DC (Dublin Core). Here we want to mention that an OAI is a protocol for metadata harvesting. This allows sites to programmatically retrieve or harvest the metadata from several sources, and tender services using that metadata, such as indexing or relating services. Such a service could allow users to access information from a large number of sites from one portal.

5. Motivation

During 2002 we received a project from MHRD with a objective of developing a Real time Multimedia for distance education with an objective of creating a virtual classroom by setting a multimedia digital library server named Ekalavya at IIT, Kharagpur and to provide linkage to intellectual contents through a high-end real time multimedia server.

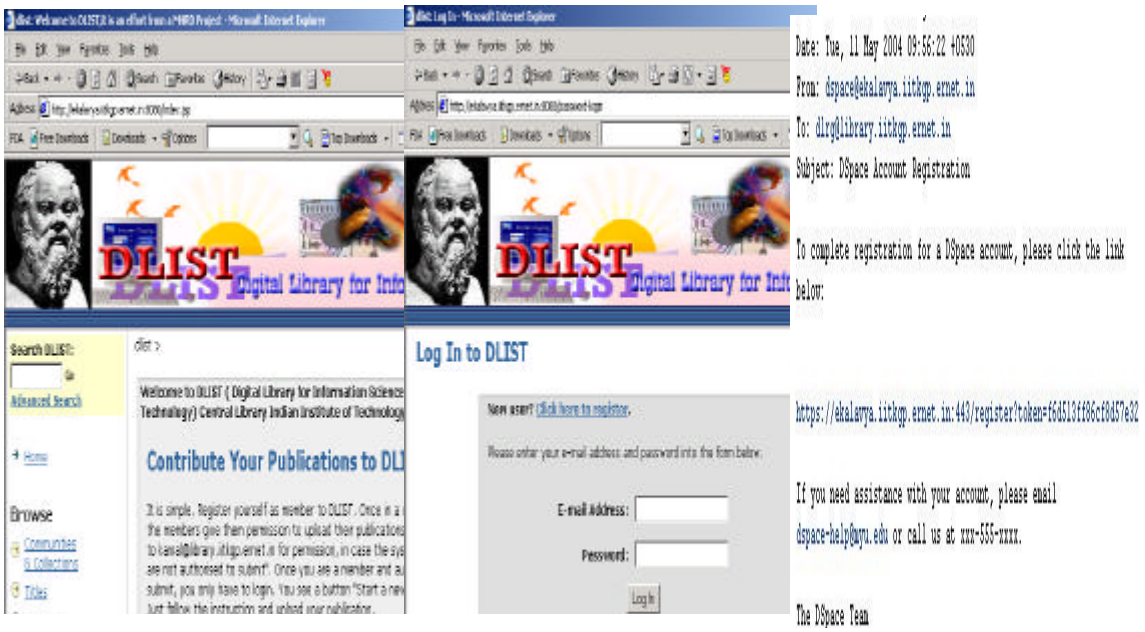


Fig: 2 Some snap sort of the IIT portal .

We are always fond of open source software. In the mean time we heard of Dspace, a digital library software platform jointly developed by Hewlett-Packard Company and MIT. In this article the authors intent to discuss the plan and success of the implementation of DSpace at Central Library, IIT, Kharagpur.

The following open sources s/w were used

•j2sdk-1_4_2_02-linux-i586	•httpd-2.0.43	•Redhat Linux 8.0
•jakarta-tomcat-4.1.27	• jk-1.2-src_connector	•openssl-0.9.7c
•apache-ant-1.5.4-bin	•dspace-1.1.1	•postgresql-7.3.4

How the portal actually works

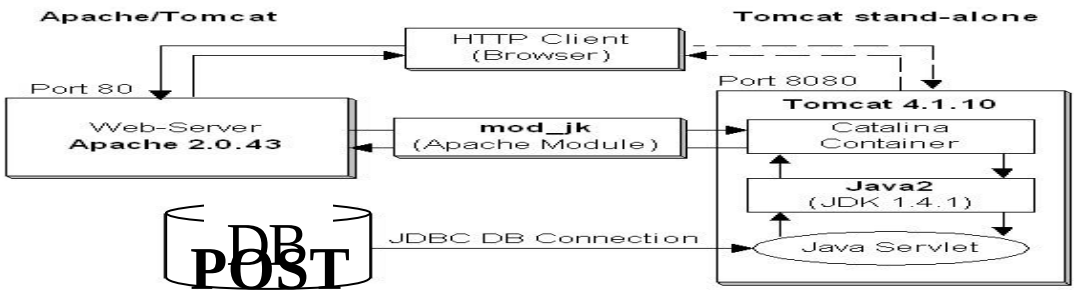
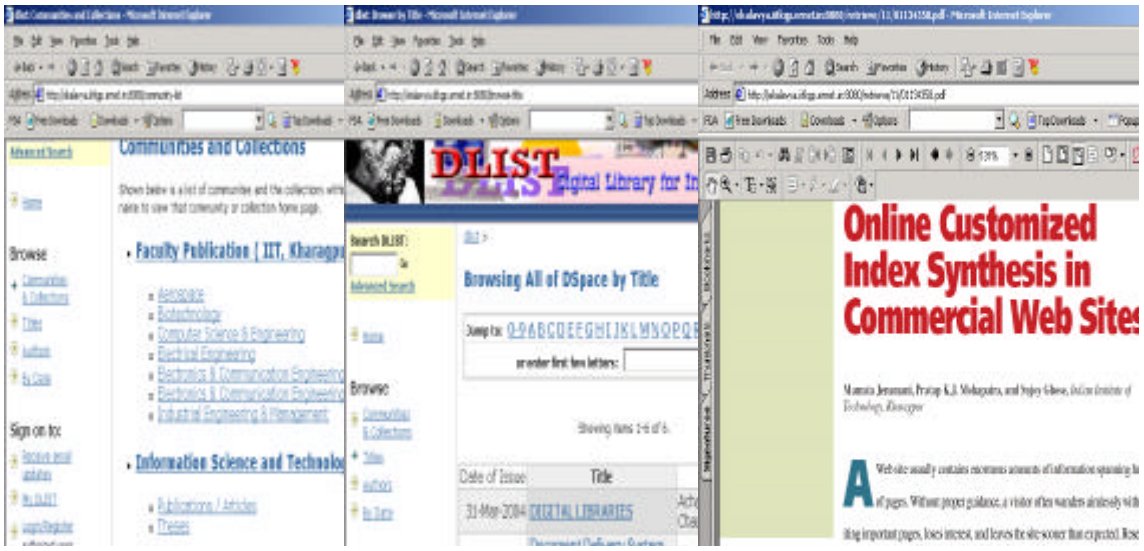


Fig: 3 How the portal works .

A complete portal has been designed, developed and deployed by a small team of IIT ,Kharagpur Library as a prototype.Document has been uploaded for test run. It has been demonstrated to the fullest satisfaction that it works like any other electronic scholarly publication. The portal at IIT, Kharagpur is capable of

- Creating e-collections, e-persons and e-groups.
- Supporting the digitization of in house resources (e.g. institute theses)
- OAI (Open Archives Initiative) with the support of DC (Dublin Core).
- Programmatically retrieving or harvesting the metadata from several sources.
- Providing various indexing services using those metadata



Creation of Community Metadata search Full text

Fig: 4 Creation of community/metadata search / access of full test document from the portal.

We have intention to help those libraries who have vary small budget and are not capable to have the costly online database and the Scholarly Community who also wants to discuss the professional issues and want to disseminate their scholarly writings to the users from various schools or colleges . We propose to publish a series of publications to give the total idea from installation of DSpace to the Organization of DSpace. In this regards this is the first effort.

6. Methodology

Fortunately we came across the IBM software products the “Quick Start” manual. This manual introduced the product through a series of common tasks, from installation to basic configuration. The manual did not go deep into the details of the software — that’s what the User Guides and other documentation were

for — but it did help new users get idea about the product quickly, and without scaring them too much. The Jakarta group produces a lot of documentation for Tomcat, but none of it seemed to be as soothingly simple to read as the "Quick Starts" and in case of DSpace, most likely the developers have an impression that the user of DSpace would have a certain echelon so they did not focus about the detailed installation of DSpace . Hence this article will prove useful. Hopefully, we have addressed the common issues facing new installation in such a way that their learning curve becomes easier to negotiate.

First we install the operating system Redhat 8.0 and on top of that we want to install Apache2

As the home of apache2 is in our case /usr/local/apache2

#tar xvfz httpd-2.0.47.tar.Z

#cd httpd-2.0.47

#./configure —prefix=/usr/local/apache2

#make

make install

It creates httpd directory at /usr/local/apache2/conf. We made the change in httpd.conf file variables Now question is how the system can know that instead of usual position (/etc/httpd/conf) httpd.conf file now we have the httpd.conf in a separate location i.e. /usr/local/apache2/conf/httpd.conf. To face these problems we have to change path of the domain at /etc/init.d/httpd. At this juncture we are also encountering error message

#/etc service httpd start

service httpd service **failed**

we have struggle a lot at last we delete the directory " httpd" under /etc

#rm –R httpd

here also we are getting error message then we delete the .exe file of httpd under /sbin/httpd and copy the same from /usr/local/apache2 to /sbin/ and now apache2 is working fine.

Now here we want to depicted the necessity of another webserver like program called tomcat. Tomcat, sometimes known as Apache Tomcat, sometimes as Jakarta Tomcat, is a Java servlet engine that is the reference implementation used by Sun for its Java servlet and JSP specifications. If you want to know details about tomcat please logon to http://jakarta.apache.org/tomcat,

7. Installation of Tomcat

tar xvfz jakarta-tomcat-4.1.27.tar.gz

mv jakarta-tomcat-4.1.27 /usr/local/

cd /usr/local

ln -s jakarta-tomcat-4.1.27 tomcat -> (sysmlink for tomcat)

Now problem is how apache and tomcat will interact with each other. Easiest way is to use a connector which will act as an agent between this two web servers. All the clients will send their requests to Apache and in process Apache will send these requests to Tomcat through the connector . There are several connectors developed by many professionals . We want to use mod-jk as connector. Before proceeding further you have to stop Apache and Tomcat .

Installation of mod_jk from rpm

rpm -ivh mod_jk-ap13-1.2.2-1jpp.i386.rpm

now you have to edit three configuration files viz. httpd.conf, workers.properties and server.xml (see Appendix 1)

httpd.conf

After mod_jk has been successfully installed, a set of entries will be appended to the end of Apache's configuration file, httpd.conf. Look for the following lines at the end of the file :

```
<IfDefine HAVE_JK>
    LoadModule jk_module modules/mod_jk.so
    AddModule  mod_jk.c
    Include /etc/httpd/conf/mod_jk.conf
</IfDefine>
Include /usr/local/apache2/conf/mod_jk.conf
RedirectMatch ^/$ https://ekalavya.iitkgp.ernet.in/
RedirectMatch ^(/[^o].*) https://ekalavya.iitkgp.ernet.in$1
RedirectMatch ^(/.[^a].*) https://ekalavya.iitkgp.ernet.in$1
RedirectMatch ^(/..[^i].*) https://ekalavya.iitkgp.ernet.in$1
```

mod_jk.conf

This file is crucial to the integration, but before we go into it, we must first understand a few things about it. In mod_jk.conf we need to do the following :

- Define where the workers.properties file is, and what it is called
- Define where the log file is, and what it is called
- What kinds of messages to record - errors only, or more ?
- Whether to enable SSL or not
- Define the web application contexts

At this point, it is a good idea to open up the mod_jk.conf that comes with the mod_jk rpm. We will examine each of the directives in the file, relating them to the tasks outlined above.

JkWorkersFile

Value: /etc/httpd/conf/workers.properties

The location of workers.properties, a file that we will examine next.

JkLogFile

Value: /var/log/httpd/mod_jk.log

The location of the log file. Depending on the error level you specify, the file will contain either very verbose information or just the critical errors.

JkLogLevel

Value: error

There are 3 available options here :

- info - logs information about mod_jk's activities
- error - in addition to activity information, errors will also be logged
- debug - logs anything and everything. As its name implies, this option is ideal when troubleshooting.

Following the first 3 lines of mod_jk.conf, we encounter a series of entries that relate to SSL. We will not go into this at the moment, because we want to get a "plain-vanilla" Tomcat-Apache integration going successfully first.

What follows after the SSL section are the "contexts". We have already seen earlier that "contexts" refers to web applications that are deployed inside Tomcat. We specified a <Context> element inside Tomcat's server.xml for every web application deployed, and we have to do the same here, because Apache needs to know how to hand-off requests for web applications to Tomcat.

For every web application, we must define the contexts to Apache, and we do this by supplying the following information :

- Where the web application is in the file system and how do we map it to a URL ?
- What additional options do we want to enable for the web application ?
- How do we "mount" it in Apache ?

This is the web application and servlet which is working properly.

dlist Web Application

Web Application Name	dlist
Location	$CATALINA_HOME/webapps/dlist
URL we want to map to	http://hostname.domain.com/ dlist/
mod_jk Worker Name (defined inside workers.properties)	ajp13

How we express this as a context in mod_jk.conf as shown below :

```
#
# The following line makes Apache aware of the location of the /dlist context
#
Alias /dlist "/usr/local//tomcat/webapps/dlist "
<Directory "/usr/local/tomcat/webapps/dlist ">
Options Indexes FollowSymLinks
</Directory>
#
# The following line mounts all JSP files, the /servlet/ uri, and all files to Tomcat
#
<VirtualHost>
```

```
JkMount /dlist/*.jsp ajp13
JkMount /dlist/* ajp13
JkMount /*.jsp ajp13
JkMount /* ajp13
JkExtractSSL On
 Jk HTTPS Indicator HTTPS
 JkSESSIONIndicator SSL_SESSION_ID
 JkCIPHERIndicator SSL_CIPHER
 JkCERTSIndicator SSL_CLIENT_CERT
# End of Tomcat mod_jk directives
#Include /usr/local/apache2/conf/mod_jk.conf

</VirtualHost>
```

That completes our configuration for mod_jk.conf. We have two more files to edit before we are ready to begin testing.

workers.properties file will be like this :

```
worker.list=ajp13
worker.ajp13.port=8009
worker.ajp13.host=ekalava.iitkgp.ernet.in
worker.ajp13.type=ajp13
```

Now the most critical part of the installation is D Space installation and this has been described in Appendix 1.

8. Conclusion

The faculty member who chooses alternative ways to disseminate his research should be recognized and rewarded by fellow colleagues. The rapid emergence of scholarly electronic publishing challenges our traditional methods of assessing professors' work for tenure and promotion purposes. We should take steps to guarantee that our evaluation practices keep pace with the adoption of new communication technologies. At the University of California, for instance, the Academic Senate supports consideration of electronic publications in academic peer review.

At the same time, we must not jeopardize the health or well-being of the scholarly societies and university presses that play so critical a role in academic life. Faculty members should continue to manage their intellectual property and copyright. They should decide which publishing organizations they will review, edit, and write for. When signing a publishing contract, they should determine whether to assign the publisher copyright and whether to seek a nonexclusive right to disseminate their work freely in an electronic form.

9. Reference

1. http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm#DefinitionsConcepts

About Authors



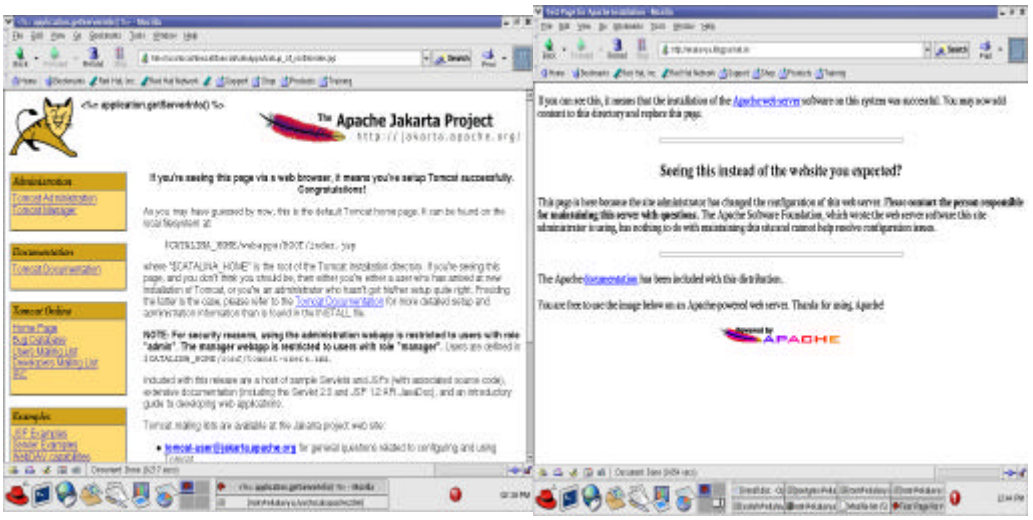
Mr. Kamalendu Majumdar Presently he is working as Assistant Librarian at Central Library , Indian Institute of Technology, Kharagpur, India. He holds MLibSc and BLibSc from IGNOU in the year 1994 and 1996. Currently he is going on under research program from Guru Ghasidas University. Bilaspur. His major research include multimedia information retrieval, wireless information retrieval, query processing and optimization in multiprocessor and distributed systems, and database management system and specialy interested in computer network . He has published over 35 papers in conferences/ journals in international and national level. He has also edited IMeL2002 conference proceeding as a member of editorial committee.
Email : kamal@library.iitkgp.ernet.in



Prof. U N Singh Presently he is the Head, Dept. of Library & Information Science and Dept. of Information Technology, G G University Bilaspur, CG. He holds M.Sc and BLibSc from BHU, Varanasi, UP in the year 1979 abd 1980, Associateship in Information Science from INSDOC IN 1983, and Ph.D in Information Science in 1992 from B I T Mesra, Ranchi. His field of interest includes information science, Information storage and retrieval, and query processing , Computer network, Data and file structure, Bibliometrics, Scientometrics and Library Automation. He has published over 29 Conference/ Journal papers in international and national conferences and journals. He has guided Six Ph.D students and many MLibSc students for their project works. He is also responsible for starting many new PG level courses in Information Technology in BIT Mesra and GG University, CG. He is also member of Board of Studies and Ph. D Committee in several Indian universities.
Email : unsingh03@yahoo.com

Appendix 1

Now the most critical part of the installation is to install the D Space. Now we have to home page Tomcat and Apache separately and both are running



Installation of Dspace

Download the binary package—> **dspace-1.1.1.tar.gz** to /usr/local as root user

]# mkdir /dspace

]# chown dspace:dspace /dspace

]# cp /usr/local/dspace-1.1.1.tar.gz /dspace

]# chown dspace:dspace /dspace/dspace-1.1.1.tar.g

Log in as dspace user

]$ cd /dspace

]$ tar xvfz dspace **dspace-1.1.1.tar.gz** (You will get **dspace-1.1.1-source** directory)

Customise the dspace.cfg file or leave this step here...

Build & install The dspace source:

Log in as dspace user

]$ cd /dspace/dspace-1.1.1-source

]$ ant

]$ ant fresh_install

Connect Dspace webapps to tomcat

]# cd /usr/local/tomcat/tomcat/webapps/

]# mv ROOT ROOT.bak

]# ln -s /dspace/jsp ROOT

]# ln -s /dspace/jsp dlist

]# ln -s /dspace/oai dspace-oai

]# chown -R dspace:dspace ROOT dlist dspace-oai

Install the Config files:

Log in as dspace

]$ cd /dspace/bin

]$ ./install_configs

Create The administrator account.

]$ cd /dspace/bin

]$ ./create_administrator

Initialise Lucence search indicies:

]$ ./index_all

Set up email subscription feature. In the dspace user's crontab insert the following:

]# su – dspace

]$ vi emailsub

Send out subscription emails at 01:00 every day

0 1 * * * /dspace/bin/sub-daily

]$ crontab emailsub

Customise the dspace.cfg file :

```
]# vi /dspace/config/dspace.cfg
dspace.url = https://ekalavya.iitkgp.ernet.in:443
           = http://ekalavya.iitkgp.ernet.in:80
]# vi /dspace/config/dspace.cfg
dspace.hostname = ekalavya.iitkgp.ernet.in
dspace.name = dlist
```

Note :When ever u write any thing in /dspace/config/dspace.cfg

U have to run install-configs from /dspace/bin

```
]# cd /dspace/bin
]# ./install-configs
```

Edition in server.xml in tomcat

```
]# cd /usr/local/tomcat/tomcat/conf
]# vi server.xml
add the following    lines between Host tags i.e   <Host>   </Host>
 <Context path="/dlist" docBase="dlist" debug="0">
       <Resources className="org.apache.naming.resources.FileDirContext"
      allowLinking="true" docBase="" />
</Context>
<Context path="" docBase="ROOT" debug="0">
     <Resources className="org.apache.naming.resources.FileDirContext"
      allowLinking="true" docBase=""/>
</Context>
<Context path="/dspace-oai" docBase="dspace-oai" debug="0">
     <Resources className="org.apache.naming.resources.FileDirContext"
      allowLinking="true" docBase="" />
</Context>
```

Edition in httpd.conf file in apache.

Add the following lines between the <VirtualHost 10.17.200.1:443> ...</VirtualHost>

```
]# vi /usr/local/apache2/conf
]# vi httpd.conf
 Alias /dlist "/usr/local/tomcat/jakarta-tomcat-4.1.30/webapps/dlist"
 <Directory "/usr/local/tomcat/jakarta-tomcat-4.1.30/webapps/dlist">
    Options Indexes FollowSymLinks
  DirectoryIndex index.html index.html index.jsp
 </Directory>
 JkMount  /dlist/*.jsp  ajp13
 JkMount  /dlist/*   ajp13
 JkMount /*.jsp ajp13
 JkMount /*  ajp13
```

]# vi httpd.conf
U also need to set up apache to understand the mime-type:
add the line :
AddType text/jsp .jsp

Note: -> At this point make sure that ur postgresql is running

login as postgres
]# su – postgres
]$ cd /usr/local/postgres/post/bin
]$ ./postmaster -i -D /usr/local/postgres/post/data

Stop the tomcat , apache & Start the tomcat ,apache

]# /usr/local/tomcat/tomcat/bin/startup.sh
wait for 30 seconds
]# /usr/local/apache2/bin/apachectl startssl

Now u can access the dspace home page on the following URLs

https://ekalavya.iitkgp.ernet.in
https://ekalavya.iitkgp.ernet.in/dlist

Congratulations..... Dlist home page

Content Management in Digital Libraries

Mohd. Nazi Faizul Nisha

Abstract

Discusses briefly the concept and characteristics of digital library. Digital library is simply an online system providing access to a variety of contents such as various kinds of electronic media(text, image, video ed),licensed databases of journals, articles and abstracts and description of physical collections. The functions of content management such as selection and acquisition,indexing,storage,retrieval,maintenance and intellectual rights are discussed. Issues regarding the research in the development and management of digital contents are highlighted

Keywords : Content Management, Digital Libraries

0. Introduction

Libraries have existed for centuries and since then they had been managed as warehouse of documents by acquiring, cataloguing and classifying books , journals and other materials and circulating them to their clients. But recent developments in Information Technology (IT), the Internet, World Wide Web (WWW), coupled with increase funding for research on creation, access and management of electronic information resources, have led to the development the new era of electronic and digital libraries. These technological innovations have improved the new breed of information professionals to select, organize, retrieve and transfer digital contents effectively and efficiently to their target audience (1).

1. Digital Libraries : Meaning and Nature

The term electronic library, digital library and virtual library have bee used interchangeably and now widely accepted as description of the use of digital technology by libraries to acquire, store, conserve and make available their content to remote users.

In a broad sense, digital library may be defined as an organized and managed collection of high quality information contents in a variety of media (text, still image, moving image, sound or combination thereof), but all in digital forms accessible over different electronic networks. Such a digital library includes a number of search or navigation aids that operate library and allow access to other collection of information connected by network world wide.

The term digital library is best defined by Christine Borgman (2) as a set of electronic resources and associated technological capabilities for creating, searching and using information … they are an extension and enhancement of information storage and retrieval system that manipulate data in any medium.

The concept of digital library is rooted in age-old dream of creating a virtual library. But digital library is different from virtual library because of its physical identification. O'Donnel (3) differentiates digital library from virtual library as it can still maintain a physical presence, whereas virtual library is a vast, ideally universal collection of information and with instantaneous access to that information wherever it physically resides.

2. Digital Library Contents

The most important components of digital library, however, is digital collection. Viability and extent of usefulness of a digital library would depend upon the critical mass of its digital contents. The information contents of a digital library include virtually any kind of electronic media(Text, image, graphics, video, etc.), licensed databases of journals, articles and abstracts and description of physical collection.

Theoretically any object from a text fragment to an animal in zoo may be rendered digitally and thus, there is no limit to the types of contents that may be held by a library. But in practice, digital contents are of three types:

- Contents created and existing primarily in machine readable format.
- Contents converted from the traditional format into digital(e.g., print text, pamphlets, manuscripts, motion pictures and recorded sound)
- Access to external contents, not held in-house, by providing pointers to web sites, publisher's services, password to consortium or other collaboration from commercial organizations

3. Management of Digital Library Contents

Contents in digital library are organised and managed for the purpose of immediate access to the target audience. How contents are developed and managed is a critical issue to the long-term success of digital library services, especially when technical resources are limited. Content management includes the following key functions:

3.1 Selection and Acquisition

Libraries select contents according to a well defined collection development policy. Such policy manifests the mission of a library and determines how budgets on materials are expended. There are two key challenges in content selection i.e. cost and quality. As soon as decision about selection is made, content must be acquired. For objects, which are already in digital form, the file transfer through networks or mass storage is straightforward as long as file formats are well specified. In case of traditional objects, digitisation must be done. Scanners for text and images range in quality on several dimensions (i.e. output resolution, value and conditions of physical objects and speed) are required. In addition to these technical challenges, policy decision must be made. For example, which resolution and formats to adopt, how much text to OCR error is acceptable, how to link different representation for multiple media from single collections.

3.2 Indexing

Once content has been selected and acquired, it will be added to the collection in such a way that users may retrieve it easily. And thus, indexing is required for digital content to search and access in a selective way like OPAC for printed content. Decisions are to be taken regarding what to be indexed (author. keywords, phrase, etc), how the content and index files are linked, what sort of access points are provided, etc. Indexing strategy comprises not only the types of fields are to be indexed, but how they are to be treated (exhaustive or sparse).

Automatic indexing techniques are used to index the content of digital library. Several www-based services use a hybrid approach by manually creating classification system and then using automatic techniques to assign objects. Most retrieval systems for images, video, audio recordings and other non-textual objects have depended on items such as title, creator name or manually assigned subject headings for retrieval.

It seams certain that digital library research and development activity of 1990s will ensure that considerable progress is made in automatic indexing for textual and non-textual objects. New indexing challenges will emerge as more dynamic objects (e.g., virtual conference proceeding, active networks) are added to digital libraries. The temporal nature of such objects will require ongoing indexing techniques.

3.3 Storage

The next thing is how to store the content of digital library. Decisions regarding procuring suitable hardware, software, networking, etc. are to be made at this stage. Storage is mainly a technical requirement, although new media may complicate storage decision and costing. When data are to deliver continuously (e.g., streaming video or audio) rather than as discrete files, alternative technologies are required (4). Large digital repositories are required like multiple level of mass storage media (e.g., disk, tape, etc) and mechanical robots to locate and mounts the media. Various supercomputer centers are using tape robots that store and access to many terabytes of data. Digital libraries will surely apply such technology just as libraries of today apply movable shelving and complex conveyer systems to move physical materials.

3.4 Retrieval

Retrieval is another major issue, as far as digital library content and its access is concerned. Ultimately, users must be able to retrieve the content, which have been selected, indexed and stored by the librarians. During 1970s to 1980s, a large number of libraries invested heavily in computerizing cataloguing and circulation functions to give users better access and services. Online Public Access Catalogues (OPACs) have long provided author, title, and limited subject access to local holdings (and more recently to union holdings across to multiple libraries). The expectation for digital collection is that catalogue should seamlessly link to the digital content itself so that remotely located users can find and display not only bibliographic records, but also primary information objects.. In physical libraries, the card catalogues or OPAC is physically distinct from the items on shelves. These distinctions are difficult to make in electronic environments because everything is displayed on the same physical screen and thus, the boundaries between metadata and primary data are often blurred. Expectation to provide primary data with metadata yields several challenges to librarians. The challenges are first to extract and provide multiple levels of representation and second to provide users with control mechanisms to move from high -level surrogates to detailed objects (5). Today most retrieval is facilitated through words, titles, captions, manually created subscriptions, automatically extracted extracted keywords and so on. There is enormous attention focused on creating non textual surrogates such as colour and shape characterisation for images and speaker identification schemes for audio recordings, but there are more difficult metadata issues looming as more contents are not stored at all but created on the fly according to the specification of the users.

3.5 Maintenance

Maintaining buildings and systems and preserving content are important and costly activities in physical libraries. Digital libraries may avoid some of the cost of wear and tear on buildings and books but still have significant maintenance costs, including some unique to electronic environments. New equipments, improved or alternative network solutions (e.g. ISDN, ATM, Wireless), and software upgrades will require excellent technical personnel.

Just as the computational system changes, digital content may also change. A digital document may have numerous versions, especially given the ease with which electronic documents may have changed. Maintaining the most essential document requires that versions be well managed, which include updating and deleting the links to those objects (6).In addition to this version control problem, digital librarians must manage the multiplicity of indices and file formats. Requirements for link managements are more

problematic, as hypertext links are created among distinct documents. Although much research and development efforts in digital libraries have been devoted to maintaining the content. but further improvements are required to maintain security, updating versions, tools for automatically checking links, database tools for property rights, etc. for the smooth library functions and services.

3.6 Rights Management

Intellectual property right and information security and authority are two global interdependent issues, which influence research and development in digital libraries. Copyright exists to promote intellectual production by providing economic incentives. Security protects unauthorized access as well as ensures the veracity and authority of digital information objects. The misuse that can be put to digital content is far more serious and voluminous than for printed content. Efforts have been made to change copyright laws to protect the illegal use of digital objects and also to develop technical solutions that protect copyright either through copy protection or automatic billing mechanism. Research on encryption algorithms, digital watermarking and electronic commerce are leading to the development of trusted system that protect intellectual property rights by managing the necessary financial transactions while protecting consumers by providing authoritative information securely (7).These techniques ensure the veracity of an object and may help to prevent copying and distribution in an open market place.

4. Conclusion

There is no doubt about the utility of digital libraries as they facilitate live and interactive access to wide variety of content online. But the problems of managing digital library content and its development are manifold. Management of digital library content requires two prolonged strategies(i) to digitize local content;(ii) to devise options for accessing external resources. Generally there is a feeling that publishers copyright most of the contents available in our library, and we are not in a position to provide online access to those contents. Though our libraries are facing a shortage of content, there is a wide spectrum of formal and informal sources available with them but could be converted into digital form by devising suitable action plan. Image format, compression schemes, network transmission, monitor and printer design, and image-processing capabilities are all likely to improve dramatically over the next decade. But technology alone will not determine the future; relationship, economic and pattern of behavior are equally important.

5. References

1. Shemeent (Jorge Reina). Encyclopedia of Communication & Information. Vol.2;2002.p.449-553.
2. Borgman (Christine L.). What are Digital Libraries? Computer Vision. Information Processing. Vol.35;1996;p.227-43.
3. O'Donnel (J.J.). The Virtual Library: An idea whose time passed. Philadelphia, University of Pennsylvania,1995,p.1.
4. Arora,(Jagdish).Building digital libraries:An overview.DESIDOC Bulletin of information technology.Vol.21 (6);2001;p3-24
5. Marchiomini (G).information Seeking in electronic environments.New York,Cambridge Press,1995.
6. Richvalsky and Walkins.Designing and implementing a digital library.ACM Crossroad student magazine,Jan 1998.http:// www.acm.org/crossroads.rds5.2/digital.html.
7. Wiederhold (G).Mediation in the architecture of future information systems.IEEE Computer,38-49 (March 1992).

About Authors

Mr. Mohd. Nazim is doing research in Department of Library & Information Science, Aligarh Muslim University, Aligarh, Uttar Pradesh. He is Junior Research Fellow of UGC.
Email : monis_naz@yahoo.co.in

Faizul Nisha is doing research in Department of Library & Information Science, Aligarh Muslim University, Aligarh, Uttar Pradesh.
Email : momi_sonu@yahoo.co.in

Knowledge Management in Bangladeshi Libraries: A Long Way to Go

Kazi Mostak Gausul Hoq M Nasiruddin Munshi

Abstract

Describes that both knowledge and information have become essential ingredients to change our society for future vision and shows the different approaches of knowledge management activities. Mentions the modules and processes of knowledge management and also discusses the knowledge management techniques in libraries. Shows the present status of knowledge management activities in Bangladesh libraries.

Keywords : Knowledge Management, Bangladesh

0. Introduction

In the present age of information technology, both information and knowledge have become essential ingredients due to multidimensional use and application in the society. They have also been playing an important role to change and improve the current society for future vision. Organizing information for its gainful use in social development interventions against the environment of information explosion all over the world has become a highly contentious issue that poses a great challenge for today's librarians and information professionals. The developed countries of the world have already realized the importance of knowledge and information and accordingly collected and organized them properly. No doubt, knowledge and information cannot be well managed until some organizations or professionals take the clear responsibility of it. Of course, library professionals are the right persons to shoulder this responsibility as a whole.

In the current atmosphere of rapid change (whether political, economical, social or technological), information has become such a key asset that teachers, students, decision makers, top executives, development activists, people form every cross-sections of the society need to be informed (Mahapatra, 1999: 7). To meet this diversified need of information, the traditional role of the library is making way for information centre that is actively involved in providing all types of information that may be in actual or potential demand.

This paper is an attempt to illustrate the real picture of present information and knowledge management situation by the libraries and information centers of Bangladesh, an emerging concept of knowledge management, modules, processes and techniques of knowledge management, various findings and important directions for further development.

1. Knowledge Management : An Emerging Concept

Knowledge Management is defined as the process of creation, capture, organizing, accessing and using knowledge to create customer's/user's value. It is also defined as the management of corporate knowledge that can improve a range of organizational performance characteristics by enabling an expertise to be more intelligent acting. Knowledge management is a revolutionary method by which, the detached knowledge available from diversified sources are captured, assimilated and converted in to a powerful competitive intelligence through various electronic devices such as CD-ROM, floppy, hard disks, software etc.(Bhunia, 2000: 59). In recent years, the term 'knowledge management' has gained worldwide prominence. For many in the academic world, and in the profession of librarianship, this is nothing new.

Librarians and information professionals consider knowledge management to be a function historically performed by librarians (Hawkins, 2000). However, in this age of digital information, the role and nature of knowledge management has diversified greatly. It has assumed new roles and responsibilities and over the years, has been accommodating many areas other than librarianship. Nowadays, it is increasingly considered to be a cross-disciplinary domain that draws ideas from a wide range of disciplines and technologies like cognitive science, library and information science, technical writing, decision support systems, computer supported collaborative work, etc.

Attempts have been made by different authors, specialists, organizations and groups to identify and define knowledge management in different ways. Gary A. LaBranche has defined it as "the process of transforming information and intellectual assets into enduring value. It connects people with the knowledge that they need to take action, when they need it. In the corporate sector, managing knowledge is considered key to achieving breakthrough for competitive advantages" (LaBranche, 2000). Skryme defines it as "the explicit and systematic management of vital knowledge and its associated process of creating, gathering, organizing, diffusion, use and exploitation. It requires turning personal knowledge into corporate knowledge that can be widely shared throughout an organization and appropriately applied" (Skryme, 1997).

While trying to define knowledge management Karl, E. Sveiby followed a different approach. He identified two tracks of activities on the subject and two different levels. (Sveiby, 2001). IT Track Knowledge Management – Management of Information. Researchers and practitioners in this field have their education in computer and/or information science. They are involved in construction of information management systems, artificial intelligence, groupware etc. To them Knowledge – Objects that can be identified and handled in information systems. People's Track Knowledge Management – Management of People. Researchers and practitioners in this field have their education in philosophy, psychology, sociology or business/management. They are primarily involved in assessing, changing and improving individual skills and/or behavior. To them Knowledge – Processes, a complex set of dynamic skills, know-how etc., that is constantly changing. Level: Individual Perspective. The focus in research and practice is on the individual. Level: Organizational Perspective. The focus in research and practice is on the organization.

Summarizing the above definitions, it may be mentioned that, Knowledge Management is a complex process, which aims to enhance the use of organizational knowledge through sound practices of information management and organizational learning. In other words, knowledge management is an allocation of knowledge including information through the application of intranet to different annexes of an organization for effective role of the executives. If we are to gainfully use knowledge management for fulfilling our organizational objectives, we must have a clear understanding of the nature and scope of information flows inside and outside the organization.

2. Modules of Knowledge Management

The radical changes during 1990's with the rapid development of software technologies in the IT sector included text, image, audio, graphics, hypertext etc. which made an optimistic contribution to knowledge management. The Internet has been playing a dynamic role in activating the technology in building the true image of knowledge management in shape of electronic storage of information, retrieval, document delivery, accessing of information through various databases, transfer of file, transfer of information etc. The knowledge management system is primarily based on eight vital modules such as, (i) Information; (ii) Expertise; (iii) Collaboration; (iv) Team; (v) Learning; (vi) Intelligence; (vii) Knowledge transfer, and (viii) Knowledge Mapping (Balaje, 2000: 15). These are:

Information : the most important bezel acts as an instant access to update and customize information;

Expertise : connects in real-time experts in an organization to members who desire assistance and even the implied knowledge can be made explicit;

Collaboration : plays an important role to facilitate online brain storming sessions and collects & preserves information;

Team : ensures efficient and systematic management among teams and shares skills;

Learning : abridges skill gap with the help of online sessions;

Intelligence : deals mainly with the explicit knowledge base;

Knowledge transfer in a structured electronic form : according to William Saffady pertains to (a) machine readable data files; (b) various online databases and CD-ROM information products; (c) computer storage devices in which information reside in shape of optical disk, juke boxes or magnetic tap autoloaders; and (d) computerized Networking Systems (Matson, 1997: 88); and

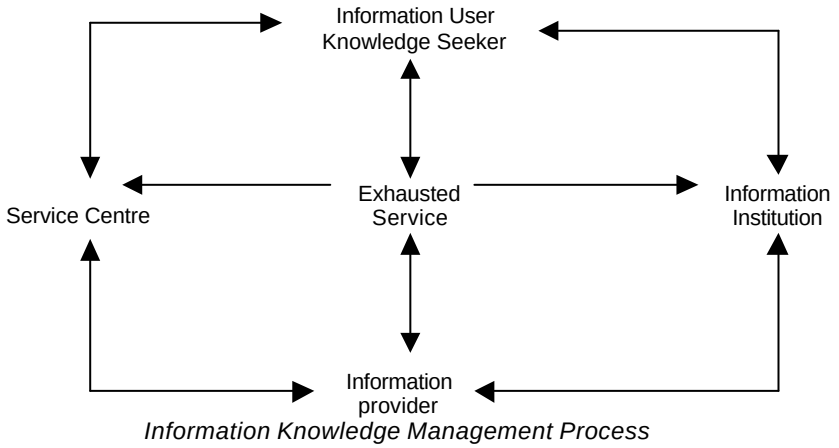
Knowledge mapping : identifies the body of knowledge within the organizations, which is primarily concerned with mute knowledge base and makes a repository of all skills and expertise in the organization.

3. Knowledge Management Process

Knowledge management is a process that helps to find, acquire, select, organize, retrieve, disseminate and transfer important information and expertise necessary for various activities such as decision-making, problem solving, planning, implementing etc. However these activities should be undertaken carefully so as right knowledge is captured & disseminated for the right users at the right time to take right decision at the right situation. Knowledge management consists of following steps (Kherde, 2004: 155):

- Create a new knowledge from all fields;
- Identify the individuals & capture the knowledge what they possess & convert it into explicit form;
- Organize the useful knowledge scientifically so that it can be easily retrieved; and
- Identify the ways by which the organized knowledge can be disseminated to the proper requests.

Knowledge management is one of the ways for the development of the nation as a whole. It is the responsibility of society to make itself aware of it. Society is made up of different organizations such as social, governmental, economical, educational, cultural, industrial, etc. All such organizations should perform the task of knowledge management and for creation of new knowledge the organization should identify the expertise working with that organization. To motivate this expertise, the activities such as seminars, workshops, symposiums, conferences etc. must be organized. On these platforms the individuals share their views, ideas and experiences. This sharing may also be useful for the better performance of their day-to-day working. From the platform of such intellectual activities, new implicit knowledge creates. The organization can convert the individual assets of knowledge is to be converted into the explicit one to transmit it to the knowledge seekers. Knowledge management process of an information institution is shown by the following model:



Information Knowledge Management Process

Information or knowledge should be organized properly so that it can be transmitted whenever demands come from the requesters/knowledge seekers. Disseminator or information provider should be able to transmit this knowledge to the right request at the right time in the right way. So far, it is highly difficult task to perform if it is not organized properly or scientifically.

4. Knowledge Management Techniques in Libraries

It has already been observed that knowledge management is something that the library and information professionals claim to be practising for a long time. But for the last few decades, phenomenal advancements in the world of information communication technologies and both quantitative and qualitative changes in the information needs of individuals and corporate organizations have encouraged the library and information scientists to assume new roles and reshape their information activities. So, that the traditional roles of librarians and the term 'librarian' are at risk of losing weight and importance. That is why an increasing number of librarians all around the world are learning new techniques (especially ICT) to cope with the challenges of modern society and keep their jobs relevant and meaningful (Sherwell, 1997: 35-36). A recent Information Service Panel Survey conducted in the USA among special librarians shows that, the librarians think that within five years they will be assuming new roles which will be altogether different from their present role. More and more librarians think in the coming days they will be acting as 'Corporate Knowledge Managers' rather than 'librarians'. Besides, increased recognition of knowledge as a valuable strategic resource would heighten the importance of information professionals (Deieckmann, 1997: 19). As a result, the librarians will be responsible for amalgamating and coordinating all information activities, e.g. library database management, competitive intelligence, marketing research, internal knowledge sharing, etc. Being the knowledge manager, the librarian will have to create value to his/her organization by facilitating access to high quality information and by networking people and their ideas together using technological and technical infrastructure. This means that, in order to become knowledge managers, librarians will have to master new techniques and beneficially use them for excelling their performance and increasing the importance of their library to the information clients.

The key to knowledge management is capturing the knowledge of technique – how organizations get their work done and how various elements of information connect to this. Here, there are two different types of knowledge, e.g. explicit and tacit. Explicit knowledge is packaged, easily codified, transferable and communicable. Tacit knowledge is personal, context specific, difficult to formalize, and difficult to communicate and transfer (Hawkins, 2000). Attempts should be made by the library and information resource centre to try and capture both tacit and explicit knowledge and after structuring and modeling this knowledge in suitable forms, making these knowledge accessible to the users through a single access point so that a comprehensive information retrieval system can be put in place with the help of electronic search engines. To make this possible, the power of ICT must be properly utilized. The digital technology is at the heart of the emerging information and knowledge economy. It is frequently known as the most important means for making knowledge management effective and meaningful.

The success of the library knowledge base depends on two things: content and access (Broadbent, 1998). The library knowledge repository should be able to ensure that the answers to the users' queries are in the repository and at the same time it should be easy to find. If these two things can be ensured, the library knowledge management system will be able to function effectively. No doubt, knowledge management is not the exclusive domain of any particular group or profession. But if librarians and information specialists want to be key players in the emerging knowledge management phenomenon, they need to understand the multiple perspectives of the other professionals. Only then they can understand their own role in knowledge management intervention and will be able to get proper indication and insight as to what their future roles and responsibilities would be. Here it is important to understand that knowledge management is not only about managing or organizing books and journals, searching the Internet for clients or arranging for the circulation of materials, but each of these activities can in some

way be a part of knowledge management spectrum and process (Broadbent, 1998). That simply means that as knowledge workers librarians' responsibilities are multi-dimensional and demand a broader understanding of communicational, technological and other competence along with traditional library and informational skills and expertise.

While designing the knowledge management system in libraries, the following factors must be considered (USSLA, 1996):

- There should be enough content to make the knowledgebase useful. People use a knowledgebase because they hope to get what they are looking for in it.
- It should be ensured that the knowledgebase will grow in time. This also corresponds with one of the laws of library science: library is a growing organism.
- The knowledgebase must be easily and effectively accessible. For this, multiple tables of contents, hot links, and a good search engine must be in place.
- The users must be allowed to access and use the knowledgebase in their own suitable ways.
- While designing the knowledgebase, common, well-understood and mainstream tools and techniques should be used, which stand the best chance of being around for a long time.

5. Knowledge Management in Bangladeshi Libraries

Knowledge management has a number of obstructions in its way. Institutional, infrastructural, organizational and psychological obstructions are posing grave challenges to the successful implementation of knowledge management system in libraries. Most of the library users and patrons are still not well aware of the potential and far reaching impact of knowledge management and hence, are yet to contribute as much as they should for making this a meaningful venture. Nevertheless, efforts are underway in developed countries to strengthen knowledge management initiatives in libraries and give this venture a formal and more institutional shape.

But quite naturally, as a developing country Bangladesh is yet to fully comprehend the notion of knowledge management, let alone be benefited from such an endeavor. Bangladeshi libraries and information centres lack adequate manpower, infrastructure, information resources, financial support, patronization from government and non-government organizations and an educated user base who would play their due roles in making libraries a centre of knowledge management initiatives. Besides, information or knowledge are yet to be considered as key development resources or commodity in Bangladesh, people and the policy makers alike are not fully woken up to the fact that if utilized effectively, information also can act as a strong economic resources like natural gas or oil.

Library and information professionals of Bangladesh have still a long way to go to better manage their resources with the help of information communication technologies for maximizing the impact and effectiveness of their library resources. Meanwhile, the patterns are shifting rapidly in every aspect of their job. The new media – audio-visuals, television, microforms and computer base communication are in competition with book. Information is being generated faster than libraries are able to organize and store it. Commercial organizations and private companies are getting into the information business. Databases are replacing catalogs (Mahapatra, 1999: 07). All this presents a depressing scenario for the libraries and librarians of Bangladesh. Under the circumstances, if the library and information professionals are to keep themselves in the broad picture with their traditional importance and relevance, they must make their presence felt in every stratum of the society and to the forces that shape and reshape the process of social advancement. Knowledge management can also give the information professionals their expected control in this quest. It holds great potential for libraries of the country like Bangladesh because

it can help library and information professionals in improving their status and turning them into a driving force of the new information age who must be taken into account for sustainable development of the society.

6. Directions

Knowledge management holds great promise for the libraries of Bangladesh for the following reasons:

1. Against the backdrop of changing socio-economical scenario, information and knowledge are increasingly considered to be key economic resources. A rich knowledge repository in libraries in every part of the country can act as an invaluable tool for the businessman, researchers, government officials, NGO personnel and community people in general for their socio-economic interventions.
2. A scientific and systematic knowledge management infrastructure will help libraries to achieve a higher status and greater importance in the society. People will cease to think libraries as unimportant or insignificant. This will help libraries to be considered as an indispensable social institution.
3. It will help libraries in assuming diversified roles and responsibilities and would give the information professional a huge influence in discharging their duties. Supported by this new power, libraries would be able to act as a platform for newer interventions like distance learning, electronic commerce, etc.
4. This would bring qualitative and quantitative differences in the library and information services. It would improve reader service by streamlining response time and providing much more relevant information from a single point.
5. Knowledge management would restructure library operations and reduce costs by eliminating excess or unnecessary processes. It would improve the performance of the employees by recognizing the value of employees' knowledge and rewarding them for it.

But all these are not easy to achieve. Keeping in mind the depressing situation of the libraries in Bangladesh, their insufficient resources, weak ICT infrastructure of the country and above all people's unawareness of the importance of information, it will be a Herculean task to put a workable knowledge management system in place in Bangladeshi libraries. It would require heroic efforts from the library and information professionals, active patronization of the government, and spontaneous participation and involvement of library users and patrons in the knowledge management process to make this a worthwhile venture. Librarians and information professionals will also have to involve other professionals especially computer scientists and programmers in this interventions and make a combined and collective effort to design and implement a successful knowledgebase and knowledge management system.

7. Conclusion

Knowledge management undoubtedly holds great potential not only for the libraries and information institutions, but also for the government agencies and corporate world. But to translate these possibilities into reality, the notion of knowledge management must be clearly understood and its real value should be evaluated. The value of knowledge management relates directly to the effectiveness with which the managed knowledge enables the members of the organization to deal with today's situations and effectively envision and create their future. This poses a considerable challenge to the persons who would design and implement knowledge management systems. The challenge is even greater for the library and information professionals of Bangladesh. The future of knowledge management practices in Bangladeshi libraries depends almost entirely on the kind of initiatives taken by Bangladeshi library professionals and their level of sincerity, skill and vigor.

8. References

1. Balaji, S. How Much does Your Company Know? Express Computer. 11(6), p 15.
2. Bhunia, C. T. The Future of Knowledge Management. Electronics for You. Nov. 2000, 59-60.
3. Broadbent, Marianne, The phenomenon of knowledge management: what does it mean to the information profession? Information Outlook, 2 (5) 1998.
4. Carillo, Javier, Managing knowledge based value systems, Journal of Knowledge Management, Vol. 1, No. 4, June 1998.
5. Corrall, Sheila, Knowledge Management: are we in the knowledge management business? Ariadne: Issue 18, December 1998, Web: http://www.ariadne.ac.uk/ issue18/Knowledge-mgt/
6. Cronin, Blaise, and Elisabeth Davenport, Knowledge Management in Higher Education, Information Alchemy: The Art and Science of Knowledge Management. EDUCAUSE Leadership Strategies, No. 3, Jossey-Bass, San Francisco, 2000.
7. Dieckmann, Heike, Information for competitive advantage, Managing Information, Vol. 4, No. 7, September 1997.http://conferences.alia.org.au /alia2000/proceedings/brian.hawkins.html
8. Knowledge Management: a library perspective, Presentation at the US Special Libraries Association (USSLA) Louisiana/Southern Mississippi Chapter Fall Program, October 12, 1996.
9. Kherde, M. R. Knowledge Management: A Need of the Day. In: Information Management Trends and Issues (Festschrift in Honour of Prof. S. Seetharama). New Delhi: 2004.
10. LaBranche, Gary A. Knowledge Management: The Killer App for the 21st Century. American Society of Association Executives. 2000, http://www.asaenet.org/sections/ membership/article/ 1,2261,50864,00.html
11. Mahapatra, P.K. and Chakrabarti, Bhubaneswar, Organising information in libraries, Vol. I., New Delhi, Ess Ess Publications, 1999.
12. Matson, L. D. Do Digital Library Needs Librarians. On-line. Nov.- Dec., 1997, 88.
13. Plater, William M. The Labyrinth of the Wide World, Educom Review, v. 30, no. 2. March/April 1995
14. Sherwell, John, Building the virtual library: the case of Smith Cline Beecham, Managing Information, Vol. 4, No. 5, June 1997.
15. Skryme, D. Knowledge Management: making sense of oxymoron, Management Insight, 2 nd series No. 2, 1997. (http://www.skyrme.com/insights/22km.htm).
16. Sveiby, Karl-Erik, What is Knowledge Management? 2001. http://www.sveiby.com/ articles/ KnowledgeManagement.htm/.

About Authors

Mr. Kazi Mostak Gausul Hoq is a lecturer in the department of Information Science & Library Management, University of Dhaka, Dhaka, Bangladesh. His teaching and research interests are information seeking behaviour, information and knowledge management systems, information technology, rural information systems and services, etc.

Dr. M. Nasiruddin Munshi is an Associate Professor in the Department of Information Science & Library Management, University of Dhaka, Dhaka, Bangladesh. His teaching and research interests are marketing of information products and services, MIS, information seeking behaviour, information and knowledge management systems, information technology, organization of knowledge, etc. He has published 20 articles in different reputed research journals from home and abroad.

A Brief Evaluation of Search Facilities and Search Results of Few Resources Accessible through INDEST Consortium

Kshyanaprava Sahoo V K J Jeevan

Abstract

The Indian National Digital Library in Science and Technology consortium, setup by the Ministry of Human Resource Development, Government of India has currently over 140 institutions as members who are taking advantage of cost effective access of premier resources in science, technology and management. Out of the different resources accessible under this consortium, the present study selected four major resources, such as ACM Digital Library, IEEE/IEE Electronic Library, ScienceDirect of Elsevier, and Springer Link to make a comparative assessment of the key features and quantity of records in these. The results identified are presented in tabular form. More exhaustive studies are further planned to categorically identify the best resource to answer a crucial query and in identifying the inherent benefits and limitations of each resource.

Keywords : INDEST, E-Resources, Evaluation of E-Resources

0. Introduction

The Indian National Digital Library in Science and Technology (INDEST) consortium was setup by the Ministry of Human Resource Development (MHRD) of the Government of India, to subscribe full-text electronic resources and bibliographic databases for 38 leading engineering and technological institutions in the country including IITs (7), IISc (1), NITs / RECs (17), IIMs (6) and a few other institutions directly funded by the Ministry of Human Resource Development (MHRD) [Arora and Agrawal]. As a result of this initiative, the full text e-journals in IIT, Kharagpur have increased more than 8 fold from 600 journals in 2000 to 5381 journals in 2003. Apart from the 38 core members, now the consortium has 66 Government engineering colleges or technical institutions that offer programmes at postgraduate level and 46 other engineering colleges and institutions (AICTE-accredited and UGC-affiliated) who joined the consortium on their own to share the benefits it offers in terms of lower subscription rates and better terms of agreement with the publishers [INDEST-Members]. Evaluating electronic resources is a major area of study in the information science. Basing three such studies [Lancaster, Nicholls and Ridley, and Lord and Ragan], we tried to evaluate the search facilities in four major resources accessed through the INDEST Consortium. We have also attempted to present a comparison of the search results obtained when searched for same terms. This paper is organized in four parts. In the second part, a brief of the INDEST consortium is presented with a list of full text and bibliographic information resources with their web site addresses. This study evaluates four premier resources, viz., ACM Digital Library, IEEE/IEE Electronic Library, ScienceDirect of Elsevier, and Springer Link, accessible through the INDEST Consortium in the third part of the paper with a critical evaluation of the search features and search results.

1. INDEST Consortium

The Indian National Digital Library in Science and Technology (INDEST) consortium has successfully demonstrated the utility of electronic information and fruitful access to premier science and technology institutions in the Country for the last two to three years of its operation in premier libraries affiliated to seven IITs, IISc, six IIMs and seventeen NITs / RECs. All electronic resources subscribed are available from the publisher's Web site. Local hosting of resources has not been considered at this stage. The INDEST consortium subscribes to the following resources for various categories of institutions as in Table 1 [INDEST-Resources]:

Table 1: Full Text and Bibliographic Resources accessible through INDEST

Full Text Resources		
No.	**Name**	**Web Site**
1	IEEE/IEE Electronic Library Online (IEL)	http://ieeexplore.ieee.org/
2	Elsevier's Science Direct	http://www.sciencedirect.com/
3	Springer Verlag's Link	http://www.springerlink.com/
4	ProQuest's ABI/ Inform Complete	http://www.il.proquest.com/pqdauto
5	ProQuest Science (formerly ASTP)]	http://www.il.proquest.com/pqdauto
	[Formerly Applied Science and	
	Technology (ASTP) Online	
6	Association for Computing Machinery	http://portal.acm.org/portal.cfm
	(ACM) Digital Library	
7	American Society of Mechanical	http://www.asme.org/pubs/journals/
	Engineers (ASME) Journals	
8	American Society of Civil Engineers	http://www.pubs.asce.org/journals/jrns.html
	(ASCE) Journals	
9	EBSCO Databases	http://search.epnet.com/
10	Emerald Full-text	http://iris.emeraldinsight.com/
11	Nature Journal	http://www.nature.com/
12	Capitaline	http://www.capitaline.com/intranet/ INDEST_consortium.htm
13	INSIGHT	http://www.insight.asiancerc.com/
14	Euromonitor (GMID)	http://www.euromonitor.com/gmid
15	CRIS INFAC Industry Information	http://www.crisil.com/
16	ASTM Standards	Intranet Version
17	Indian Standards	Intranet Version
	Bibliographic Databases	
No.	Name	Web Site
1	COMPENDEX on EI Village	http://www.ei.org/ev2/home
2	INSPEC on EI Village	http://www.ei.org/ev2/home
3	SciFinder Scholar	http://www.cas.org/SCIFINDER/SCHOLAR/ index.html
4	Web of Science	http://isiknowledge.com
5	J-Gate Custom Content for Consortia (JCCC)	http://jccc-indest.informindia.co.in/
6	J-Gate for Engineering and Technology (JET)	http://jet.informindia.co.in/

2. Resources Studied

We found that four resources, ACM Digital Library, IEEE/IEE Electronic Library, Elesevier's ScienceDirect, and Springer-Verlag's Link are used more heavily than others and hence decided to study the features of them. ACM publications such as journals, magazines, transactions, special interest group (SIG) newsletters, proceedings, and publications by affiliated organizations are the premier source of information for researchers in the fields of computer science and information technology. Through its technical publishing, conferences and consensus-based standards activities, the IEEE produces 30 percent of the world's published literature in electrical engineering, computers and control technology

and holds annually more than 300 major conferences and has nearly 900 active standards with 700 under development. Since its launch in 1997, Science Direct has evolved from a web database of Elsevier journals to one of the world's largest providers of scientific, technical and medical (STM) literature covering over 1,800 journals having around 6 million articles and over 60 million abstracts from all fields of science. With a collection of journals and book series that account for over 300,000 documents in Springer Link, the browse and explore functions help users to get quickly to the information and titles they need. With Springer Keyword alerts, users can register a keyword and each time the keyword appears in a publication on Springer Link, the user receives an e-mail notification. A comparison of major features of these resources is highlighted in Table 2, search fields supported in Table 3, search facilities available in Table 4, and a comparison of search results is presented in Table 5.

Table 2: Features of ACM Digital Library, IEEE Electronic Library, Elsevier's ScienceDirect, and SpringerLink

Parameters	**Science Direct**	**Springer Link**	**IEEE**	**ACM Digital Library**
Coverage	Covers 24 subjects. Includes over 1800 Journals online. About 6 million articles and over 60 million abstracts from all fields of science are available	Currently offers over 500 fully peer-reviewed journals and a growing roster of series, comprising more than 2400 books online	IEEE provides full-text access to IEEE transactions, journals magazines and conference proceedings published since 1988 plus select content back to 1950, and all current IEEE Standards. Contains more than 770,000 articles in over 12,000 individual publications. IEEE adds about 25,000 new pages to the database per month.	Advancing the arts, sciences, and . applications of information technology
Accessibility	Access through Web	Accessed directly or through links provided by Abstracting and Indexingservices, agencies or through CrossRef. Free access is provided to search functions tables of contents, as well as keyword and tables of contents alerts	Access to all or part of the collections based on whether you are a member or whether your organization subscribes to all or part of the collections	Members can access through Subscription, license, or transaction fee
Authority	Reputable. Sources are provided with	Offers electronic & printed literature from Springer-	IEEE publications provide quality, depth, and valueto	Reputed. Suitable references are

	references	Verlag, a preeminent scien-tific publisher with A reputation for excellence spanning more than 150 years.	electrical engineering computing collection	provided with the references.
Searching	An easy-to-use, powerful search engine with both basic and enhanced search capability	Search engine is supported.	Easy search option helps to get required data	Search engine helps to identify useful records.
Browsability	Browsing facility is there for Abstract databases, Book series,Journal homepages, Reference works, etc.	Browse and explore functions help users to get quickly to the information and titles the users need	Browse functions are available such as searching the database by specifying one or more authors, index terms, and other criteria	Browse through Keyword Index.
Archiving	Currently, Science Direct contains over 1,800 Elsevier journals, most from 1996 forward and some from 1993 forward. Back file conversion from volume 1, issue 1 ongoing.	Scientists and researchers can access over a century of scientific evolution and complete historical information.	Access to IEEE journals, magazines & conference papers back to 1988, select titles to 1950.	50 years of archives
Organization	Resources are organized in a logical manner	Organized resources	Resources are organised	Arranged logically
Currency	Daily updated	As soon as the new information comes	Depending on the type of document, currency of the content varies.	Weekly
Documenta-tion	User can print or download Content from the site	Subscribers to a journal title receive online access to that title whether by a print plus online or an electronic-only subscription	Provision to view and print individual articles and papers, search results lists, tables of contents, bibliographic records, and full-page images with no limit on the number of prints.	Users can make digital or hard copies of the individual articles as long as they bear the ACM copyright notice
Ease of use	Resources are novice friendly	Has a User friendly interface	Very much easy to use	Training is not required to use

Links to other resources	The site contains Hyperlinks to other sites or resources.	Users can expand their research with the reference linking found within articles in SpringerLink.	This Site contains links to other sites. There is a facility for OPAC linking. Links may be created at the title level (for journals/ magazines, conference proceedings and standards) or at an issue's table of contents level (for journals/ magazines only.)	Users can assemble and distribute links that point to works in the ACM Digital Library.
Required environment or platform computing	Users should connection to the Internet. In order to possess a good access certain content and to make use of the full functionality and advanced personalization features of the site password is necessary	Required Internet connection only. To use the person-alized features of this site, such as 'Favorites' or 'Table Of Contents Alerting', registration is required.	-Internet browser - Connection to an Internet Service Provider - For best dial-up performance a 56.6 or higher modem is recommended - Adobe Acrobat Reader 5.X or higher. Direct parallel or LAN-attached printer with at least 300 dpi resolution, Compatible mouse	With Internet connection one can access. To access the Portal/ Digital Library, members must have an ACM Web account. To get the Personalized services offered by ACM, Membership login is required
Stability Uniqueness	Resources are very much Stable Articles are available online before appearing in print. Online access to multimedia features not available in print journals, such as: video files, audio files, Excel spreadsheets and Word files	Stability is there Interdisciplinary research is a key feature of Springer Link. Subscribers have a vast universe of information at their disposal with 11 online libraries that enable them to obtain critical data in many fields.	Stable resource Cost effective and premier resource base for electrical, electronics and computer engineering	Yes Users can make digital or hard copies of the individual articles that they are entitled to access for personal or classroom use, as long as the copies are not made or distributed for profit or commercial advantage.
Networkable	Yes	Yes	Yes	Yes
Indicativity of record Response time	Default is 200 items per page Very quick	Default is 10 items per page Fast, reliable, and powerful access 24 hours a day.	Displays 25 search results per page Quick response	Shows 20 items per page Fast access

Help features	Online help is available for various tasks.	Help feature is very much useful	Help function is very satisfactory	Help option is available
Value added	Resources contain supplements to the textual materials such as graphics, search engines etc.	Electronic suppl-ementary materials such as color images, simula-tions, video and sound, so that researchers not only read about the research in the article but can see, and often hear the research as it happens	Contains complete original page images, including all charts, graphs, diagrams, photographs, and illustrative material, from an integrated-circuit schematic to a topographic map to a photograph of a new crystalline structure.	Links can be created to citations. ACM encourages the widespread distribution of links to the definitive versions of its copyrighted works.
Indexing & Vocabulary factors	- Abstracts, titles, Keywords, and authors within the selected contentby subject, publishers search based on Boolean logic, title search etc.	Access points such as –Author, Title, Publishers, Keyword	- By Author- Basic Search- Advanced Search- Cross Reference	Searching with words, phrases, Topics, Wildcards, cross reference

Table 3: Search Fields in ACM Digital Library, IEEE Electronic Library, Elsevier's ScienceDirect, and SpringerLink

Fields	**Science Direct**	**Springer Link**	**ACM Digital Library**	**IEEE**
Author	Author, title, keywords	Author	Author, Editor	Author
Title	Title	Title	Title	Title
Year	Publication date	Year	Publication date	Year
Source	All journal, my favourite journal, subscribed journals			
Abstract	Abstract	Abstract	Abstract	Abstract
Language	Limit to English language documents			
References	References			
Full Text	Full Text			
Descriptor	Keywords	Descriptor		Index terms
Author affiliation	Affiliation		Affiliation	Affiliation
Document type	Document type (used within limit field)		Publication type	Publication type
ISSN	ISSN	ISSN	ISBN/ISSN	

Table 4: Search Facilities in ACM Digital Library, IEEE Electronic Library, Elsevier's ScienceDirect, and SpringerLink

Description	**ScienceDirect**	**Springer Link**	**ACM**	**IEEE**
Operators: Boolean	AND, OR, AND NOT	AND, OR, AND NOT	AND, NOT	AND, OR, NOT, ADJ
Sort results by			Relevance, Title, Publication, Publication date	
Display results			Expanded form, Condensed form	
Wildcards (search terms marked inside?	*, !	*	*	*, ?
Operators: proximity			-, +, ""	-,
Searching	Go	Go	Search	Find
Complex searching	Search terms can be linked with operators	Search terms can be linked with operators	Search terms can be linked with double quotation marks	Search terms can be linked with operators and double quotation marks
Saving search strategy and search results	Save	Save	Download	Save
Printing	Printing	Printing	Printing	Printing
Hit-term searching	No specific name while displaying record	No specific name		
Display	Display	Results	Display	Results
Search terms entry	Search terms can be typed directly or can be browsed or selected from Index, glossary	Typed directly	Typed directly	Typed directly
Truncation of word roots	!			
Help facility	Is available	Not available	Is available	Is available

Table 5: Search Results from different search engines

Search Terms	ScienceDirect	IEEE	Springer Link	ACM Digital Library
Digital resource management	90	75	34	0
Spread spectrum communication	66	126	18	1
Information Retrieval	3463	1036	1000	4322
Error control coding	96	198	69	12
Wireless Network	678	1420	900	1669
Wireless Communication	495	640	430	207
IIT, Kharagpur	10	96	0	67
Soumitro Banerjee	2	4	0	0

3. Conclusion

This study has to be extended to include more parameters identified for evaluating information resources. Also search features are to be critically examined with user interfaces in each of the resources to judge which resources are more user friendly and have better search functionalities. In the era of more and more interdisciplinary research taking place in many institutions, identifying best resource to answer a particular query needs searching different resources with same terms and examining the results obtained. Resources selected for access through INDEST was selected by an expert committee drawn from the premier institutions and they have carefully screened the different resources available and selected the best possible ones available. This fact is evident from the standard, quality and usage of these resources. More involved studies on the subject matter covered in the different resources and if possible a term mapping would help to identify what resources are of interest to which specialists. Researchers too involved in their research and academic work can be helped by libraries by venturing into such value added information support.

4. References

1. Arora, Jagdish and Agrawal, Pawan, Indian Digital Library in Engineering Science and Technology (INDEST) Consortium: Consortia-Based Subscription to Electronic Resources for Technical Education System in India: A Government of India Initiative, International CALIBER 2003, Ahmedabad, 13-15 February 2003, Conference Volume - Mapping Technology on Libraries and People, Ahmedabad: INFLIBNET, 2003, pp. 271-290.
2. INDEST-Members, http://paniit.iitd.ac.in/indest/members.html
3. INDEST-Resources, http://paniit.iitd.ac.in/indest/eresources.html

4. Lancaster, F. W., The Evaluation of Machine Readable Databases and of Information Services derived from these Databases, in Lancaster, F W, and Cleverdon, ed., Evaluation of Scientific Management of Libraries and Information Centres, Netherlands: Noordhoff, 1977, pp. 73-100.

5. Lord, Jonathan and Ragan, Bart, Working together to develop Electronic Collections, Computers in Libraries, 21 (5), May 2001, pp. 40-44.

6. Nicholls, Paul and Ridley, Jacquline, A Context for Evaluating Multimedia, Computers in Libraries, 16 (4), April 1996, pp. 34-39.

About Authors

Ms. Kshyanaprava Sahoo is Professional Trainee in Central Library, Indian Institute of Technology, Khadagpur, West Bengal..
Email : sony_prabha@rediffmail.com

Mr. V K J Jeevan is presently working as a Assistant Librarian at Indian Institute of Technology, Kharagpur, West Bengal. He has presented number of papers in seminar, conferences and journals. He is also a member of many professional bodies.
Email : vkj@library.iitkgp.ernet.in

The Needs for Content Management with Special Reference to Manuscripts of Manipur

T Satyabati Devi T A V Murthy

Abstract

Manuscripts are one of the precious materials of our cultural heritage. They are valuable sources for the reconstruction of the history and culture of a country. They reveal their contemporary society and provide vital link to culture and knowledge. In order to manage and preserve our cultural heritage for use now and for future generation it is necessary to create the context in which cultural heritage agencies and organizations can pursue the rising standard of stewardship .A great deal of work has to be done to improve the level and profile. Ensuring that the public gets access to these information resources has become one of the priorities for those providing services. This paper highlights the importance of manuscripts collection and the necessity to preserve them for the future research and references. It also gives a bird's eye view on the existing conditions of the Manuscripts collection of Manipur. The central theme of the paper discusses the need for content management of these Manuscripts.

Keyword : Manuscript, Content Management

0. Introduction

Manuscripts are invaluable source for the reconstruction of the history and culture of the country. These reveal their contemporary society and provide vital link to culture and knowledge. It was during the reign of Khagemba that manuscripts were taken seriously. Up to the close of the 17th century before the advent of Hinduism, a good number of manuscripts were written on different topics like chronicle, mythology, administration, astrology, pure literature dealing with romance and heroism. Unfortunately, this work does not carry the names of writers and dates of composition. The manuscripts are mostly written in Meitei script. In Manipur, a large number of manuscripts are still lying scattered in every known and unknown place in organized/Unorganized conditions. Many eminent scholars have collected the Meetei scripture of the early medieval period which are the Meetei counterparts of the epics, literary heritage and similar early written evidences of human civilization and they have even transcripted in Bengali script. The variety and richness of the historical literature is a striking feature of early Manipuri literature.

1. Manuscripts of Manipur

The manuscripts of Manipur are found mostly in the private and public custody. The custodians keep the manuscripts as a sacred entity with proper care. They are not allowed to be used anytime we want. They did their own processing to differentiate the subjects with which they deal. Translation and transcription are done by some of the eminent scholars and published already and there are many more not yet published. Though the state archives and some museums collect the manuscripts they cannot estimate the number of manuscripts lying scattered in every known and unknown places. The custodians played an important role in safely keeping these invaluable manuscripts saving our cultural heritage. But there is scare that these manuscripts are slowly decaying and vanishing day by day and there is a need to take remedial measures for their preservation before they become totally useless. Preservation of these precious gifts presents a great challenge to us but still the IT offers the best solution for preservation and enhancement of them for wide access i.e. through digitization. The proliferation of development in digital

technology makes choosing the right method of digitizing collection an increasingly complex process for information organization. As the information age is creating a digital dilemma, the content management through digitization should be the best solution.

2. Detailed Account of Manuscripts Available in Manipur

As per the catalogue published by Manipur Sahitya Parishad, Imphal and State Kala Academy, there are some 1000 manuscripts available at present in the custody of Manipuri scholars. Many eminent scholars have collected the manuscripts and kept them in their custody. Shri. N. Khelchandra, an eminent scholar, has collected about 500 Meitei scriptures of the early medieval period which are the Meitei counterparts of the epics, literary heritage and similar early written evidence of human civilization. Other scholars who had a good number of manuscripts collection are B. Kullachandra Sharma, M. Chandra Singh, O. Bhogeswar, R.K. Sanahal, T. Madhob, N. Indramani etc. Most of the works of the early medieval Manipuri literature contain no particulars about their author, compilers and editors. This omission is accounted for by the fact that it was then a literary tradition of not disclosing authorship of their works and some works whose authorship was dedicated by the writers to their royal patrons. The names of the authors and scholars have therefore to be ascertained form indirect sources like the Royal chronicles, Clan chronicles. Some of the works of the later medieval period contain particulars of their authors. The variety and richness of the historical literature is a striking feature of the early Manipuri literature. The subject coverage of the manuscripts available in Manipur, ranges from administration, arts and culture, astrology, charms and mantras, creation, lexicography, fine arts, earth science, genealogy, poetry, prediction, prose, religion and philosophy, Meitei scripts, supernatural stories, Meitei confederacy to family genealogies. Here I am enclosing some photographs of the Manuscripts with their brief description.

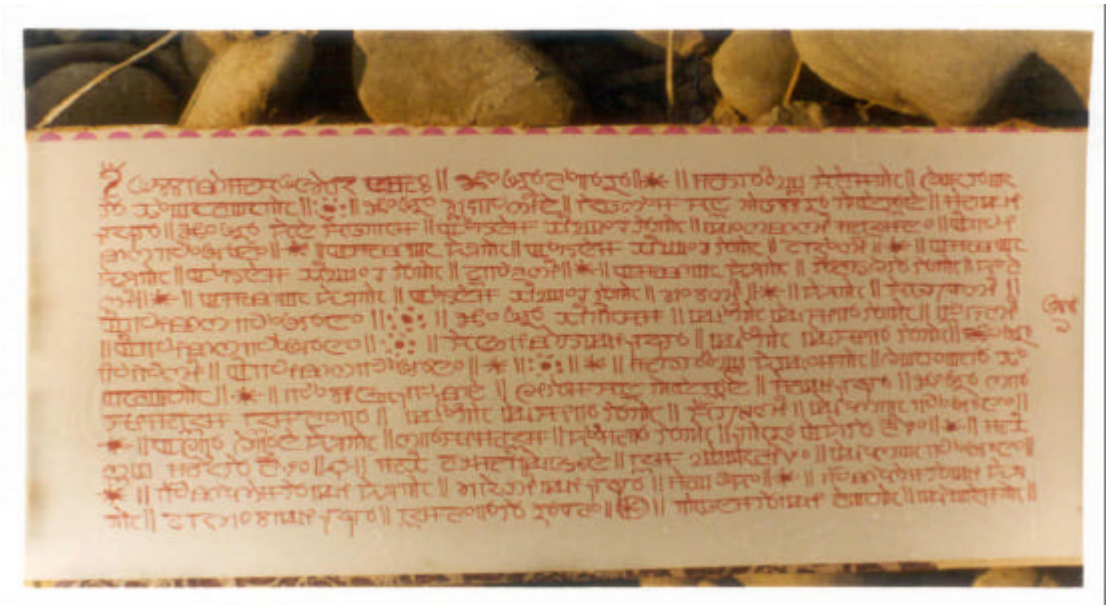


Fig. 1. This Manuscript deals with the Immigrants of the Kshetriyas

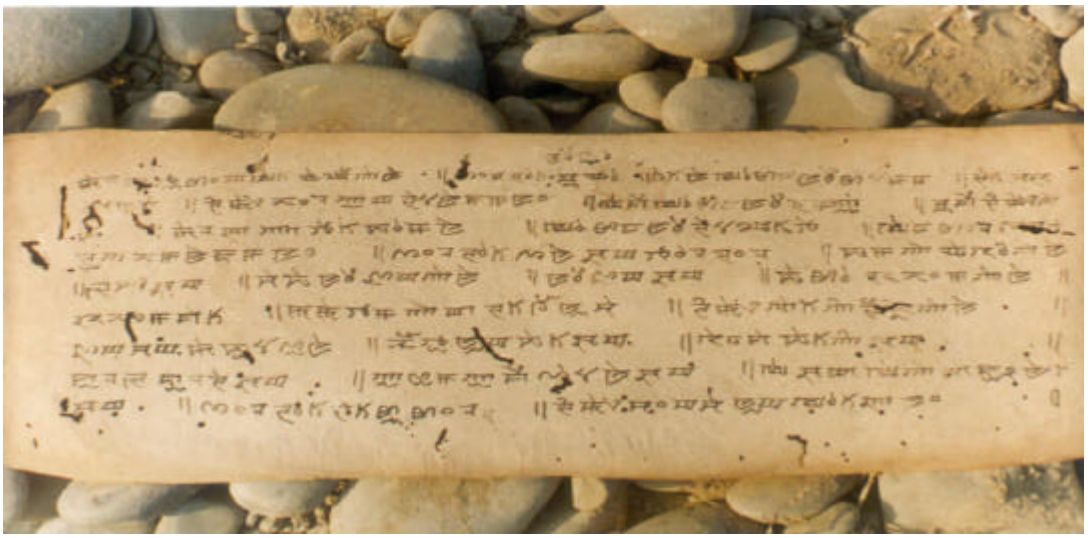


Fig. 2. It deals with the account of the Nongmaijing hills

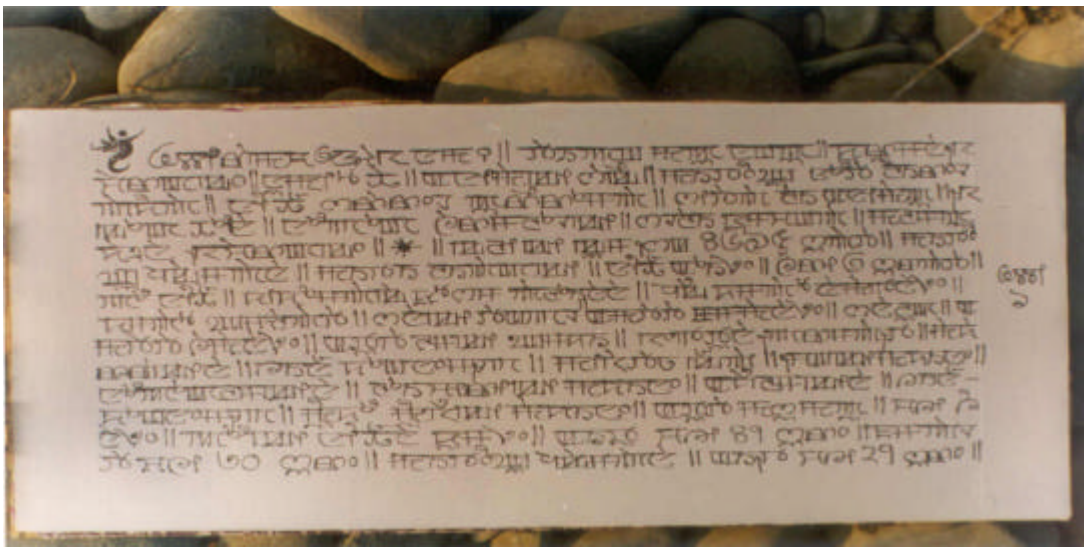


Fig. 3. It is the account of division of land and exchange of cultural materials between Meitei and the Shan

3. Content Management

Content management is a framework to generate, administer, distribute and create possibilities of using and processing electronic content located on the Internet, Intranet or in corporation wide system. It also refers to the process of capturing, storing, sorting, codifying, integrating, updating and protecting any information. Documents have always been at the heart of the organization, originating from a variety of sources like traditional paper documents such as letters, invoices, orders, checks, and other structured business forms and today, many documents originate from electronic formats such as fax, e-mail, and images or data keyed into database, word processor, and spreadsheet files. No matter from where the document originates, the first priority of document management is to get it into a database— whether a relational or object database. Only then we can intelligently manage the document data. As a result, one of the biggest challenges facing document management vendors is providing a standard way of accepting document data from all of these sources and integrating it into one "hub" database for ongoing document management.

Content management involves managing the content through its entire life cycle from creation to archiving. Of course, between these two steps there are various intermediate activities like modification, replication etc. Document management solution aims to streamlines these activities and give users greater control over each one of them. Technology has made document management easy as well as difficult. The advantages of technology in this aspect are lower cost of maintaining documents over long period of time,safely from nature's wrath and easy searching and archiving. But the rise in the popularity of communication media like email etc. has meant that corporate communication is lot less structured than it used to be and thus keeping a track of it isn't always easy.

A typical paper document has three key events: receipt, review, and ready-to-file. These events represent the document life cycle in a nutshell. The transitional document life cycle or workflow of a document takes place in the review stage. This may involve moving the document from an in-box, then stamping, annotating, and linking it to other attachments. The document life cycle may also involve incremental changes and additions. In any case, the document is eventually batched with others and archived in filing cabinets or on microfiche.

4. Benefits of Content Management

- ✍ It provides direct cost benefits by cutting down on the cost of paper that is wasted in storing multiple copies of the same document.
- ✍ It frees up the precious real estate that is needed to store these documents
- ✍ It saves money and time. Any decent document management system would cut down the amount of time you spend digging through the archives looking for that elusive document
- ✍ It further saves time by making reproduction of documents faster than traditional means
- ✍ It also makes your document secure from unauthorized access while keeping them conveniently accessible to authorized users only.
- ✍ E- Document management system offers much better and faster recovery than those possible with paper based ones.

5. Essential Elements of Content Management

A decent content management solution can be significant investment to justify the management of rare documents. Here are some points to remember while going through solution from different vendors.

- Integration
- Scalability
- Easy to use
- Web Based
- Vendor Support
- Cost

After having all this essential elements we should justify each point and see whether we can go through the features we need.

5.1 The Features of content management:

Status reporting : This feature should be able to provide details like when the particular document was created ,who created and when it was modified by whom so that it will help the end user to identify the owner of the document and the various stakeholders in it.

Access Control : A user may have the full right to add, delete or modify a document but in access control we can limit the users on their access for a particular document.

Version control : Document management system should be capable of storing various versions of the same document by keeping track of the current and old.

Retention management : The important function of document management system is to provide an archive of the document for retention purpose.

Disaster Recovery : They should support taking regular backup and quick recovery in case of breakdown with minimum downtime.

6. Content Management Lifecycle

All content management systems are focused on four keys processes which relate to managing each content throughout its life cycle.

- Input/bringing in document
  - Scanning
  - Conversion
  - Importing
- Storing Document
- Indexing Document
  - Index Field
  - Full text indexing
  - Folder/File structure
- Search/Retrieve document

Once the management life cycle is completed then we need to select the hardware/software to achieve the objectives of each of the above stages. Then only, the content management solution will be benefited.

7. Conclusion

Preserving the contents of our world heritage in their original form for infinite future is not only difficult but rather impossible. Thus, we should at least work towards preserving these contents in different formats. Our locally owned collection having traditional formats will continue to be essential, so it is important to realize that the community is interested in this kind of special collection. It is in fact our cultural heritage, the story of our past, reflected in the things that were made by natural or social forces. The value of collection content derives from access and if we are not able to make the people know about the content available to user it is of no use. If it is of no use then there is no value. We have to move forward with a mechanism to support the work of the collection. Then only we are able to achieve our goal. Content management through digitization will change the way in which collections are used and accessed.

8. References

1. Stewart Mc Kie . A New Era of Document Management downloaded on 10th Dec, 04 http://www.dbmsmag.com/9506d14.html
2. Khelchandra Singh (N) Ed. A Catalogue of Manipur Manuscripts. Manipur Sahitya Parishad, Imphal,1984.
3. Content, Computing and Commerce Technology and Trends. The Gilbane report Vol.8 (2) October, 2000 downloaded on 10th October, 2004 http://www.gilbane.com
4. Dua Kunal (2004) Document Management. P C Quest September, 2004

About Authors



Ms. Thiyam Satyabati Devi is currently working as a Project Scientist (LS) in the HRD Division in INFLIBNET Centre. She holds B.Sc. Biochemistry, MLISc and DCA. She had published around 10 papers in Regional, National and as well as in International level. She attended a number of seminars and conferences. She has also worked on a project on Bibliography Compilation of Manipuri Literature under Sahitya Akademi, New Delhi.
Email : satyabati@inflibnet.ac.in



Dr. T.A.V. Murthy is currently the Director of INFLIBNET and President of SIS. He holds B Sc, ML I Sc, M S L S (USA) and Ph.D. He carries with him a rich experience and expertise of having worked in managerial level at a number of libraries in many prestigious institutions in India including National Library, IGNCA, IARI, University of Hyderabad, ASC, CIEFL etc. and Catholic University and Case Western Reserve University in USA. His highly noticeableContributions include KALANIDHI at IGNCA, Digital Laboratory at CIEFL etc. He has been associated with number of universities in the country and has guided number of PhDs and actively associated with the national and international professional associations, expert committees and has published good number of research papers. He visited several countries and organized several national and international conferences and programmes.
Email : tav@inflibnet.ac.in

Streaming Communication for Web Based Training

E Jayabalan R Pugazendi
A. Krishnan

Abstract

Streaming audio became available on the Web in 1995, but with the development of the Synchronized Multimedia Integration Language (SMIL), the technology has reached a new level of maturity. SMIL is based on the XML standard, and allows audio, video, images, and text to be integrated. The implications for Web based pedagogy are tremendous. We now have the opportunity to do training on the desktop in a feature-rich open environment. We will give a basic background in streaming media technology, discuss the standards, the current state of the art, our experience with RealNetworks, how to take advantage of it in an intranet environment, and touch on future developments, including the integration of testing engines.

Keywords : Multimedia, Multimedia Integration Language, Streaming Media

0. Introduction

In the online world of 1999, what you can do is both empowered and constrained by the technology. A good understanding of the limits of the viewing software, end user hardware, and intervening network allows the instructional designer to make the best possible use of what is available - to push the limits while balancing speed and usability. We are now entering a new world where those limits are being pushed back rapidly, and anything is possible. It is still very important to understand the limits, for now, but it is equally important to break down our preconceived notions of what is possible. For a moment, imagine anything is possible!

The new SMIL (Synchronized Multimedia Integration Language) standard allows multimedia content, including text, pictures, sound, and video to be synchronized for a coherent learning experience. Control of all these media is contained in a simple text file (although the format is quite complex). Tools to simplify creation and editing are rapidly being developed. SMIL can greatly reduce the bandwidth required while delivering an experience similar to watching a fully interactive television channel.

1. Definitiions, History and Current Status

Streaming media is defined as network based data, which can be presented to the user before the whole data file has finished transferring. If you see a picture begin to appear on your screen before the transfer completes (e.g., a PNG or some JPG files), or hear an audio file start playing as soon as you click it (e.g., a RealAudio file), that is an example of streaming media.

The primary advantage of streaming is that large audio and video files can be played as they arrive on the computer rather than having to wait for the file transfer to complete. For training in particular, this means that the user interface is much more responsive.

Data can be streamed in a variety of ways:

Pre recorded /	Real-time / on-on-demand	Real-time / on-demand - pulled	Real-time/ lived pushed
Web server	×		
RealMedia™ server	×	×	×

If served via the web server, when a user clicks on a web link to a RealAudio sound file, the data is delivered over the web's HTTP protocol. One kind of data can be presented. If served via the RealMedia server, the data is delivered back to the web browser via a streaming protocol like RTSP (Real Time Streaming Protocol) or UDP. Pre-recorded content could include online training materials, and real time content could include classes and meetings.

SMIL is important for several reasons. First, it integrates various kinds of media. Second, it is an open standard that can be leveraged across all platforms. Finally, it is derived from the W3C Extensible Markup Language (XML), standard; this is important because anything can be defined in XML, and it can be extended on the fly simply by defining new tags.

A SMIL player can act like any other Web browser plug-in, and displays SMIL content over an HTTP connection. However, it can also subscribe to a host group and view an IP multicast, or negotiate the control connection and open a unicast RTSP connection to stream the data.

The TCP/IP protocol upon which the Internet is based is reliable over a wide variety of physical networks because the packets can be retransmitted by TCP, the Transmission Control Protocol. When delivering streaming data over a high bandwidth corporate or university campus network, however, the high reliability of TCP is not required, and retransmissions can slow down performance and take up too much network bandwidth. RTSP is designed to degrade gracefully even if a few packets get lost, and therefore delivers the data faster with lower overhead. Early papers on Real-Time Video concluded that the Web was not suitable for high bandwidth media, because of the inherent delays. The new protocols help deal with these problems.

With IP multicast, a streaming server can send a broadcast message across the network, allowing multiple computers to receive it. The new protocols can allow some packets to expire without retransmission, and new routers can allow the data to go across the network to multiple destinations with only one destination address, the multicast host group address, in the header.

If you want to take advantage of these capabilities, first decide on your needs, then take the time to understand the options. You can then talk with your networking group for help in configuring routers, or to help you decide what technology best suits your needs.

2. SMIL, The Best Direction for Web Based Training

Just as web based training (WBT) has some important gains over traditional classroom training, the use of SMIL allows the Instructional Designer (ISD) to take the training experience one step further. SMIL builds on the existing base of XML standards, tools and experience. It allows for very easy indexing and editing because the control files are all plain text with tags, similar to HTML. It can be used inline within an HTML page, and allows simple extensibility for other applications (such as testing engines) within a well-defined framework. Open standards tend to be simple and durable.

Poorly designed WBT can be a waste of time and money as well as an ineffectual tool for training. Many of the streaming videos foisted upon unsuspecting viewers as WBT show a video of a slide presentation

with an audio narration, or even worse a subject matter expert as a "talking head" giving a lecture. To take advantage of SMIL, build a presentation including the audio track of the lecture combined with streaming text of the speaker's notes, streaming JPEGs of any presentation slides, and perhaps short video segments of any animated processes required to illustrate the topic. Besides being more informative and imminently more useful to the end-user, the performance gains of the SMIL presentation over the pure video are tremendous. The network bandwidth required for audio playback with text and JPEGs can be less than half that of a single video stream. Due to the fact that the JPEGs will take up less bandwidth than the video, the quality of the JPEGs can be much higher and small details on the presentation slides, like text, will actually be readable by the end-user. The audio and text tracks could be localized and presented to the user in the language of their choice without having to re-author the entire presentation in a single monolithic chunk. SMIL gives you the ability to split these data types into separately maintained files and maintain full control over how and when they are displayed to the user. One of SMIL's strongest abilities is controlling when something happens on-screen. The content author can precisely control all events, effects and transitions.

Applications that support open standards tend to be available at low or no cost for academic uses. Many of the instructional design tools available today use proprietary code and custom designed Java plug-ins to display the coursework. Some software companies charge far too much for their WBT solutions and may have simply repackaged their custom computer-based training software engines into a complicated browser plug-in. Using this type of approach to WBT can lead to a variety of problems. Trying to deploy the specialized plug-ins to your client base and dealing with unforeseen incompatibilities caused by these plug-ins can quickly become a maintenance nightmare. Better to rely upon a solution based on open standards where the browser plug-ins are freely available and tested by the Internet at large.

There are still some issues to be resolved with SMIL. Drag-and-drop tools to automatically generate SMIL code are still in their first-generation or in beta testing. Complex SMIL presentations still require hand-coding or at the very least, some hand tweaking to perfect and debug. At the time this paper was authored, writing SMIL presentations complex enough to be called WBT requires knowledge and experience beyond the capabilities of the typical instructional-designer. For now, the use of an experienced web site programmer or a staff member who can be dedicated to learning the technical aspects of SMIL is recommended.

From an authoring standpoint, the emerging collaboration and streaming technologies in Microsoft Office 2000 appear to be simple to use and well integrated into the traditional Office suite of tools. It remains to be seen if the Microsoft products continue to have the server scalability and quality issues which plagued earlier streaming technology releases.

When choosing a streaming technology for WBT many factors must be weighed and evaluated. There are a variety of technical factors such as existing network infrastructure between you and your audience, server platform and availability, and client software maintenance. These will be discussed in "The Real Nitty-Gritty" section below. Other factors, which can sometimes be more important to overcome than the technical issues, include personal experience and comfort level with the technology, institutional politics, and any existing corporate relationships. First and foremost, you need to be comfortable and familiar with the technology you implement. If your instructional designers are all familiar with a specific tool set and very comfortable with the processes and procedures surrounding your existing traditional or CBT training methods, there will probably be a substantial resistance to change. Overcoming any internal training paradigm "inertia" will definitely be an obstacle. Often, training the trainers is the hardest job of all. Convincing the management that a new method of training is needed and it may cost them some money to overcome the technical issues can sometimes be an insurmountable hurdle. Presenting the idea to management requires careful analysis of the actual costs involved. Another potential issue when choosing

a streaming technology within an organization can be factoring in any pre-existing corporate relationships. If your organization has, for example, a strong relationship with Apple Computers, then trying to justify a WBT solution utilizing SMIL instead of QuickTime for streaming media may require good analysis and strong justifications.

Building a WBT module using SMIL follows the same process as building traditional WBT with a few specialized requirements. The first phase of any project is the conceptual brainstorming and storyboarding. This is best done on paper for speed and easy reference. The first step is to define the objective of the WBT module. Decide what information is to be conveyed to the viewer and be specific. Define the project scope, setting definite boundaries encompassing just enough detail to properly cover your objectives. Keep the focus tight and stick to your stated objectives. Next, set requirements for the user experience, thinking about not just what the user will be learning, but how you want them to learn. The idea is to lead them through your learning materials in an organized and straightforward manner. This will help you design the navigational methods used within the WBT module. Try to design the framework first. Don't worry about the graphics yet, work on the layout first. Designing a common look-and-feel that can be reused across modules will help reduce development time on subsequent projects and lend a consistency to your training. Consistent look-and-feel across WBT modules gives the viewer a higher comfort level knowing that even though the subject matter may be new, the process of learning throughout your modules is familiar.

From your storyboard, build a timeline and organize the presentation of your learning materials. Plan the layout of the learning materials and the navigational items. Decide when a particular item needs to be displayed. This will help you determine load orders of your media assets and help identify any constraints imposed on your load order by your target bandwidth. With SMIL different items within your presentation can be specified to load serially or in parallel. It's usually a good idea to make sure that your navigation buttons and other graphics and text load before the user gets to view the video animation on the first page of your presentation.

After the layout and storyboard have been finalized, the next two steps for building your presentation are the design of the interface and the content design of the learning materials. These steps frequently happen in parallel. While one group of graphics artists work on the backgrounds, buttons, graphics and other window dressing, the instructional designers work with the media production staff to plan and create the learning materials.

3. Interface Design

When designing the interface for SMIL presentations, one must consider how the presentation will be displayed. SMIL can be embedded into a web page or displayed stand-alone in the RealPlayer®. The choice can be simple depending on the level of integration desired, use of any courseware testing engines, and finally personal preference. Either way, standard web design rules definitely apply. To be effective, the interface must be clean and uncluttered. The design should encourage the user to explore while intuitively leading them safely through the learning materials in the appropriate order to effectively teach them what they need to know. The design must support the learning materials. Layout of the presentation should lead the user to focus on the learning materials. With time-based control over all display elements, SMIL provides the ultimate in design flexibility.

As with any web based design project, the module must be designed with the lowest common denominator client system in mind. If your audience is within your organization's intranet and there are hardware and software standards in place to assure that all of your users have at least a certain minimum configuration then it is relatively simple to plan your WBT module to fit within those requirements. Typically, a SMIL

presentation should be designed to fit within a 640x480 VGA screen. Remember that the actual usable space within a browser is smaller than the full screen resolution. For a 640x480 VGA display with the web browser window maximized, with default menu settings, the usable screen real estate is approximately 600x300 pixels. When adding a 320x240-pixel video, not much room is left vertically for titles and text.

4. Learning Materials Content Desing

While the graphics artists are busy with the interface, the source materials for the video, audio, and other media clips must be recorded and encoded. Plan and conduct source material recording sessions. Once the project storyboard and layout are finalized, it's time to build the actual "meat" of the presentation. Successfully planning and producing the actual presentation material is simple if you, the designer, have control over the material being presented. More often than not, the audio and video have to be recorded onto cassette or videotape and then digitized and encoded for use on the web.

When dealing with video as a streamed medium, many factors influence the final stream quality and playback rates. The well-known rule of "Garbage in, garbage out." applies to streaming video. The higher the quality of the recording used as source material, the smaller and faster the streaming video file will be. The differences in signal-to-noise ratios and overall resolution between VHS, S-VHS, 8mm, High-8, Mini-DV, BetaCam-SC, and DV-PRO video formats (these are listed in increasing degree of quality) directly influence the playback frame rate and encoded file size of the streaming video file. The better the format you can afford to record in, the cleaner and better your video will stream to your clients.

For important high-bandwidth content, the use of a professional video production staff equipped with proper lighting and recording equipment will always yield a higher quality recording than a consumer-quality video camera. This by no means should be interpreted to mean that low cost, consumer-quality equipment is incapable of producing satisfactory results. However, to provide Internet-based video streams larger than a postage stamp at acceptable quality when network bandwidth is at a minimum, starting with premium quality video recordings is essential. The objective here is to plan the multimedia source materials appropriately taking into consideration the time, resources, and funding required to realistically achieve your design goals.

5. Building Teh Realtext, Realpixm and Smil Files

After the actual content files have been created, the SMIL presentation files need to be created. This is quite similar to creating HTML pages, except SMIL is time sensitive and requires specific timing for each event and transition, and the files need to live on the streaming server, not the web server. For instructions on how to code SMIL files, the SMIL technical documentation can be found at the World Wide Web Consortium Architecture for Synchronized Multimedia. Technological Issues

The topics covered within this section will address issues surrounding manufacture of streaming media for intranet use where high-bandwidth network connections are available. Although SMIL presentations can be adapted to incorporate different sized videos for either high or low bandwidth use, that is outside the scope of this paper.

Once the instructional designer has obtained the source materials for the videos, they need to be digitized. The format into which the video is digitized will affect the final encoded output. Always digitize video uncompressed at 30 frames per second and in Stereo at 16-bit 44-KHz sampling rates. Let the streaming format encoder software have the best quality input so it has the all of the data it needs to provide the highest-quality output. The more data the encoder has to work with, the fewer assumptions the

compression routines need to make. This will result in smoother, cleaner, and smaller encoded video output.

Digital editing of video and audio sources before encoding is usually required. Certain optimizations such as video cropping and audio normalization can be made to provide optimal output upon playback. Applications for video and audio editing include Adobe Premiere® and Sonic Foundry's SoundForge®.

6. Encoding Content to Realneworks Formats

Audio and video can be encoded into the RealNetworks® RealMedia® format using a variety of third party applications. The easiest encoder to use is the RealProducer® Plus G2 from RealNetworks. It has many different stream options. The RealMedia G2 SureStream™ format option allows multiple streams at different bandwidths to be encoded into the same file. This allows the RealPlayer® and RealServer® to better negotiate how much data to send the player based on network performance. For example, a video may be encoded for 28.8K modem, 56K modem, 64K Single ISDN, 128K Dual ISDN, 220K xDSL and Cable Modem, and 150K Corporate LAN data rates all within a single file. Depending on the available network bandwidth, the player will switch between these different encoded formats dynamically as the user watches the video. This feature provides much better playback than older streaming technologies, which only adapt to changing network conditions by dropping frames or "fuzzing-out" the video into large indistinguishable blocks. Depending on the resolution and frame rate of your video source files it may not make sense to encode at the higher bandwidth settings, such as 220K and 150K, and at the low bandwidth settings, 28.8K, 56K, and single ISDN in the same file. The lower settings may not have enough available bandwidth to stream the file.

Streamed animations can be produced using Macromedia's Flash® technology. Flash is in widespread use for non-streamed web based animations. The same animations can be included into your SMIL presentation after a simple encoding procedure into the RealFlash™ format. Now the use for Flash animations is no longer limited to the realm of the static web page and can be unleashed into the dynamic environment of a SMIL presentation. The RealNetworks site has some good examples of RealFlash™ SMIL presentations.

Bandwidth between you and your target audience is the limiting factor on SMIL design. Designing SMIL presentations includes tradeoffs for each data stream sent to the player. The designer must balance data stream buffering times versus compression and the number of streams being loaded simultaneously. These calculations are also affected by the resolution of the data to be streamed. Resizing a video originally intended to stream at 320x240 pixels down to 160x120 will reduce your bandwidth requirements by a factor of four (assuming constant compression rates). The RealNetworks SMIL kit has exhaustive information on this topic.

7. Uploading Files in Streaming Server

As the streaming media files are created, they need to be stored on a separate server running the RealNetworks RealServer® G2 server software. The content creator will need to place the files in a subdirectory off the mount point for the server, and will need the address and port number of the server, as well as whether the Ramgen file system, for sending temporary small files, is in use. The files can then be linked to from any web page. Links can be of the format http://server/ramgen/MountPoint/virtual_directory/filename, and once within SMIL, individual components can be specified in a very similar format, rtsp://server/MountPoint/virtual_directory/filename.

8. Hardware Requirements

Hardware requirements vary widely depending on your application. Four sets of hardware requirements are involved: network infrastructure, web server, stream server, and client browser/player. The web server requirements and configuration are outside the scope of this document. Many of the issues discussed here are particularly important to corporate implementers who have controlled environments into which they wish to introduce streaming technologies. The only successful way to implement streaming media in a corporate environment is to work with the network and computer infrastructure organizations within your company to understand and proactively adapt to the additional requirements imposed by the technology.

Network infrastructure: Both Internet and intranet bandwidth demands should not be underestimated. Careful analysis of current network loads and capacities can help determine how much streaming traffic can be handled before network upgrades are required to provide adequate quality of service to all users. Network upgrades are expensive and time consuming. Always consult with your network operations staff before implementing any streaming technologies on a widespread basis across your network.

Client browser/player: The RealNetworks RealPlayer G2 will currently run on any PC-compatibles running Windows 95, 98, NT4.0, and Power Macintosh. Performance will vary depending on CPU speed and available memory. For fast, responsive control and playback, we recommend a minimum of a 166-Mhz Pentium with 32Mb of RAM Slower machines will provide sub-optimal playback. PCs also need to be MPC-2 compliant and have appropriate sound-cards, drivers, and headphones or speakers installed. For corporate intranet sites, overcoming the current installed base of non multi-media equipped PCs can be a significant challenge.

Stream server: The RealNetworks RealServer G2 products run on a variety of UNIX platforms as well as Microsoft NT. Hardware requirements vary depending on expected number of users. Consult the RealNetworks website for details

9. Tools

SMIL Authoring	RealNetworks RealProducer® Pro G2 Digital Renaissance TAG Author® 2.0 Veon Interactive V-Active® for RealSystem G2
Audio/Video Editing	Sonic Foundry's Sound Forge® 4.5 Adobe Premiere®
Streaming Media Encoding	RealNetworks RealProducer® Plus G2

10. Conclusion

Most currently available web based testing software requires custom format files, special software, and is not standards-based. New products such as TopClass use plain text and HTML, and are much better suited to integration within a SMIL framework. Testing is the next step.

The evolution of the tools currently available will no doubt give rise to a suite of powerful and easy-to-use tools for creating SMIL presentations. Ongoing development of SMIL with ratification via W3 will assure interoperability.

Better integration and tools will allow the potential we see to be realized. Right now, streaming media is at the level of maturity the web was in 1994. The standards are there, and the tools are coming. You can go out and use the technology now. Let us know what you do with it!

11. References

1. W3C Recommendation: Synchronized Multimedia Integration Language (SMIL) at http://www.w3.org/AudioVideo/
2. RealNetworks HTML+TIME at http://www.real.com/
3. W3C Recommendation: Extensible Markup Language (XML) at http://www.w3.org/XML/
4. Comer, Douglas E. The Internet Book: Everything You Need to Know About Computer Networking and How the Internet Works. Prentice Hall, August 1997
5. "Real-Time Video and Audio in the World Wide Web" by Chen, Tan, Campbell, and Li. Proceedings of the Fourth International World Wide Web Conference, December 1995.
6. RealServer Administration Guide at http://service.real.com/help/library/servers.html
7. WBT Systems TopClass Overview at http://www.wbtsystems.com/

About Authors

Mr. E. Jayabalan is a Research scholar in K.S.R. College of Technology, Tiruchengode, Tamil Nadu
Email : ej_ksrcas@rediffmail.com

Mr. R. Pugazendi is a Research scholar in K.S.R. College of Technology, Tiruchengode, Tamil Nadu
Email : pugazendi_r@rediffmail.com

Dr. A. Krishnan is Principal in R.R. Engineering College, Tiruchengode, Tamil Nadu
Email : a_krishnan26@hotmail.com

Streaming Media to Enhance Teaching and Improve Learning

E Jayabalan R Pugazendi
A. Krishnan

Abstract

In a search for more natural way of learning, streaming media is one of the ways. One old Chinese statement says: "a picture is worth a thousand words if used correctly". Today we can ask ourselves how many words worth a video sequence? The use of online courses continues to grow worldwide, yet it is still not clear whether such online learning environments enhance the learning outcome of students or even meet the level of success of traditional classrooms. Many online courses represent no more than electronic versions of traditional classes without, of course, face-to-face interaction between instructor and students. These electronic courses often contain Web materials that lack any significant level of creativity or interactivity . Cognitive research suggests that the addition of multimedia can *actually improve the learning process if certain methods are employed. By using auditory and visual methods of presenting information, students can process that information more quickly, often fostering an enhanced learning process.*

Keywords : Streaming Media, Real-time Multimedia, Multimedia Integration Language, E-Learning

0. Introduction

Steaming media is a method of creating digital video, audio, graphics and text so that it is distributed in "real-time" (synchronously) over the Internet. This means that packets of data are sent or "streamed" from a computer serving the data in real-time. The end user doesn't have to wait for the files to download or store the files on the hard drive in order to view them. A media file is quite large, so the advantage of streaming packets of information is that the end user views the media as the data is received by the user's computer from the server. This technology tool is a powerful asset in delivering instruction from a distance.

Whereas the technology that allows media to be distributed in real-time is a very recent development, the manner in which we apply this technology tool to teaching is not so new. Educational research of the past ten years informs us as to which practices best help students learn using technology tools. The key to the success of using streaming media will depend on how it is integrated into the course's over all instructional design and applied to learning activities.

1. How can streaming media be used for instruction?

When considering the best practices of the use of streaming media, the pedagogy should reflect what we know to be effective teaching and learning practices regardless of delivery method of the instruction. The application of streaming media should be an intricate part of the instructional design and support those essential activities in which the student must engage to achieve the learning outcomes. We know that engagement in interactive activities in which the learner receives feedback regarding performance from either a computer, peer or teacher is a best practice.

Streaming media can be used to support activities such as cooperative learning projects, online discussion or individual practice in applying skills and knowledge. For example, several short case studies which

demonstrate concepts related to the learning objectives might be presented in streaming video and audio format. Students are then instructed to discuss the case online, perhaps answering direct questions or offering solutions to a specific problem. Alternatively, an audio visual segment may be used to present instructional information related to course learning objectives. After viewing the media clip, each student can use a computer-based quiz to assess his/her own learning by answering a series of questions that check comprehension of the learning objectives associated with the media clip.

The benefit of streaming media is that the student can control the pace of the learning process if learning materials and activities are designed to foster interactivity. A learner can also select the media which match the preferred personal learning style: text, audio or visual. Animated models, charts and graphs, which learners can manipulate, will appeal to kinesthetic learners. Streaming media can present information to fit many learning styles.

2. Building Skills and Knowledge

A best practice is that all presentations of information should be clearly associated with specific learning objectives and have some form of interactions to practise the new skills or knowledge before that knowledge is assessed via exams.The more interactions a student has in order to practise new skills and knowledge, the more the student has opportunity to engage in construction of knowledge. Imagine a construction scaffold that vertically spans several stories of a 10 - story building. The top of the building is where a student must be to succeed in your class. At the bottom of the scaffold is the foundation that is necessary to support each of the next levels.

A student may have some background knowledge that is the foundation for adding new knowledge. However, an instructor may discover that he or she will need to provide remediation for the prerequisite skills and knowledge before some students can acquire new ones. It is through guided practice and interactions with new information that a learner constructs meaning. Passively reading or viewing information is usually not enough interactivity for all students to master procedural or conceptual knowledge.

3. Levels of Media Interactivity

- Level 1- student controls stopping and starting of segment
- Level 2- student implements an action and gets a “canned” response
- Level 3- student implements an action and gets a unique response
- Level 4- student inputs unique data and gets a unique response

Level 1 : Lecture, as an instructional strategy, is not interactive. It is presentation of information in audio and visual mode. Research has shown we learn 50% of what we see and hear.

In a face- to- face lecture, a low level of interaction can be incorporated by posing questions to students or considering questions from students. Unless a student gets an actual response to an individual question, lecture is not interactive. If the equivalent of a “lecture” is used in distance learning, then it recommended to pose questions that students have asked ahead of time or ones that the instructor predicts would be asked. An example of level 1 interactivity is a student starting, stopping, rewinding, etc. the RealMedia player while viewing a presentation.

Level 2 : student implements an action and gets a “canned” response (all students get same response) This level can range from multiple choice questions with corrective feedback to intricately designed simulations or role playing activities. An application of this level of interactivity would be the use of

scenario or role play that was created with branching or different paths depending on the response of the student.

For instance, in one such activity designed with the purpose of demonstrating the influences of institutional discrimination, the student views several media clips of a young black woman who is making choices about school and her future. The student, who plays the role of this young woman selects an option that is presented regarding jobs, housing employment , etc. Depending on the choice that is made after viewing each scene, new options become available to the student. Choices are somewhat limited because of certain societal restrictions. For example, the role-playing student experiences “driving while black”. The outcome of the role play is dependent on the path that the student takes as a result of choices.

Level 3 : Student implements an action and gets a unique response - Examples of this level of interaction are similar to those of search engines and at sites like about.com or Ask Jeeves. The computer-based response will be unique to the individual’s search request. High level interactive programming, such as .cgi (common gateway interface) is used for database search. An instructional use of this level might be to have a database of media which the student can search with keywords specific to a certain problem that needs to be solved.

Level 4 : Student inputs unique data and gets a unique response - This level is an example of artificial intelligence or intelligent tutoring systems which are capable of generating a unique response to a specific inquiry. These systems are usually in place at large research institutions or in training programs at NASA. Complex computer programs enable the learner to tailor individual instruction to fit specific needs.

Good uses of streaming media should support students in achieving learning outcomes aimed at:

1. Procedural Knowledge ie. a skill with narrative (text/audio) of procedures : For example a nursing distance learning program uses streaming media to demonstrate how to prepare a microscopic slide with a blood smear. The student can repeat the viewing as many times as necessary. The student then goes to a hands-on lab and practices the skill.
2. Conceptual Knowledge ie a concept such as a case study or problem-based scenario with interactive opportunities to explore different outcomes, or a Flash graph that changes when you input different data

4. Cognitive Apprenticeship Model (Collins, Brown and Newman 1989)

A good model that emphasizes the role of practice is the Cognitive Apprenticeship Model. The analogy of “apprenticeship” suggests that acquisition of thinking and reasoning processes can also be learned by observing an expert and practicing with the guidance of a “master” or expert.

1. Modeling : Teacher gives examples and non-examples of concept or demonstrates skill. One technique is to “think aloud” as the expert proceeds through the steps of cognition.
2. Coaching : Provide students with opportunities to practise newly acquired knowledge, skills and provide feedback, offer suggestions.

   First the practice is carefully guided by the expert and as the learner gains competence, the learner begins to practice independently. Use elaborative feedback rather than just “Yes, that is right.” or “No, that is wrong”.
3. Articulation : Students discuss problem-solving process, knowledge or reasoning.

4. Reflection : Student assesses own cognitive processes by comparing with another student or expert.

5. Exploration : Students pose own problems and continue the quest, asking questions themselves. (Collins, Brown, Newman, 1989, 481-482).

For example, in the context of language learning, a short video and audio segment may demonstrate two persons engaged in a conversation in the new language. The learning objectives are clearly defined for the students. One of the speakers falters over understanding a word that was used. This speaker would think out loud, modeling various ways to try to determine the meaning until the speaker finally comprehends. The targeted vocabulary words appear in text on the screen as a caption as the word is applied in the context of the conversation. After viewing the media segment, the student engages in activities requiring the student to answer questions about the scene and receives feedback. The student can replay the video and audio or select key words from a text list to hear the words pronounced, see definitions, etc.

After practice exercises, small groups of students discuss online their processes of comprehension and mastery of the learning objectives. Finally, students formulate new avenues to explore related to the topic.

5. Conclusion

Streaming media is a powerful tool for learning environments when used effectively. Rather than seek to merely replace face-to-face lectures with audio and video lectures which have few interactions, if instructional developers keep the focus on the learner and what the learner will do with the information, the use of streaming media can reach students with a variety of learning style preferences.

Streaming media encompasses the use of text, audio, video, graphics, animations, simulations so that learners can not only control when they interact with instructional materials and how long, but they can choose the preferred mode of learning: audio, visual, kinesthetic. Steaming media is powerful as it can demonstrate both procedural and conceptual knowledge.

There is a range of interactivity available to the end-user of streaming media. The more a learner can practice new skills, the better the chance for achievement. Since not all learners learn all subjects best by reading texts, by offering alternative learning activities, more learners can be reached.

The instructional design is important to the learner's achievement. Creating apprenticeship experiences which allow the learner to practice skills and knowledge along with an expert is a good model.

In the 21st century, educators are challenged to prepare a diverse workforce to be skilled and knowledgeable employees. We have the technology tools to reach more learners than before. By keeping the focus on the learner and not on the technology alone, educators will be more likely to use sound practices that can help learners with diverse learning styles and ensure success.

6. References

1. W3C Recommendation: Synchronized Multimedia Integration Language (SMIL) at http://www.w3.org/AudioVideo/

2. RealNetworks HTML+TIME at http://www.real.com/

3. W3C Recommendation: Extensible Markup Language (XML) at http://www.w3.org/XML/

4. Comer, Douglas E. The Internet Book: Everything You Need to Know About Computer Networking and How the Internet Works. Prentice Hall, August 1997

5. "Real-Time Video and Audio in the World Wide Web" by Chen, Tan, Campbell, and Li. Proceedings of the Fourth International World Wide Web Conference, December 1995.

6. RealServer Administration Guide at http://service.real.com/help/library/servers.html

7. WBT Systems TopClass Overview at http://www.wbtsystems.com/

8. Collins A., Brown, J.S., & Newman (1989). Cognitive Apprenticeship: Teaching crafts of reading, writing and mathematics. In L.B. Resnick (Eds.), Knowing, learning and instruction: Essays in honor of Robert Glaser (pp. 450-494). Hillsdale, NJ: Erlbaum.

About Authors

Mr. E. Jayabalan is a Research scholar in K.S.R. College of Technology, Tiruchengode, Tamil Nadu
Email : ej_ksrcas@rediffmail.com

Mr. R. Pugazendi is a Research scholar in K.S.R. College of Technology, Tiruchengode, Tamil Nadu
Email : pugazendi_r@rediffmail.com

Dr. A. Krishnan is Principal in R.R. Engineering College, Tiruchengode, Tamil Nadu
Email : a_krishnan26@hotmail.com

Information Life Cycle Management for LIS Professionals in the Digital Era

Ramesh R Naik

Abstract

Information life cycle management (ILM) is a comprehensive approach to managing the flow of an information system's data and associated metadata from creation and initial storage to the time when it becomes obsolete and is deleted. Unlike earlier approaches to data storage management, ILM involves all aspects of dealing with data, starting with user practices, rather than just automating storage procedures. Also in contrast to older systems, ILM enables more complex criteria for storage management than data age and frequency of access. This paper traces various concepts and process of ILM for Library and Information Science (LIS) professionals.

Keywords : Knowledge Management, Information Management

0. Introduction

Information Life Cycle Management (ILM) is a new stage in the era of data storage. It is the latest buzzword in the storage-networking world. These days information must last longer. The value of data changes faster and is often unpredictable. But it must be readily available to provide new openings for growth. As the value of information changes, it makes sense to move data different online and off line storage media that provide the right levels of protection, replication and recovery at the low cost.

Traditionally, data was moved, from primary to secondary and archival storage, based on the files size and age. But the emerging demand for instant access of relevant data has generated a need for storage systems. To get the maximum value from their information at a lowest possible cost, naturally the sectors like banking, healthcare, finance are taking a close look for implementing ILM.

1. What is Information Life Cycle Management?

It is important to understand that ILM is not a piece of computer software, but a process. It is a framework or methodology that enables users to segment their data by value, In other words it is a means of proactive management of information. It aims at improving the effectiveness of organizations by managing information as a resource. Some information professionals state that this is a complex and time-consuming process, but ILM gives the ability to define a set of practices around any set of data, so that anybody can move this data into different classes of storage over a period of time.

2. Purpose

Data is ever growing, and it is not advisable to keep old and current data together. Therefore there should be a policy for shifting old data to separate storage location, The ILM establishes management policies, procedures, and practices governing the initiation, definition, design, development, deployment, operation, maintenance, enhancement, and retirement of automated information systems.

3. Objectives

The idea of an ILM is derived from record management, where the idea of a document life cycle is central to the overall process. The primary objectives of ILM in any organization are:

- ✍ It provides visibility and comprehensive information to functional and technical managers.
- ✍ It focuses on the plans and activities that need to be performed to control organizations records.
- ✍ It delivers quality systems within the cost limitations.
- ✍ It establishes appropriate levels of management authority to provide timely direction, coordination, and control.
- ✍ It ensures an organizational and project management structure and keeps its accountability throughout its life cycle, also identifies project risks early and manages them before they become problems.
- ✍ It delivers systems that work effectively, efficiently and within the planned infrastructure.
- ✍ It ensures the integrity of information.
- ✍ It makes useful information available round the clock and also ensures their protection.

4. Steps for adoption of ILM

Information Lifecycle Management (ILM) has emerged as an approach to library storage that is designed to align users needs and storage practices by infrastructure decisions largely on the value of information. Moreover, librarians should leverage expertise found within the records and information management community, which has long understood that all information has a “lifecycle”. The ILM views records as a cyclical process.

Most organizations will go through three steps when trying to implement ILM; each step requires the planning, organization, coordination and control of a number of activities supported by emerging Information technology, the first step is to eliminate any direct attached storage. Storage needs to be fully networked and then resources can be managed effectively .As a proper subset of communication, information storage and retrieval can be subjected to further analysis. An organization must go through data classification, cataloguing and organizing data according to its value, type and requirements, and also its availability, recovery, security, cost etc.

The second step is to target a number of applications and fix polices for various information types. The third step is to create a tiered storage infrastructure. This is an opportunity to use less expensive disks for data that is used less frequently. Thus ILM is the key for sustaining knowledge creation and application in organizations .The human element is considered in techniques such as information audit and mapping.

5. Phases of ILM

ILM includes six phases, during which infrastructure is created or modified.

5 .1 Initiation Phase

The purposes of this phase are (i) to identify and validate an opportunity to improve accomplishments of the organization;(ii) to identify significant assumptions and constraints on solutions, (iii) to recommend the exploration of alternative concepts and methods to satisfy the need.

5 .2 Concept Phase

This phase will determine whether an acceptable and cost-effective approach can be found to the need or not.

5.3 Detailed Analysis and Design Phase

The purposes of this phase are to further define and refine the functional and data requirements. At the end of this phase, the system is described by a completed high-level architecture and logical design.

5.4 Development Phase

The purposes of this phase is to design, develop, integrate, and test the infrastructure system, and to update and finalize plans to deploy further..

5.5 Deployment Phase

The purposes of this phase is to ensure that (i) the infrastructure system is installed as planned and (ii) the end users are trained; and (iii) supporting organizations are prepared to accept the system.

5.6 Operations Phase

The purposes of this phase is to operate, maintain, and enhance the infrastructure system.

6. Implementation of ILM in the Library

ILM changes the way the information is managed. The parameters set would include retention polices and other rules need to put in place. An ILM based solution stores data, retrieves it and re-tiers it according to how valuable it is to the any given point of time. Before undertaking an ILM approach, an organization needs to know what kind of data it has, how that data is used, and its value at any point of time, based on the requirements for a specific applications-it is really this combination that defines information. The starting point is to understand the types of data associated with each application structured (e.g databases), or semi-structured (e.g.E-mail) or unstructured (e.g files). In addition, the organization needs to determine whether data is transactional or referential data type. Information life cycle will vary from organization to organization depending on the nature of information. The Information professional or Librarian has specific roles and responsibilities related to the management of information, which include:

- Identifying, selecting, acquiring and preserving records, in all media, considered to be of enduring value;
- Developing tools, standards, guidelines, and practices to support institution-specific records and information life cycle management initiatives;
- Serving as a leader in building records management and as a credible resource on records management;
- Managing and protecting less frequently referenced and essential records of institutions in a network of centres ; and building skilled manpower for effectively management of information life cycle.

.ILM may be the right answer to the growing data management gap. The idea is that by matching data to appropriate storage products in a tiered fashion, library administrators will not be only able to better manage increasing volumes of information over time, but they will also increase overall value to the library system. ILM products automate the processes involved, typically organizing data into separate tiers according to specified policies, and automating data migration from one tier to another based on those criteria. As a rule, newer data, and data that must be accessed more frequently, is stored on faster, but more expensive storage media, while less critical data is stored on cheaper, but slower media. However, the ILM approach recognizes that the importance of any data does not rely solely on its age or how often it's accessed. Users can specify different policies for data that declines in value at different rates or that retains its value throughout its life span. Effective evaluation of ILM yields the key outputs that are helpful in identifying areas of strength and deficit, and that point towards actions that can be taken to strengthen the process.

7. Conclusion

The librarians and information professionals need to improve their professional competencies in which scientific, research, methodological, managerial and economical skills are integrated with communicative, navigational, information seeking, retrieval, and analytical design knowledge and also must have a sense of purpose and professional commitment. ILM is a best tool in order to achieve all these more effectively.

8. References

1. Prytherch, R. Information Management and Library Science: a guide to the Literature. Mumbai. Jaiko Publishing House.1997. p1-18
2. Davanport, T. and Prusak.L. Working Knowledge: How Organisations Manage, What they Know. Boston. Harvard Business School Press. 1998. P 5
3. Banka Bihari Chand. Knowledge Management: Tools and Techniques for Librarians. J.of Library Information Science. V.26, N2, Dec 2001. P175-186
4. Burton, P. Information Technology and Society: Implications for the Information Professionals. London Library Association Publishing. 1992

About Authors



Dr. Ramesh R. Naik is a Lecturer in Department of Library and Information Science. Karnatak University, Dharwad. He holds MSc (Botany), MLIS and Ph.D in Library and Information Science. He has presented number of papers in seminar, conferences and journals. He is also a member of many professional bodies.
Email : rameshrnaik@yahoo.co.in

Challenges of Multimedia Watermarking Techniques

E.Jayabalan R.Pugazendi
A. Krishnan

Abstract

Data transmitted through a network may be protected from unauthorized receivers by applying techniques based on cryptography. Only people who possess the appropriate private key can decrypt the received data using a public algorithm implemented either in hardware or in software. Fast implementation of encryption-decryption algorithms is highly desirable. Data-content manipulation can be performed for various legal or illegal purposes (compression, noise removal or malicious data modification). The modified product is not authentic with respect to the original one. The technology of multimedia services grows rapidly, and distributed access to such ser-vices through computer networks is a matter of urgency. However, network access does not protect the copyright of digital products that can be reproduced and used illegally. An efficient way to solve this problem is to use watermarks. A watermark is a secret code described by a digital signal carrying information about the copyright property of the product. The watermark is embedded in the digital data such that it is perceptually not visible. The copyright holder is the only person who can show the existence of his own watermark and to prove the origin of the product Reproduction of digital products is easy and inexpensive. In a network environment, like the Web, retransmission of copies all throughout the world is easy. The problem of protecting the intellectual property of digital products has been treated in the last few years with the introduction of the notion of watermarks.

Keywords : Multimedia Encryption, Watermarking, IPR, Copyrights

0. Watermarking Algorithm

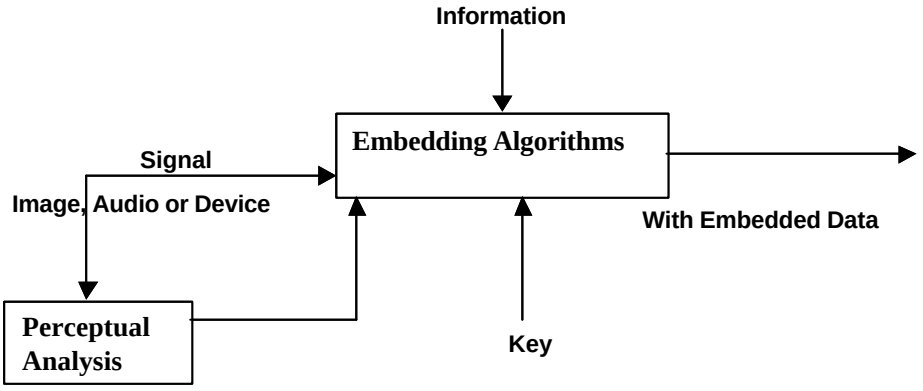
The following requirements should be satisfied by a watermarking algorithm

- Alterations introduced in the image should be perceptually invisible.
- A watermark must be undetectable and not removable by an attacker.
- A sufficient number of watermarks in the same image, detectable by their own key, can be produced.
- The detection of the watermark should not require information from the original image.
- A watermark should be robust, as much as possible, against attacks and image process-ing, which preserves desired quality for the image.

Watermarks slightly modify the digital data to embed non perceptible encoded copyright information. Digital data embedding has many applications. Foremost is passive and active copyright protection. Digital watermarking has been proposed as a means to identify the owner or distributor of digital data. Data embedding also provides a mechanism for embedding important control, descriptive or reference information in a given signal. A most interesting application of data embedding is providing different access levels to the data. Most data-embedding algorithms can extract the hidden data from the host signal with no reference to the original signal.

The first problem that all data-embedding and watermarking schemes need to address is that of inserting data in the digital signal without deteriorating its perceptual quality. We must be able to retrieve the data from the edited host signal. Because the data insertion and data recovery procedures are intimately related, the insertion scheme must take into account the requirement of the data-embedding applications. Data insertion is possible because the digital medium is ultimately consumed by a human. The human hearing and visual systems are imperfect detectors. Audio and visual signals must have a minimum intensity or contrast level before they can be detected by a human. These minimum levels depend on the spatial, temporal and frequency characteristics of the human auditory and visual systems. Most signal-coding techniques exploit the characteristics of the human auditory and visual systems directly or indirectly. Likewise, all data-embedding techniques exploit the characteristics of the human auditory and visual systems implicitly or explicitly. A diagram of a data-embedding algorithm is shown in figure. The information is embedded into the signal using the embedding algorithm and a key. The dashed lines indicate that the algorithm may directly exploit perceptual analysis to embed information. In fact, embedding data would not be possible without the limitations of the human visual and auditory systems.

Data embedding and watermarking algorithms embed text, binary streams, audio, image or video in a host audio, image or video signal. The embedded data are perceptually inaudible or invisible to maintain the quality of the source data. The embedded data can add features to the host multimedia signal, for example, multilingual soundtracks in a movie, or they can provide copyright protection



(Block diagram of a data-embedding algorithm)

1. Watermarking Techniques

Different watermarking techniques have been proposed by various authors in the last few years. The proposed algorithms can be classified into two main classes on the basis of the use of the original image during the detection phase: the algorithms that do not require the original image (blind scheme) [3.147, 3.148, 3.149] and the algorithms where the original image is the input in the detection algorithms along with the watermarked image (nonblind scheme). Detectors of the second type have the advantage of detecting the watermarks in images that have been extensively modified in various ways.

Watermarking embedding can be done either in the spatial domain or in an appropriate transform domain, like a DCT domain, a wavelet transform domain or a Fourier transform domain. In certain algorithms, the imposed changes take into account the local image characteristics and the properties of the human visual system (perceptual masking) in order to obtain watermarks that are guaranteed to be invisible.

The DCT-based watermarking method has been developed for image watermarking that could survive several kinds of image processings and lossy compression. In order to extend the watermarking techniques into video sequences, the concept of temporal prediction exploited in MPEG is considered. For intraframe, the same techniques of image watermarking are applied, but for non-intraframe, the residual mask, which is used in image watermarking to obtain the spatially neighboring relationship, is extended into the temporal domain according to the type of predictive coding. In considering the JPEG-like coding technique, a DCT-based watermarking method is developed to provide an invisible watermark and also to survive the lossy compression.

The human eyes are more sensitive to noise in a lower frequency range than its higher frequency counterpart, but the energy of most natural images is concentrated in the lower frequency range. The quantization applied in lossy compression reflects the human visual system, which is less sensitive to quantization noise at higher frequencies. Therefore, to embed the watermark invisibly and to survive the lossy data compression, a reasonable trade-off is to embed the water-mark into the middle-frequency range of the image. To prevent an expert from extracting the hidden information directly from the transform domain, the watermarks are embedded by modifying the relationship of the neighboring blocks of midfrequency coefficients of the original image instead of embedding by an additive operation.

For example, The original image is divided into 8x8 blocks of pixels, and the 2D DCT is applied independently to each block. Then, the coefficients of the midfrequency range from the DCT coefficients are selected. A 2D subblock mask is used in order to compute the residual pat-tern from the chosen midfrequency coefficients.

Let the digital watermark be a binary image. A fast 2D pseudorandom number-traversing method is used to permute the watermark so as to disperse its spatial relationship. In addition to the pixel-based permutation, a block-based permutation according to the variances of both the image and watermark is also used. Although the watermark is embedded into the mid-frequency coefficients, for those blocks with little variances, the modification of DCT coefficients intro-duces quite visible artifacts. In this image-dependent permutation, both variances of the image blocks and watermark blocks are sorted and mapped according to importance of the invisibility. After the residual pattern is obtained for each marked pixel of the permuted watermark, the DCT coefficients are modified according to the residual mask, so that the corresponding polarity of residual value is reversed. Finally, inverse DCT of the associated results is applied to obtain the watermarked image.

For example, The extraction of a watermark requires the original image, watermarked image and also the digital watermark. At first, both the original image and the watermarked images are DCT transformed. Then, we make use of the chosen midfrequency coefficients and the residual mask to obtain the residual values. Perform the EXCLUSIVE-OR operation on these two residual patterns to obtain a permuted binary signal. Reverse both the block and the pixel-based permutations to get the extracted watermark.

A video sequence is divided into a series of Group of Pictures (GOP). Each GOP contains an interframe (I-frame), forward-predicted frame (P-frame) and bidirectional predicted/interpolated frame (B-frame). P-frame is encoded relative to intraframe or another P-frame. B-frame is derived from two other frames, one before and one after. These non-intraframes are derived from other reference frames by motion-compensation that uses the estimated motion vectors to construct the images. In order to insert the watermark into such kind of motion-compensated images, the residual patterns of neighboring blocks are extended into the temporal domain and other parts of the image. Watermarking techniques can be applied directly into non intraframes.

For a forward-predicted P-frame, the residual mask is designed between the P-frame and its reference I- or P-frame, that is, the watermarks are embedded by modifying the temporal relationship between the current P-frame and its reference frame. For a bidirectionally predicted or interpolated B-frame, the residual mask is designed between the current B-frame and its past and future reference frames. The polarity of the residual pattern is reversed to embed the water-mark

2. Main Features of Watermarking

Watermarks are digital signals that are superimposed on a digital image causing alternations to the original data. A particular watermark belongs exclusively to one owner who is the only per-son that can proceed to a trustworthily detection of the personal watermark and, thus, prove the ownership of the watermark from the digital data. Watermarks should possess the following features

- ✍ Perceptual invisibility : The modification caused by the watermark embedding should not degrade the perceived image quality. However, even hardly visible differences may become apparent when the original image is directly compared to the watermarked one.
- ✍ Trustworthily detection : Watermarks should constitute a sufficient and trustworthily part of ownership of a particular product. Detection of a false alarm should be extremely rare. Watermark signals should be characterized by great complexity. This is necessary in order to be able to produce an extensive set of sufficiently well distinguishable watermarks. An enormous set of watermarks prevents the recovery of a particular watermark by trial-and-error procedure.
- ✍ Associated key : Watermarks should be associated with an identification number called watermark key. The key is used to cast, detect and remove a watermark. Subsequently, the key should be private and should exclusively characterize the legal owner. Any dig-ital signal, extracted from a digital image, is assumed to be a valid watermark if and only if it is associated to a key using a well established algorithm.
- ✍ Automated detection/search : Watermarks should combine easily with a search procedure that scans any publicly accessible domain in a network environment for illegal deposition of an owner's product .
- ✍ Statistical invisibility : Watermarks should not be recovered using statistical methods.For example, the possession of a great number of digital products, watermarked withthe same key, should not disclose the watermark by applying statistical methods.Therefore, watermarks should be image dependent.
- ✍ Multiple watermarkings : We should be able to embed a sufficient number of different watermarks in the same images. This feature seems necessary because we cannot pre-vent someone from watermarking an already watermarked image. It is also convenient when the copyright property is transferred from one owner to another.
- ✍ Robustness : A watermark that is of some practical use should be robust to image modifications up to a certain degree. The most common image manipulations are com-pression, filtering, color quantization/color-brightness modifications, geometric distor-tions and format change. A digital image can undergo a great deal of different modifications that may deliberately affect the embedded watermark. Obviously, a watermark that is to be used as a means of copyright protection should be detectable up to the point that the host image quality remains within acceptable limits.

3. Conclusion

Adapting signal compression to networked applications may require some changes in the fundamental approach to this problem. The compression and transmission aspects have generally been treated as separate issues. The first problem with this approach is that the resulting compression algorithms usually do not address the needs of networked transmission. A successful compression algorithm removes all the redundancy, and, hence, the compressed data must be delivered error free. Another consideration in designing compression techniques for network use is to identify the impact of losing different portions of a compressed stream. It is preferable to have the important parts of the compressed stream concentrate into a short and identifiable segment.

Signal-processing techniques can be valuable for hiding a watermark (or identifying information) in the media. Watermarks can play a number of roles. First, a watermark can mark or identify the original owner of the content, such as the image creator. Second, it can identify the recipient of an authorized single-user copy. Third, a watermark can be used to identify when an image has been appreciably modified. An appropriate solution for the watermarking problem requires understanding of both the signal coding and networking or security issues.

Multimedia processors that realize multimedia processing through the use of software include those for bit manipulation, arithmetic operations, memory access, stream data I/O and real-time switching. The programmable processors for multimedia processing are classified into media-enhanced microprocessors (CISC or RISC), embedded microprocessors, DSPs and media processors.

Many critical research topics remain yet to be solved. From the commercial system per-spective, there are many promising application-driven research problems. These include analysis of multimodal scene-change detection, facial expressions and gestures, fusion of gesture/emotion and speech/audio signals; automatic captioning for the hearing impaired or second language television audiences; multimedia telephone and interactive multimedia services for audio, speech, image and video contents.

From a long-term research perspective, there is a need to establish a fundamental and coherent theoretical ground for intelligent multimedia technologies. A powerful preprocessing technique capable of yielding salient object-based video representation would provide a healthy footing for online, object-oriented visual indexing. This suggests that a synergistic balance and interaction between representation and indexing must be carefully investigated. Another fundamental research subject needing our immediate attention is modeling and evaluation of perceptual quality in multimodal human communication. For a content-based visual query, incorporating user feedback in the interactive search process will be also a challenging but rewarding topic.

4. Reference

1 R. B. Wolfgang, C. I. Podilchuk, and E. J. Delp, “Perceptual Watermarks for Digital Images and Video”, submitted to the Proceedings of the IEEE, 1998.

2. F. M. Boland, J. J. K. ´ O Ruanaidh, and W. J. Dowling, “Watermarking digital images for copyright protection,” in Proc. Int. Conf. Image Processing and Its Applications, vol. 410, Edinburgh, U.K., July 1995.

3. M. Barni, F. Bartolini, V. Cappellini, and A. Piva, “A DCT-domain system for robust image watermarking,” Signal Processing (Special Issue on Watermarking), vol. 66, no. 3, pp. 357–372, May 1998.

4. P. Bas and J.-M. Chassery, “Using fractal code to watermark images,” in Proc. Int. Conf. Image Processing (ICIP), vol. 1, Chicago, IL, 1998.

5. P. Bassia and I. Pitas, “Robust audio watermarking in the time domain,” in Proc. European Signal Processing Conf. (EUSIPCO 98), Rhodes, Greece, Sept. 1998.

6. W. Bender, D. Gruhl, and N. Morimoto, “Techniques for data hiding,” in Proc. SPIE, vol. 2420, San Jose, CA, Feb. 1995, p. 40.

About Authors

Mr. E. Jayabalan is a Research scholar in K.S.R. College of Technology, Tiruchengode, Tamil Nadu
Email : ej_ksrcas@rediffmail.com

Mr. R. Pugazendi is a Research scholar in K.S.R. College of Technology, Tiruchengode, Tamil Nadu
Email : pugazendi_r@rediffmail.com

Dr. A. Krishnan is Principal in R.R. Engineering College, Tiruchengode, Tamil Nadu
Email : a_krishnan26@hotmail.com

Automatic Ontology Generation for Semantic Search System Using Data Mining Techniques

K R Reshmy S K Srivatsa Sandhya Prasad

Abstract

Here we present about automatically generated ontologies for a semantic web search system using data mining techniques. This will improve the query process and will get better semantic results. Ranking algorithm[1] is used to search and analyze web documents in a more flexible and effective way. Hyperlink structure of web document is utilized to rank the results. We use association rule mining to find the maximal keyword patterns. Clustering is used to group retrieved documents into distinct sets. This will extract knowledge about query from the web, populate a knowledge base. The search engine that searches the web documents so far are syntactic oriented. Here we develop a searching system that semantically searches the documents. The semantics of the terms is achieved using the ontologies. Ontology serves as Meta data schemas, providing a controlled vocabulary of concepts, each with explicitly defined meaning. Ranking algorithm used here is the hyper textual ranking algorithm that scans both the contents of the documents and also the reciprocally linked documents. This technique has several advantages that include providing better semantic notion during the search. It also serves for multiple frame documents. There is a need for automatic generation of ontologies when using the semantic searching system. The paper here focuses on how the automatic generation of ontologies could be done for a semantic search system using data mining techniques.

Keywords : Ontology, Data mining, Semantic web, Information retrieval

0. Introduction

World wide web is the most excited society in the last 20 years. Web has turned to be the largest information source available in the planet. It is the huge, explosive diverse, dynamic and mostly unstructured data repository, which supplies incredible amount of information and also raises the complexity of how to deal with the information from the different perspectives of view-users, web service providers, business analysts. The users want to have effective search tools to find relevant information easily and precisely. Web mining is the term of applying data mining techniques to automatically discover extract useful information from the web.

The web mining taxonomy

The web mining is has three fundamental dimensions, namely web content mining, web structure mining and web usage mining. The in-depth taxonomical classification is depicted in Fig. 1:

Web content mining

The web documents have heterogeneous structure which makes difficult to categorize, filter or interpret documents. So a more intelligent tool is needed for information retrieval such as intelligent agents. Moreover advanced database and data mining techniques are required to provide higher level organization of semi-structured data available on the web. Some efforts include:

Agent based approach

This involves development of sophisticated AI systems that can autonomously or semi-autonomously discover and organize web-based information on behalf of a particular user. The agent-based web mining system can be categorized as follows:

Intelligent Search Agent

This system has the characteristics of a particular domain to organize and interpret the discovered information. It relies on pre-defined and domain specific information about particular types of documents or on the coded models of the information sources to retrieve and interpret documents. Several intelligent web agents have been developed that search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information. Traditional search techniques use keywords as input to find the information that a user wants. But here only very little relevant portions only the user will get. Here we use the data mining techniques to search the web documents in a more efficient way. We use a new hyper-textual ranking algorithm to look much deeper into the content of linked documents. If a ranking algorithm superficially considers that all the linking documents generated automatically by web document publishing tools. To rank multiframe web documents also we can use the new ranking algorithm. The new ranking algorithm utilizes the hyperlink and hyper document information to address the need to include ranking criteria for rich and relevant content in the search for information on the web. Web mining is the data mining techniques applied on the web. We use three algorithms association rule mining, sequential pattern mining, and clustering. Weighted association rule mining is used to retrieve the frequently accessed keywords in retrieved web documents. This is helpful since it also shows the relationships between data items. The best techniques to mine the online documents is clstering.That will segement the data into groups.so here we can use clustering to group related words together.Fuzzy C clustering algorithm is used to divide the retrieval documents into a user specified number of groups.

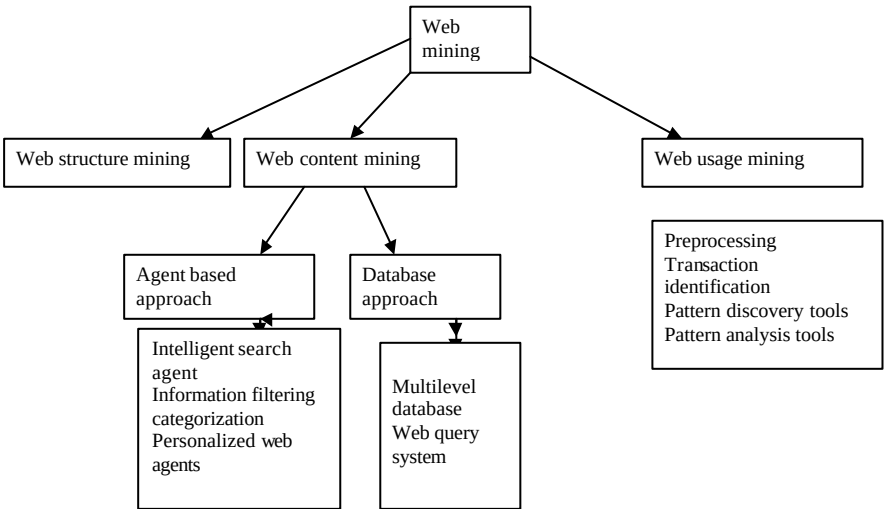


Fig. 1 Web mining taxonomy

The technique we have used is the efficient web content mining. Here an intelligent searching algorithm based on hyper textual information is used. It is coupled with automatic generation of ontologies.

1. Existing technologies

Every search engine must have 3 main components: crawler & indexer, searcher & ranker and interface.

1.1 Crawler and indexer

A crawler is also called as a robot, agent or spider .It is an unattended program that works continuously and automatically .It automatically scans the websites and collects web documents. From the links the crawlers find the other related documents The order to visit the linked documents is done by depth first and breadth first searches coupled with heuristics .Then the documents are indexed.

1.2 Searcher, ranker and interface

The user enters the keyword. Then the searcher scans the indexed documents matching the keyword. The ranker performs the ranking function and hence determines the order in which the document must be displayed to the user.It receives the queries and displays the result. The interface must be simple, intuitive and easy to use.

1.3 Ranking algorithms

- Vector space model: Most of the engines use these ranking algorithms .It is based on the hyperlink structure (the quality measure of a web document is determined by counting the number of documents that has links to a document).
- Page rank algorithm: This algorithm is used in google search Engine.
- HITS: Hyper Link Induced Topic search. This algorithm finds the authoritative documents .these documents are said to be information rich. The algorithm also tracks the hub documents i.e. documents that have links to many authority documents. Some other ranking techniques include edge-weighted strategies, extended HITS etc.
- Hyper textual ranking algorithm: The hyperlinks between the hyper documents contains useful information .This information is utilized in the hyper textual algorithm. The contents of the linked documents are also evaluated. It provides better semantic notion.

1.4 Reasons for selecting the algorithm

- It provides better semantic notion by retrieving relevant documents.
- It uses ontologies to identify and rank relevant web documented semantically.
- Thus the problems of polysemy, synonym and content sensitivity are prevented.
- It also ranks multi frame web documents.
- The text and hypertext are not just evaluated but also the contents in the reciprocally linked documents.

1.5 Data mining algorithms

To enhance user friendly searching some data mining algorithm are used. The 2 most significant algorithms are

- Associate rule mining: These algorithm explorers the frequently used keywords set which can be used for subsequent querying.
- Fuzzy c-Means clustering algorithm

1.6 Ranking process

The new Ranking Algorithm used here [1] Inthis the compounded ranks into three categories of sorted order.These three categories are1)documents that contain all main search concepts.(2) documents that contain some of the main search concepts and have linkage relationships with other concepts.(3)documents that have linkage relationships with some of the main search concepts. Here hypertext characteristics of web documents and ontology are used to model the ranking algorithm to provide more flexibility.

2. Proposed Architecture

The semantic web search system that uses the hyper textual ranking algorithm discussed so far has several advantages. The proposed intelligent search system has 6 main components:

They include crawler, language processor, interface, query engine miner and db connector.

2.1 The crawler

It is also known as agent,an unattended program that works continuously ,and automatically ,having the essential role of locating information on the web and retrieving it for indexing.

2.2 Language processoR

It is used by all theother components to process textual information.

2.3 Interface

it provides a user with a way to input query terms,request mining process, and display query and mining results.

2.4 Query Engine

It is the heart of the system.i t searches the inverted file indexes,which are created by the crawler in our index databasefor efficient retrieval of the documents matching the query terms provided by the user.It uses the linking structure of the retrieved documents to expand the query results.The new ranking algorithm[1] helps to display the order based on the degree of how well the results match the user's query.

2.5 Miner

It provides several kinds of data mining techniques .clustering groups the retrieved documents for a user's query.association rule mining uses the retrieved documents for more specific documents.

2.6 Data base connector

There are five databases that are connected by connection threads to enhance efficiency of the system. The databases are

3. Ontology

The ontology in this search system act as conceptual backbone for semantic document access by providing a common understanding and conceptualization of a domain. ontology consists of two main components; term, & term relationship. Term is the basic terms comprising the vocabulary of a domain. Term relationship is asset of relationship between terms. Populating ontologies with a high quantity and quality of instantiations is one of the main steps towards providing valuable and consistent ontology based knowledge services. Manual ontology population is very labour intensive and time consuming. Some semi automatic approaches have been presented, but are not adequate. Here we present a fully automatic approach of feeding the ontology with knowledge extracted from the web. Information is extracted with respect to a given ontology and provides XML files, one per document, using tags mapped directly from names of classes and relationships in that ontology. The following figure shows an example of the XML representation of the extracted knowledge and how it is asserted in the ontology.

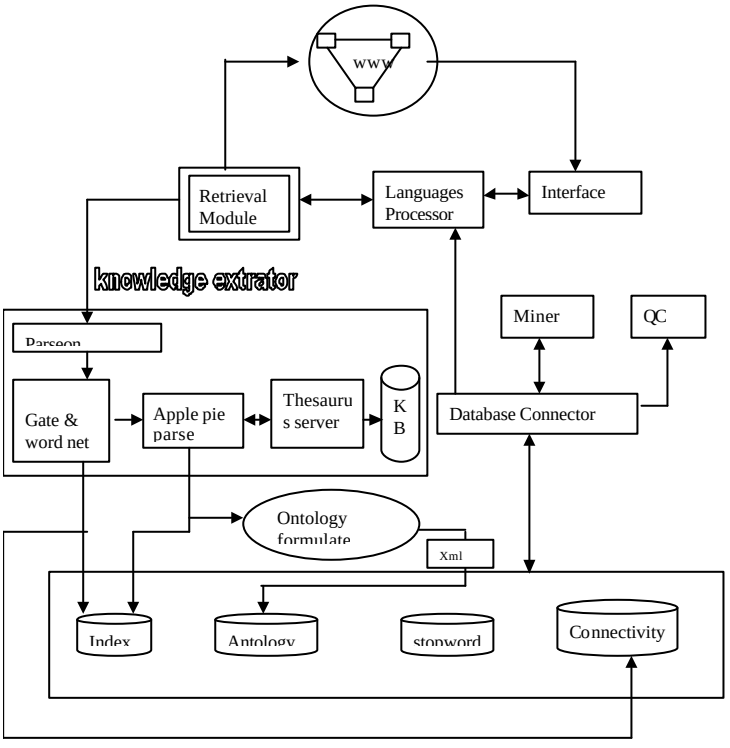


Fig.2 Proposed system architecture

The system has been expanded to 8 components

3.1.1 Crawler

The crawler retrieves documents and sends for indexing to the index db. It has four modules retrieval, URL listing, formatting and modeling & hypertext parser. Retrieving module is sed to retrieve information from the web. This module fetches URLs from large storage of candidate URL's stored in the URL listing module. The hypertext parser module processes retrieved resources. It will

1. determine the retrieved data type.
2. parsing the retrieved hypertext documents
3. extracting the hyperlinks and specified structures in the documents.

The results are then passed to the Formatting and Indexing Module. The processing module adds the parsed URL'S to the URL listing module. The Language processor efficiently and effectively converts the retrieved text into uniform expression used for data mining and information retrieval. After converting the text the formatting and indexing module updates the index database to index the gathered web documents for later searches, adds the hyperlink structure information to the connectivity data base, and splits all the sentences in the acquired documents for the sentence database. The URL listing module feeds the retrieving module, and makes some decisions for selecting URLs from the processing module to be added to the pol of candidate URLs.

3.4.2 Language processor

It processes the textual information for the other components. Our system can process data in English language. Three processes are there. Case translator, Word Stemmer and Stopword Filter. To convert all the retrieved English words into a lower case, case translator is used. Word Stemmer reduces words to their morphological root. Stopword filter removes insignificant words.

3.4.3 Knowledge extractor

Documents on the web use limiteless vocabularies, structures and composition styles. This make it hard for any IE technique to cover all variations of writing patterns. traditional IE systems lack the domain knowledge required to pick out relationships between the extracted entities.

Here ontology is coupled with a general purpose lexical database(word Net) and an entity recogniser (GATE) as guiding tools for identifying knowledge fragments consisting of not just entities, but also the relations between them.. Then performs knowledge extraction. The output of the extraction process is an XML representation of the facts, paragraphs, sentences and keywords identified in selected documents. It has

- Parser- to parse documents into paragraphs and sentences.
- Apple-Pie Parser groups grammatically related phrases to derive relationships.
- GATE and Word Net identifies terms
- Thesaurus server- To query the thesaurus knowledge base.

The following figure shows an example of knowledge extraction.

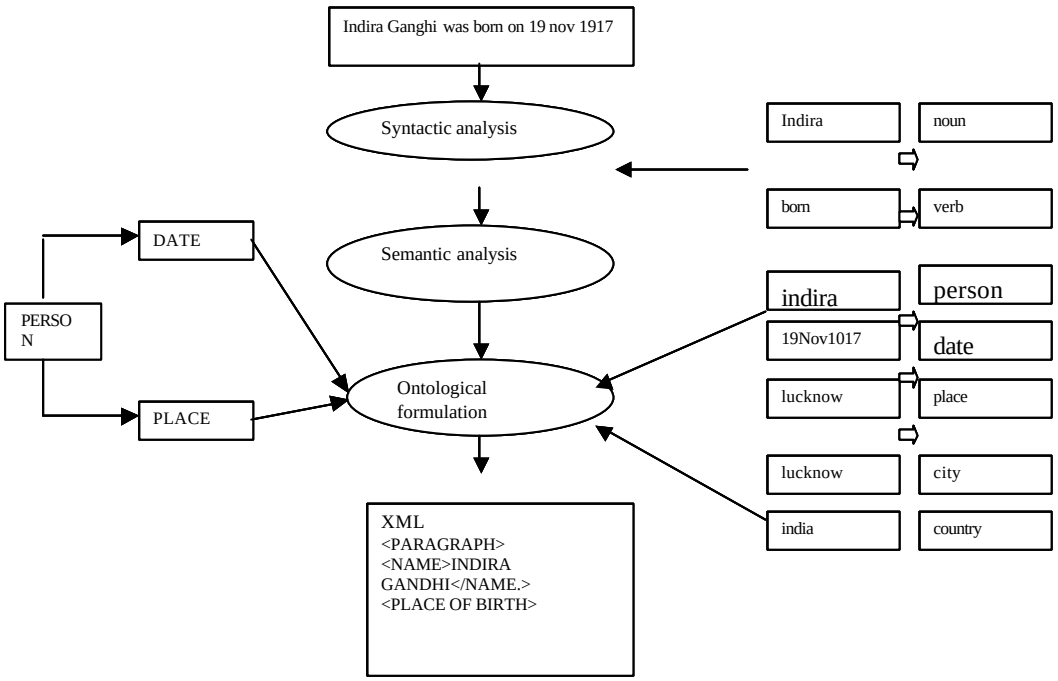


Fig (3) example for knowledge extraction

3.4.4 Db connector It connects all the four data bases.

Index db -This db has term table and documents table. The documents table has document Id and posting id columns. Posting id represents the position of the document identified by the doc id. The term table has term id and posting id. Where the posting id specified the document id which has the corresponding term id.

- Connectivity db : It also has two tables such that a doc-id can point to a given doc-id quickly and vice versa.
- Stop word db : It has 615 insignificant terms for stop word filtering.
- Ontology db : This db has domain specific ontologies. They are automatically fed with knowledge extracted from the web.
- Thesaurus db -This has record of words and related words and their degree of semantic correlation.
- Thesaurus server : To query the thesaurus knowledge base.

3.4.4 Query Engine

The Query engine consists of ranker and searcher. It searches the inverted file indexes, which are created by the crawler in the index database for efficient retrieval of the documents matching the query terms provided by the user. It uses the linking structure of the retrieved documents to expand the query result. The ranking algorithm displays the order of web documents based on how well result matches the user query. The new Ranking Algorithm is used here. [1] In this the compounded ranks are separated into three categories of sorted order. These three categories are (1) documents that contain all main search concepts. (2) documents that contain some of the main search concepts and have linkage relationships with other concepts.(3) documents that have linkage relationships with some of the main search concepts. Here hypertext characteristics of web documents and ontology are used to model the ranking algorithm to provide more flexibility.

3.4.5 Miner

It provides several kinds of data mining techniques. Clustering groups the retrieved documents for a user's query. Association rule mining uses the retrieved documents for more specific documents. The miner groups the retrieved documents in clusters as specified by the user. Then the top 10 representative keywords and corresponding documents are displayed. It also traces the frequently used keyword sets for subsequent usage.

4. Searching Process

The retrieval module retrieves URL s from URL listing module. The hypertext parser parses and fetches the links which are added to the listing pool. The gathered web documents are given to the index and forming module that will be send to the indexed databases for indexing. After a user inputs several keywords for searching relevant web documents, our searcher performs a lookup of the terms in the ontology databases to get the ontology containing these words. The Ontology will be created automatically. The searcher scans the search index in the index database for every key term in search concepts to obtain all the of the conceptually related documents. Then the ranker uses these documents and ontology for ranking and filtering on order to get a sorted document list for all of the relevant documents corresponding to the user's query.

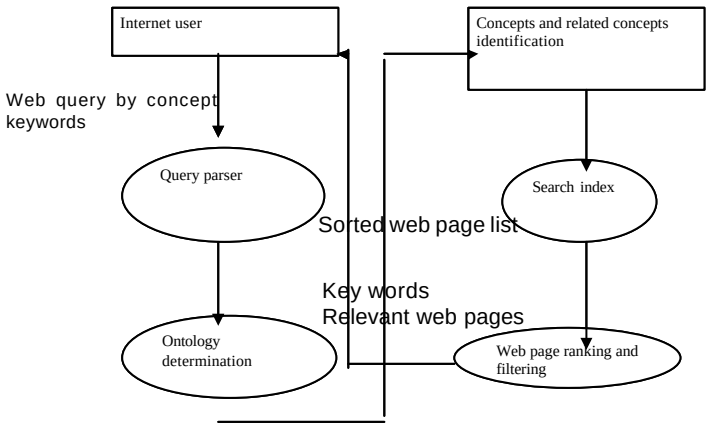


Fig 4 searching process

6. Ontology

Ontology is the conceptual back bone for semantic document access Here ontology is created automatically.

6.1 Automatic Ontology Population

Manual ontology population is time consuming. Populating ontology with a high quantity and quality of instantiations is one of the main steps towards providing valuable and consistent ontology based knowledge services. The ontology is selected only if it already exists. If the term is new then human intervention is needed. So there is a need of automatic approach of feeding the ontology with knowledge extracted from the web. There are certain semi-automatic methods which are at infancy. In such cases document annotations were created and results were stored as assertions in an ontology. Here we present a fully automatic approach of feeding the ontology with knowledge extracted from the web. Information is extracted with respect to a given ontology and provides XML files, one per document, using tags mapped directly from names of classes and relationships in that ontology.

The following Figure shows an example of the XML representation of the extracted knowledge and how it is asserted in the ontology.

```
paragraph>
<url>http://www.kings.edu/womens-history/igandhi.html>
<text>Indira Gandhi was born on nov 19,1917 at Lucknow and would be the only child of Jawaharlal and Kamala Nehru.Being inspired by her father Mr.Jawaharlal Nehru,Ingira Gandhi rose to power in India and eventually became the Prime Mininster of INDIA.>
</text>
<sentence>
<text> Indira Gandhi was born on nov 19,1917 at Lucknow </text>
<person>
<name>Indira Gandhi</name>
<place-of-birth>lucknow</ place-of-birth>
<date-of-birth>
<day>19</day>
<month>11</month>
<year>1917</year>
</ date-of-birth>
</person>
</sentence>
…..
<sentence>
<text>she was inspired by her father Mr.Jawaharlal Nehru</text>
<person>
<name>Indira Gandhi</name>
<inspired-by> Mr.Jawaharlal Nehru</inspired-by>
<person>
</sentence>
…….
</paragraph>
```

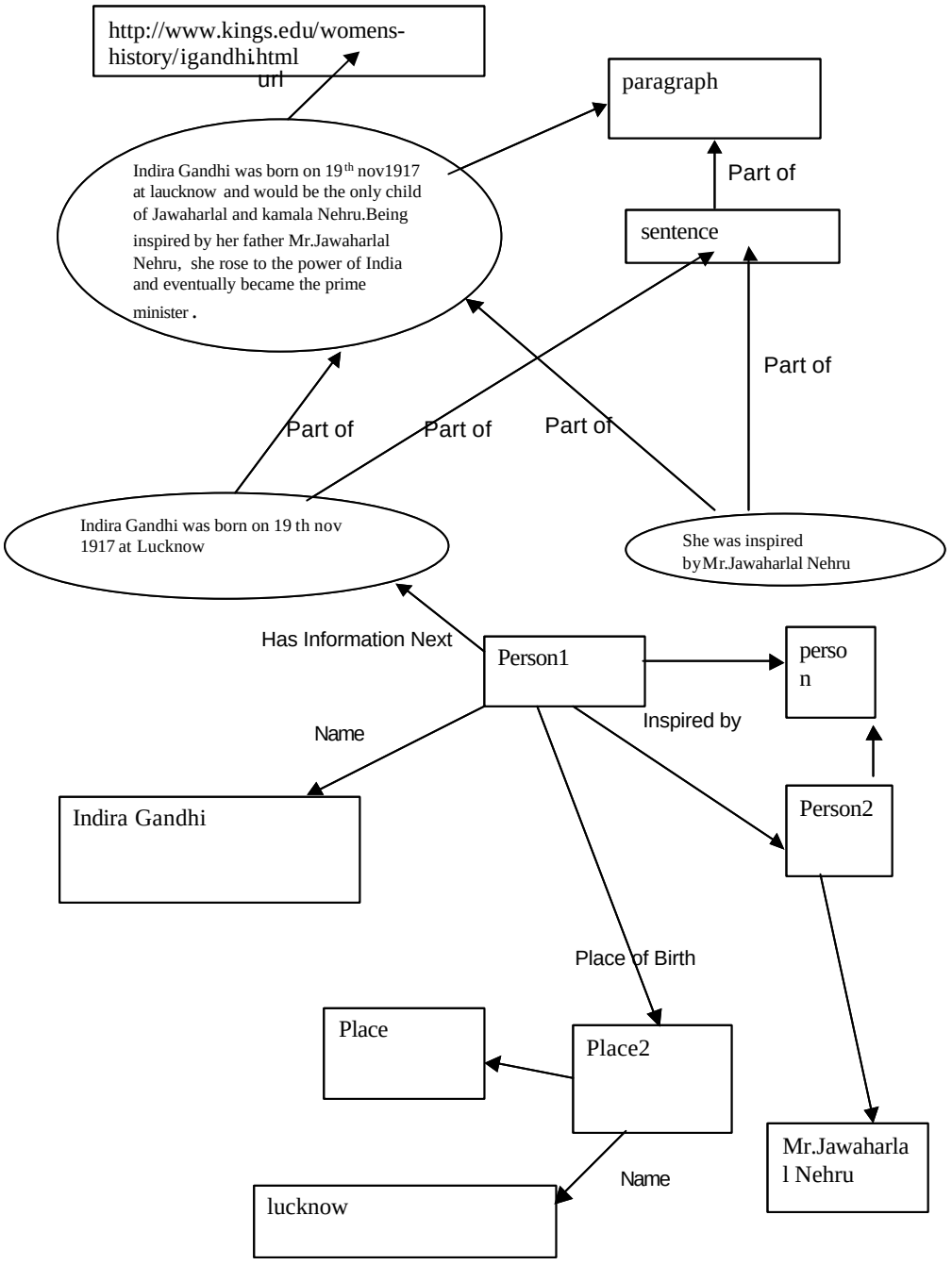


Fig. 6 Ontology representation for terms like JSP and tags

The above diagram represents the basic ontology representation for the terms JSP and tags. The semantic degree is the value between -1 to 1 where it represents the relevancy of two terms (synonym) or irrelevancy (polysemy).Semantic degree is a property of the term ontology. The terms are updated into the index database. The term name, relationship form APPLEPIE and the SD from the thesaurus KB are updated into the ontology database. The terms with properties are formulated into an XML format and updated in the ontology database. Thus the new ontologies are generated automatically into the database to provide fast access to frequently used information via SQL queries.

7. Conclusion and Future Work

The system we produced is more effective and flexible for web search with more semantic notion. We utilize hypertext characteristics of web documents and ontology to model the ranking algorithm to provide more flexibility. This system allows users to select the desired search domains that can correctly locate the documents they are looking for based on relevancy. The new ranking algorithm to look much deeper into the content of linked documents. In addition we used automatically generated ontology to solve traditional problems in text search that involve synonymy, polysemy and sensitivity. Data mining techniques are used to refine our search engine. Three useful techniques were used. association rule mining to explore primary keywords of retrieved documents, fuzzy c-means clustering to provide an overview of the desired documents. We implemented the system and tested it with English web documents from our university web sites. Preliminary results show that our web search system is effective and efficient. The system we produced integrates a variety tools in order to automate an ontology – based knowledge acquisition process and maintain a KB. In the future we can improve the flexibility of our system even further. Issues of duplicate information across documents and redundant annotations are still major challenges of automatic ontology population. Automatically populating an ontology from diverse and distributed web resources poses significant challenges.

8. References

1. Using data mining to construct an intelligent web search system. Yu-Ru Chen, Ming-Chuan Hung and Don Lin Yang, 2003
2. Automatic knowledge based extraction and tailored biography generation from the web Harith Alani, David Le Millard, Sanghe Kim, IEEE Intelligent system 18(1),2003,14-21.
3. R. Agrawal and R. Srikant, “Mining Sequential Patterns”, in Proc. of the 11th international conference On Data engineering,1995, PP.3-14.
4. http://www.google.com/
5. http://www.yahoo.com/
6. Fciravegna,A.Dingli,Y.Wilks,and D.Petrelli,’Timely and Non- Intrusive Active Document Annotation via Adaptive Information Extraction,’’Workshop on semantic authoring,Annotation & Knowledge Markup ,15th European Conference on Artificial Intelligence Lyon,France,2002.
7. S. Handschuh, S. Staab, and F. Ciravegna, “S-CREAM - Semi-Automatic Creation of Metadata,” Semantic Authoring, Annotation and Markup Workshop, 15th European Conference on Artificial Intelligence, (ECAI’02), Lyon, France, 2002, pp. 27-33.
8. K. Lee, D. Luparello, and J. Roudaire, “Automatic Construction of Personalised TV News Programs,” Proc. 7th ACM Conf. on Multimedia, Orlando, Florida, 1999, pp. 323-332.

About Authors



Mrs. K R Reshmy is working as Assistant Professor in Department of Information Technology, Sathyabama Deemed University, Chennai.She holds B.E (CSE)ME (CSE). She has contributed number of papers in seminars, conferences and journals and she is also member of professional associations.

Dr.S.K.Srivatsa is Professor in Department of ECE, MIT, Chennai. He has contributed number of papers in seminars, conferences and journals and he is also member of professional associations.



Ms. Sandhya Prasad is in Department of Information Technology, Sathyabama Deemed University, chennai.She has contributed number of papers in seminars, conferences and journals .

Digital Libraries in Knowledge Based Society : Prospects and Issues

Om Vikas

Abstract

Our information and knowledge environment has been and continues to be changed by the development of the Internet and ubiquitous communication technologies. Information and knowledge are replacing capital and energy as the primary wealth-creating assets, just as the latter two replaced land and labor many years ago. In addition, technological developments in the 20th century have transformed the majority of wealth-creating work from physically-based to "knowledge-based." Technology and knowledge are now the key factors in development of economy of the country. With increased mobility of information and the global work force, knowledge and expertise can be transported instantaneously around the world. We are now an information society in a knowledge economy where knowledge management is essential. The paper presents an overview of the role of digital libraries in the knowledge economy, its prospects and issues.

Keywords : Digital Library, Knowledge Management, Information Management

0. Knowledge Economy Steps In

Economic development began with harnessing natural resources to get more food to eat, to get higher speed to reach the goal, to preserve resources longer and to achieve better living. Industrial revolution automated a number of processes, and enticed the society for newer products and services.

Joseph Schumpeter, the economist, saw capitalism moving in long waves. Every 50 years or so technological revolution would cause "gales of creative destruction", in which old industries would be swept away and replaced by new ones. To illustrate,

1st long wave of harnessing steam power during 1780s to 1840s drove industrial revolution

2nd long wave of harnessing Railway was during 1840s to 1890s

3rd long wave of harnessing Electric power prevailed during 1890s to 1930s

4th long wave of availability of cheap oil and automobiles during 1930s to 1980s

5th wave of computing power with rapidly increasing performance-price ratio set in the Information Revolution in 1980s. If this was due to microprocessor, next technological revolution may based on nano-technology.

The societies, which participated in the process of knowledge generation, became advanced. Parity in sharing of knowledge is distancing the societies.

The process of technology adoption by the society and thereby technological transformations are speeding up. Just 4 years after its inception, the World Wide Web had 50 Million users. The number of Internet users now doubles every quarter. A quarter-century ago, it took a laboratory two months to sequence 150 nucleotides. Now, scientists can sequence 11 Million nucleotides (molecular letters that spell out a gene) in a matter of hours. Cost of DNA sequencing has also dropped from US$ 100 per base pair in 1980 to less than a penny by 2005.

ICT (Information and Communication Technology) is buzz word in modern digital economy. ICT emerges as an enabling technology to improvise productivity and quality of life. Computers process digital information very fast, communication channels provide larger bandwidth to pass on vast amount of digital information very fast. Distances shrink. Globalization sets in. Time zones promote business collaborations aiming at 24x7 hours a week operation.

Information revolution is transiting into knowledge revolution. Businesses begin to follow knowledge management practices. Knowledge based society is emerging. Knowledge is not scarce in traditional sense. The more you pass it on, the more it proliferates. It is "infinitely expansible" or "non-rival in consumption". It can be replicated cheaply and consumed over and over again. However, knowledge is more difficult to measure than traditional inputs such as steel or labor. The economist Brain Arthur argues "increasing returns of knowledge economy will magnify the market leader's advantage".

Future property of rich economies will depend both on their ability to innovate and on their ability to adjust to change.

1. Is there gain in knowledge or loss of knowledge ?

UNESCO study (1999) of 65 languages reveals that: 49 of the languages (75%) had experienced real decline in number of works translated from these languages into other languages.

The proportion for English arose from 43 percent in 1980 to over 57 percent in 1994.

The share held by top four translated languages (English, Spanish, French and German) rose from 65 percent in 1980 to 81 percent in 1994.

According to an UNESCO study involving world's 140 most published authors; 90 out of 140 were English writers in 1994 compared to 64 out of 140 in 1980.

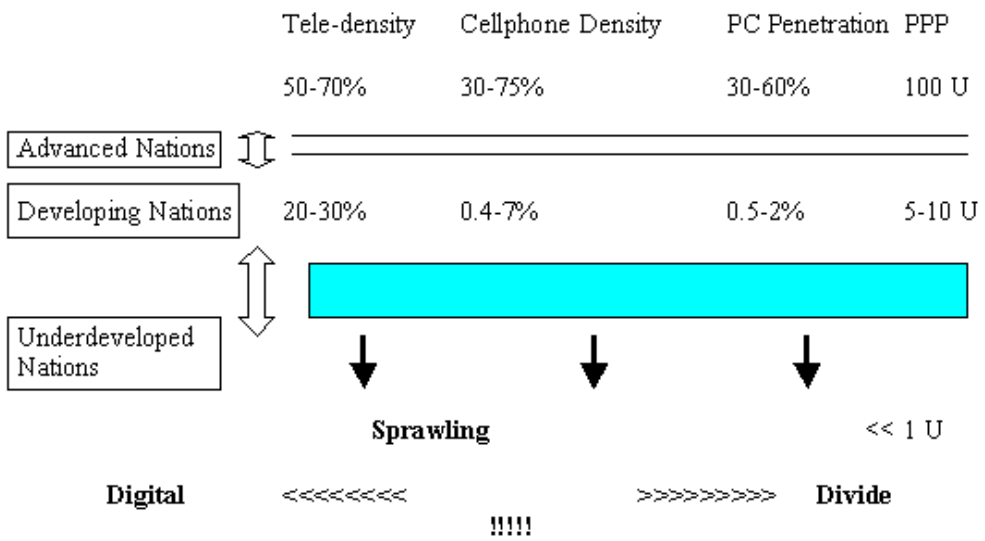
There is gradual collapse in authorship, quantity and quality of translation in other languages.

There is tendency from being creators to consumers at the time when technology could have amplified our creative capacities.'

We notice erosion of Cultures - languages - and indigenous knowledge skills.

2. World Divides Digitally

ICT Indicators and PPP (Purchasing Power Parity) are compared here below for underdeveloped, developing and advanced nations.



In comparison to advanced nations, PPP is around 10 percent for developing nations, and less than 1 percent for underdeveloped nations. For rapid penetration of ICT, PPP is key factor in evolving action plan during catch up phase of economic development. Affordable cost may be determined on this basis. For example $400 PC may be low cost PC in advanced nation, but it must cost less than $40 in developing nations. Communication technology will soon be suitable. However, computer technology may pose some problems in input & output, representation & manipulation of information in non-Roman scripts.

The price and the language processing ability will determine ICT efficacy in a local situation.

Linguistic Divide on Internet is obvious with the following statistics:

Latin Alphabet users have 39 % of the global population, and enjoy 84% of access to the Internet

Hanzi-users (in CJK) have 22% of global population, and enjoy 13% of Internet access

Arabic script users have 9% of global population, and enjoy 1.2 % of the Internet Access

Brahmi-origin scripts users in South-east Asia and Indic scripts users occupy 22 % of the world population whereas they have just 0.3 % of Internet access.

More than 65% of the content on Internet is in English.

[*according to IBM's Web Fountain analysis, 2003*]

Digital Divide as They Behold

Perception	**Developed Nations**	**Developing Nations**
Why discussed?	Desire to capture larger markets	Fear of lagging behind in economic race
Policy	Information explosion	Localization
Results	Increasing use of English and thrust of western culture.	Erosion of local languages and culture.
Consumer nature	"substitute the old"	"Upgrade the Old"
Technology development	IPR-Centric	Open source technology
Low cost PC	$400	less than $ 40
Access cost	100 U	less than 10 U
Reason: PPP : (15:1) GNP : (75:1)	 34260 (USA) 24260	 2400 (India) 460
Focus	Digital divide Access to information Wider control	Digital Unite Universalisation of creativity Share the Knowledge clustering

Low affordability means low ICT penetration and sprawling Digital Divide

3. New Order : Rise, Raise & Race

Shift from Creativity to Consumerism is alarming. This needs to be arrested for sustainable holistic development. Notion of competition should not widen gaps in society; it should rather accompany notion of cooperation to achieve objectives of Sarve bhavantu sukhinah (all be happy) and sah veeryam karvaavahai (let's strive together).

Knowledge based society will aim at universalisation of creativity. To achieve that there is need for openness of knowledge resources as well as human attitude of "Rise, Raise & Race". Raise others, and work in collaboration. Alternatively raise to rise. Time is critical factor. Race to limits of innovation.

Innovation follows on stretching our imagination to limits.

Let all the communities the world over catch up to the *basic technology* absorption capability and use it for improving quality of life of the people at large. There is need to reverse the trend from '*being consumer*' into '*becoming creator*'. This necessitates to innovatively design ICT tools to facilitate creation and access to knowledge across geographic boundaries and linguistic barriers.

Moreover, attitude needs to change. Promote collectivist culture rather than individualistic culture. Think globally but act locally to ensure relevance of technology based solution. As the real life problems become complex, and time is a critical factor, there is need to collaborate for innovation. However, scope remains

for competing for excellence. Further there is need for paradigm shift in our learning and teaching process. Learning has to be life long. Teacher acts as a facilitator, but also bears role of a guru or mentor to teach wisdom – the encapsulated knowledge that holds good across several context domains. Knowledge is contextual. The world is undergoing the turmoil of violence and terrorism. Efforts are being made at UNESCO and country levels to promote international understanding and values education for peace, human rights, democracy and sustainable development.

5. Technology Races to Human Brain

There is paradigm shift in computer processing. In the recent past, there was focus on 'data'; and R&D topics included databases and data processing. Currently focus is on' information', and R&D topics include Internet tools, content creation design of user-friendly systems (at physical level). In the near future, focus will be on knowledge aiming at wisdom. Hence R&D topics may include knowledge manipulation and development of human inspiring system at cognitive level. With the convergence of computer, communication, consumer electronics and content technologies, Information Technology makes information available at any time, at any place, in any form and on any device. Multi-lingual Multimedia technologies combine text, still pictures, moving pictures, sound animation and content in different languages. Internet brings such rich content accessible at every place. Storage, processing and retrieval of such rich content emerge as new topics for research and development. Like database management, new area of *C*ontent Management is growing.

Prof. Raj Reddy of Carnegie Mellon University predicts that after 10 years from now we shall be getting at the same cost the processing power 100 times, the storage 1000 times, and the band-width 10,000 times. ICT will be affordable, easy to use and pervasive.

Ray Kurzweil, an informatics guru, predicts that within 10 years, a 1000-dollar computer will be able to perform more than one trillion calculations a second, that well within the first quarter of the 21st century, a similarly priced computer will match the human brain.

Future Direction: Information Interspace

The Interspace represents the third wave in the ongoing evolution of the Global Information Infrastructure, driven by rapid advances in computing and Communication Technology.



The technological progress of knowledge exchange - from e-mail in Arpanet (1965-1985) to Document browsing in the Internet (1985-2000) to Concept navigation in the forthcoming Interspace (2000-2010) - has occurred in three waves, each building on the previous one.

The convergence of computing and networking is more evident in the phenomenal growth of the World Wide Web. Gorden Moore, founder of Intel corporation postulated in 1965 that the microprocessor chip would double in performance (as defined by the number of transistors on a chip) every 18 months, that is 58 percent compounded annual growth rate. Historically, the semiconductor industry has kept pace by continuously shrinking feature size to increase the number of transistors on a chip, and thus increasing the speed of the circuits

Technology roadmap for semiconductors:

Characteristic	**1997**	**1999**	**2001**	**2006**	**2012**
Process technology (nano meter)	250	180	150	100	50
No.of logic transistors (million)	11	21	40	200	1,400
Across chip Clock speed (MHz)	750	1,200	1,400	2,000	3,000

Beyond 2006, physical barriers ultimately include atomic properties that will come to fore with aggressive device shrinkage.

Metcafe's law predict the power of a network of computer (p) as square of the number of connected computers (n) [p is proportional to n^2]. Gilder (1993) predicted that the communication bandwidth will triple every year until 2020 AD. Network link throughputs are fast outstripping processor performance and memory capacities. There is increasing mismatch between fiber-optic transmission bandwidths and computer speeds, pushing computing further away from the network core. Whereas a high-end workstation today has a throughput of one gigabit per second, commercially available OC-192 Synchronous Optical Network (sonet) links operate at about 10 GIGAbit-per-second serial through put. Wave division multiplexing (WDM) optical systems can deliver aggregate throughputs of more than 200 GIGA-bits per second. Transmission bandwidth increased from 50 KBPS used by POTS (plain old telephone service) or ISDN (integrated services digital network) to 10 MBPS by Ethernet to 10 GBPS by OC-192 Sonet. As we move toward the ultrafast, fibre-optic systems found in network backbones, computing is increasingly relegated to the peripheries of the network. On the one hand, the Web's popularity and growth has been fueled largely by desktop applications consuming bandwidth intensive images and videos. On the other hand, thin-client computers are becoming more commonly used as edge-of-network devices, often connected by wireless technology.

There is increasingly shift from Operating System to processor to network to storage. Storage is increasingly strategic to businesses. Information centric computing include Operations such as Find, Create, Store, Retrieve, Manage. Data is more valuable than processing. Internet provides new challenges for storage: A4 data accesses (Anywhere, Anytime, Anyone, Any device); 24x7x365 hours uptime dynamic scalability; lower costs, independence from legacy systems. Areal density on magnetic hard disk drives have advanced 2 million times since the first disk drive by IBM in 1957. DVD (Digital Video/ Versatile disc) can store up to 17 billion bytes of data on 4.75 inch platter. Areal density for DVD-type products is targeted to 50 Gb/ in^2 for multimedia applications. This may further be pushed to exceed 100 Gb/ in^2 using e-beam lithography micor-fabrication techniques. Optical storage techniques are reported to provide terra bits/ in2 areal density. Current storage media can be classified into 3 classes : magnetic, optic and solid state. A relatively new approach to information management known as the SAN (Storage Area Network) provides high-speed any-to-any interconnection of servers and storage elements. Solid-state storage technology is approaching the density and cost of magnetic mass storage. FLASH memory is now replacing hard disks in some applications and may replace floppy disks.

5. Digital Library Brings Knowledge at Door Steps

Notion of digital library include electronic ("digital") storage of materials. Newby categorizes into major approaches of data store, electronic access to traditional library material, and scholarly archives.

Data store focuses on digitization, indexing & retrieving and standards for data organization. This is more dominated by DL researchers' view. Electronic access to traditional materials are geared more towards general public, whereas data store is for specialized user groups. Scholarly archives bypass publishers for quick, ready and equitable access to scholarly works. However editorial efficiency is necessary for maintaining good quality control.

Book costs money. Production cost of a book is about 20% of total cost. One model for use of a digital book may be "pay as you go". Publishers would favor that. But a library follows "buy once, use as many" model.

There are technologies for restricted viewing, restricted reproduction and retransmission. Legal copyright restrictions need to be evolved to prevent piracy. The third model "scholar as publisher" need to be evolved. This is somewhat like open source. There had been a number of open access initiatives declaring international policy on open access.

Timeline of International Policy on Open Access:

February 14, 2002 Budapest Open Access Initiative

December 17, 2002 Howard Hughes Medical Institute makes commitment to cover open-access publication fees for its own researchers

April 11, 2003 Bethesda Meeting on Open Access Publishing

October, 1 2003 The Wellcome Trust position statement in support of open-access publishing

October 22, 2003 Berlin Declaration on Open Access to Knowledge in the Sciences & Humanities endorse open access, encourage scientists to publish open-access papers

December 5, 2003 JISC announces funding to help publishers transition to open-access

December 12, 2003 UN WSIS Declaration of Principles includes support for open access initiatives

January 30, 2004 OECD Committee for Scientific and Technological Policy adopts Declaration on Access to Research Data from Public Funding

February 24, 2004 IFLA Governing Board adopts Statement on Open Access to Scholarly Literature and Research Documentation

Open Access Initiatives :

Budapest Open Access Initiative

(BOAI www.soros.org/openaccess)

Recommends and supports two strategies to get to open access: (1) Self-archiving, (2) New open access journals

Scholarly Publishing and Academic Resources Coalition

(SPARC www.arl.org/sparc/)

Promotes fundamental changes in scholarly publishing. Offers practical support to initiatives that bring down the cost of scholarly publishing

Public Library of Science

(PLoS www.publiclibraryofscience.org/)

Calls for scientists to pledge only to publish in, edit, review for, subscribe to, these journals that are making research material available in open access within six months of publication

6. Managing Information in Distributed Digital Library

Public awareness of the Internet as a critical infrastructure in the 1990s has spurred a new revolution in technologies for information retrieval in digital libraries. Many believe we are now at the start of the Net Millennium, a time when the Net forms the basic infrastructure of everyday life. Collections of all kinds must be indexed effectively, from small communities to large disciplines, from formal to informal communications, from text to image and video repositories, and eventually across languages and cultures. The Net needs new technology to support this new search and indexing functionality.

Digital library is a form of information technology in which social impact matters as much as technological advancements. The best way to develop effective new technology is by undertaking multi-year large-scale research projects that develop real-world electronic test beds used by actual users and by aiming at developing new, comprehensive, and user-friendly technologies for digital libraries.

DARPA's Information Management program (www.dapra.mil/ito/research/in) address core digital library issues requiring revolutionary research in technology. These include:

- Federated repositories. The organisation of distributed repositories into a coherent virtual collection is fundamental
- Scalability. Managing billions of digital objects and millions of sources poses challenges in identifying, categorizing, indexing, summarizing and extracting content.
- Interoperability. Digital libraries require semantic interoperability among heterogeneous repositories distributed across the network.
- Collaboration. Analysts work in distributed teams, building on each other's knowledge experience and resources.
- Communication. Timely dissemination of research results is the focus of D-Lib.

Problems generic to digital libraries for any domain include behavior and cognition issues, lack of standards, legacy systems, distributed data, the need to network among heterogeneous systems, inefficient information retrieval and privacy concerns.

The Illinois DLI project (http://dli.grainger.uiuc.edu) chose as its research paradigm and complete manipulation of structured documents-namely, the search and display of engineering journal articles

encoded in Standard Generalized Markup Language (SGML). The project aimed at developing and experimentally testing new technology for federated search by deploying real collections to real users on a production basis.

The Illinois D-Lib take SGML directly from the publisher's collections, converting it into a canonical format for federated searching and transforming tags into a standard set. The coming widespread availability of rich markup formats, such as XML (eXtensible Markup Language) - nearly complete instance of SGML will likely make such formats the standard for open document systems.

UDL project at CMU identifies the research challenges concerning:

Input : low cost scanning, formats conversion, color representation, graphics file formats, archiving;

OCRs for Indian languages yet to mature; structured matter such as musical notation, chemistry, 3D items, web documents

Navigation : keyword searching does not scale, browsing, finding, searching, flying, zooming, view whole collection or one glyph; Fractal view – granularity and connectivity of keys, Hyperbolic trees, virtual reality, discovered similarities, user defined catalogues, searching mathematical expression.

∞

$\int_0$ $[e^{-x^2} \sin x^2]\, dx$ may be represented as

Integrate [Times [Power [e, Times [-1, Power [V_1, 2]]] Times [sin [power [V_1, 2]]],

{ V_1, 0, infinity }]

Multilingual issues: character sets (Unicode, ISCII), Multilingual navigation, translation assistance.

Synthetic Documents: derived automatically from retrieved information via intelligent agents, abstracts, summaries, glossaries, translations, critical reviews, encyclopedia-on-demand.

Aboutness is central to cataloging and retrieval. Suppose a topic T is subset of W (all words/book). P is about the topic T if P is subset of W, and P<intersection> T = <nonempty subset>.

Thesaurus is topic-hierarchical with numbered entries. Thesaurus + aboutness hierarchy can be used to disambiguate meanings without "understanding". Topic numbers are language independent.

Improving Web searching beyond full-text retrieval requires using document structure in the short term and document semantics in the long term. Interspace, the future Internet, is developed where each community indexes its own repository of its own knowledge. Information infrastructure must provide substantial support to community amateurs for semantic indexing and retrieval. Interspace focuses on scalable technologies for semantic indexing that work generally across all subject domains. We can use concept spaces - collections of abstract concept generated from concrete objects-to boost searches by interactively suggesting alternative terms. We can use category maps to boost navigation by interactively browsing clusters of related documents. Scalable semantics is used to index the semantics of document contents on large collections. These algorithms rely on Statistical techniques, which correlate the context of phrases

within the documents. Concept spaces use text documents as the objects and noun phrases as the concepts.

The Interspace consists of multiple spaces at the category, concept, and object levels. Within the course of an interaction session, users will move across different spaces at different levels of abstraction and across different subject domains. Such a fluid flow across levels and subjects supports semantic interoperability. Interspace navigation enables location of documents with specific concepts without previous knowledge of the terms within the documents.

Federating the search at a semantic level is an area of active research in digital library community. Statistical approaches lead toward scalable semantics - indexing deeper than text word search that is computable on large real collections. Concept spaces for semantic retrieval which capture contextual information, have been computed for collections of millions of documents.

It is necessary to evolve metadata standard and interoperability framework. Metadata Encoding and Transmission Standard (METS) schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library expressed using the XML schema language of the World Wide Web Consortium (W3C). The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation.

The Open Archives Initiative Protocol for Metadata Harvesting provides an application-independent interoperability framework based on metadata harvesting. There are two classes of participants in the OAI-PMH framework: Data providers and service providers.

In the 21st century, there will be a billion repositories distributed over the world, where each community maintains a collection of they own knowledge. Semantic indexes will be available for each repository, using scalable semantics to generate search aids for the specialised terminology of each community. Concept members of one community to easily search the specialized terminology of another. Information analysis will become a routine operation in the Net, performed on a daily basis worldwide.

Future knowledge networks will rely on scalable semantics, on automatically indexing the community collections so that users can effectively search within the Interspace of a billion of repositories. Just as the transmission networks of the Internet are connected via switching machines that switch packets, the knowledge networks of the Interspace will be connected via switching machines that switch concepts. Connectivity and training continue to be the principal barriers to integrating the global network of libraries.

7. Digital Library Initiatives

Six major projects were launched during 1994-1998 under DLI (Digital Library Initiative) funded by the NSF, DARPA and NASA in the USA.

Digital Libraries Initiative-phase 2 (DLI-2) is an NSF led initiative that builds on the successes of DLI-1. DLI-2 is supported by many funding agencies like NSF, DARPA, National Library of Medicine, Library of congress National Endowment for the Humanities. DLI-2 will investigate digital libraries as human-centered systems.

JSTOR (Journal Storage) project started at University of Michigan with the grant of the Andrew W Mellon Foundation. JSTOR database total 450,000 articles and 2.7 million pages created via a combination of page images and full-text scanned-in files, the database is growing at a rate of 100,000 pages per month. JSTOR serves more than 350 academic institutions around the world. JSTOR should be usable by any

browser that supports HTML 3.2 standard. The JSTOR (Journal Storage) project was intended to become a commercial service. The chose the mature technology of digitized bitmaps (page images) rather than the immature technology of SGML markup. The www.jstor.org URL links to three server machines: two at University of Michigan, a third at Princeton University. Distributed mirrors offer increased reliability, accessibility, and capacity. The round robin feature of DNS (Domain Name Service) provides a single Web service from multiple locations.

The Informedia Project at Carnegie Mellon University has created a terabyte digital video library in which automatically derived descriptors for the video are used for indexing, segmenting, and accessing the library contents. Artificial Intelligence techniques have been used to create metadata - the data that describes video content. Powerful browsing capabilities are essential in a multimedia information retrieval system because the underlying speech, image and language processing are imperfect and produce ambiguous incomplete metadata.

The Carnegic Mellon DLI project searched multimedia, particularly video segments, by generating text indexes using speech understanding. The Stanford DLI project searched across different engines using multi-protocol gateways. Other even harder issues remain untouched, such as multicultural search across context and meaning.

The importance of D-Lib research is spreading beyond the US. European research in Digital Libraries is funded by the European Union as well as national sources. DL projects have supported by the Information Engineering, (www.echo.lu/ie), Language Engineering (www.echo.lu/langeng/en/lehome.html), and Esprit (www.cordis.lu/esprit) programs in Europe. Under NSF-EU collaboration, five working groups has been formed in the key technical areas of Interoperability, Metadata, IPR, Resource indexing and discovery, and multilingual information access.

Since 1995, D-Lib research has become a national grand challenge in several countries in Asia. Most projects can be classified into the following categories:

- Nationwide D-Lib initiative and special purpose digital libraries-for example, the library 2000 Project in Singapore (to link all library resources) and Financial Digital Library at the University of Hong Kong (to serve the needs of HK stock market and users)
- Digital museum and historical document digitalization-fox example, Digital Museum Project of the National Taiwan University and Digitalization of art collection of the Palace Museum in Taipai by IBM.
- Local language and multilingual information retrieval-for example, the Net Compass Project of Tsinghua University in China, Chinese Information Retrieval at the Academia Sinica, Taiwan, and New Zealand's multilingual project.

Local language processing and historical cultural content could be the most immediate Asian contribution to the international DL community. An Asia Digital Library consortium is fostering long-term collaboration and projects in DL-related topics in Asia (www.cyberlib.net/adl).

The New Zealand D-Lib (http://www.nzdl.org) currently offers about 20 collections, varying in size from a few documents upto 10 million documents and several gigabytes of text. The documents written in many different languages, including English, French, German, Arabic, Maori, Portugese and Swahili. The D-Lib provides interfaces to the collections in several languages. To accommodate blind users (with speech synthesizers) and partially sighted users (with large-font displays), NZ D-Lib provides text only version of the interface for each language.

Design is based on collections-set of like documents. The documents come in a variety of formats: plain ASCII, Post Script, PDF, HTML, SGML and Microsoft Word for textual documents. Collections invariably undergo a building process to make them suitable for search, retrieval, and display.

Managing the complexity of multiple collection, multiple languages, and multiple interface options presents a significant challenge. For example, document items that have not yet been translated to other languages need to default to English. Non_ASCII languages like Arabic and Chinese need special text positioning and justification.

Digital Library projects were initiated by the Department of Scientific & Industrial Research (DSIR), the Department of Information Technology (DIT) and the Department of Culture (DoC). DSIR funded project on Digital Library of Traditional Heritage knowledge; DIT launched Digital Library of India initiative; Department of Culture support DL activities at Indira Gandhi National Center for Arts, launched a comprehensive National Mission for Digital Libraries that synergizes with other mission such as National Mission for Intangible Cultural Heritage (ICH) and National Mission on Manuscripts.

DLI (Digital Library of India) Initiative was launched in September 2003 by President of India. DLI portal (http://www.dli.ernet.in) is operational. By mid 2004, 84000 book (~2.8 million pages) were scanned and cropped in various languages, viz English, Telugu, Tamil, Sanskrit, Kannada, Hindi. There are 4 regional mega centers and 20 scanning centers. The mega centers are responsible for content development of around 14 million pages resulting into a total of 56 million pages and scanning centers would contribute about 15 million pages. Hence 250,000 books are targeted. The mega centers will develop requisite access technologies such as Cross-Lingual Information Access, Multilingual Crawler, OCR with workflow, Multimedia Interface for physically challenged, Automatic Search Indexing tools, Multilingual and multi-modal authoring tools, Text summarisation with focus on nine languages to begin with, Hindi, Marathi, Punjabi, Bengali, Assamese, Sanskrit, Telugu, Kannada and Malyalam. DLI is being implemented in close collaboration with UDL (Universal Digital Library) project (http://www.ulib.org) at Carnegie Mellon University. Overall coordinator of UDL project is Prof. Raj Reddy at CMU, whereas Prof. N Balakakrishnan is coordinator of the India nodal center at Indian Institute of Science, Bangalore.

Heavy duty scanners, Minolta PS7000, have been provided to the scanning & mega centers under the UDL project. Over 100 scanners are operational in India by mid 2004. Along with scanner, Abby Fine Reader 6.0 and Scanfix software have also been provided. China preferred to use portable flatbed scanner AVA3+ of Sharp Corpn.. UDL aims at digitizing 1 million books which are only 1% of all books available in the world. There is good scope of research in the domains of Universal access, Design of distributed cached servers, multilingual information retrieval, Machine Translation and Summarization technologies. Simultaneously, efforts need to be renewed towards improvement of OCR technology for Indian languages. About 5% books are out of copyright; 92% of the books are out of print but they are under copyright, and 3% of the books are in print and copyrighted. Selection of books may follow Gresham's law that is, convenience displaces quality. Present focus is on evolving metadata standards and Standard Operating Procedures (SOP), improvising OCR in Indian languages, developing Indian language search tools, design system architecture taking into account the storage bandwidth, and connectivity requirements and the web-services. Other policy & management issues of copyright, classification of resources, duplication of content, delivery & web-services also need immediate attention.

10,000 pages may be scanned per scanner per day in 3 shifts. Images are stored in TIFF (Tagged Image File Format), OCRed text is stored in HTML, TXT, RTF & JPAG formats for searching purpose. Metadata is in XML (Dublin Core) format scanned. For a book of 500 pages, image is 50-150 KB, RTF/HTML text file 8-15 KB, average size of digitized book is about 60 MB. Formats for audio are WAV, MP3, RA. Formats for video are MPEG-1, MPEG-2, MPET-4, AVI, QT, H.263.

Research challenges include Input (scanning, digitizing, OCR), Metadata creation, Data representation, Navigation and search, Multilingual issues, Output (voice, pictures, virtual reality.

8. Meta Data for Efficient Accessibility

The Web's creator Tim Berners-Lee considers the Web not to be the technology but connection of all things enabled by it. Issues of irrelevant search results spelling mistakes during search necessitate standardization of metadata that is descriptive information about the web resources. This may be added to the web page during the coding of the web page or afterward. Metadata do not appear in document display and do not affect the browser's display at all; however it provides lot of useful information to web-robots and search engines about the web pages.

Mainly there are 3 standards of Digital Library: Dublin Core Standards, OCLC Standards & Information Retrieval Standard.

Dublin Core Metadata Initiative began in 1995 to develop conventions for resource discovery on the World Wide Web. DC Metadata set is about semantics of 16 core data elements. The simplicity of creation & maintenance, commonly understood semantics, International Scope and Extensibility are the underlying goals of DC Metadata set.

8.1 Dublin Core Standards

The Dublin Core metadata element set is a standard for cross-domain information resource description. Here an information resource is defined to be "anything that has identity". This is the definition used in Internet RFC 2396, "Uniform Resource Identifiers (URI): Generic Syntax", by Tim Berners-Lee et al. There are no fundamental restrictions to the types of resources to which Dublin Core metadata can be assigned.

DC is based on the principle that each data element is optional, repeatable and may have any field legth.

i. Simple Dublin Core Standard : 15 data elements and those are expressed as "attribute-value" pairs, without using quantifiers.

The Elements of Dublin Core:

Element Name	:	Title
Definition	:	A name given to the resource.
Comment	:	Typically, Title will be a name by which the resource is formally known.
Element Name	:	Creator
Definition	:	An entity primarily responsible for making the content of the resource.
Comment	:	Examples of Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.
Element Name	:	Subject/ Keywords
Definition	:	A topic of the content of the resource.
Comment	:	Typically, Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.

Element Name	:	Description
Definition	:	An account of the content of the resource.
Comment	:	Examples of Description include, but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.
Element Name	:	Publisher
Definition	:	An entity responsible for making the resource available
Comment	:	Examples of Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity.
Element Name	:	Contributor
Definition	:	An entity responsible for making contributions to the content of the resource.
Comment	:	Examples of Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.
Element Name	:	Date
Definition	:	A date of an event in the lifecycle of the resource.
Comment	:	Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [W3CDTF] and includes (among others) dates of the form YYYY-MM-DD.
Element Name	:	Resource Type
Definition	:	The nature or genre of the content of the resource.
Comment	:	Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the DCMI Type Vocabulary [DCT1]). To describe the physical or digital manifestation of the resource, use the FORMAT element.
Element Name	:	Format
Definition	:	The physical or digital manifestation of the resource.
Comment	:	Typically, Format may include the media-type or dimensions of the resource. Format may be used to identify the software, hardware, or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types [MIME] defining computer media formats).
Element Name	:	Resource Identifier
Definition	:	An unambiguous reference to the resource within a given context.
Comment	:	Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. Formal identification systems include but are not limited to the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL)), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN).
Element Name	:	Source
Definition	:	A Reference to a resource from which the present resource is derived.

Comment	:	The present resource may be derived from the Source resource in whole or in part. Recommended best practice is to identify the referenced resource by means of a string or number conforming to a formal identification system.
Element Name	:	Language
Definition	:	A language of the intellectual content of the resource.
Comment	:	Recommended best practice is to use RFC 3066 [RFC3066] which, in conjunction with ISO639 [ISO639]), defines two- and three-letter primary language tags with optional subtags. Examples include “en” or “eng” for English, “akk” for Akkadian”, and “en-GB” for English used in the United Kingdom.
Element Name	:	Relation
Definition	:	A reference to a related resource.
Comment	:	Recommended best practice is to identify the referenced resource by means of a string or number conforming to a formal identification system.
Element Name	:	Coverage
Definition	:	The extent or scope of the content of the resource.
Comment	:	Typically, Coverage will include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names [TGN]) and to use, where appropriate, named places or time periods in preference to numeric identifiers such as sets of coordinates or date ranges.
Element Name	:	Rights Management
Definition	:	Information about rights held in and over the resource.
Comment	:	Typically, Rights will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights. If the Rights element is absent, no assumptions may be made about any rights held in or over the resource.

Qualified Dublin Core Standard includes an addition element, Audience, as well as a set of element quantifiers to further refine meaning and scope of the data element.

The core data element may be grouped under three broad categories:

Content 8 data elements: Coverage, Description, Type, Relation, Source, Subject, Title and Audience

Intellectual Property: 4 data elements: Date, Format, Identifier, and Language

There are two types of quantifiers for the above core data elements and these are specified as sub-fields

Element Refinement to specify meaning

Element Encoding Schemes to specify the encoding scheme used

For example, Date element qualifier sub-field may specify data created/issued/ modified/copyrighted/ submitted.

Date element encoding schemes sub-field may include DCMI, Period, W3C-DTF ISO-860 format (YYYY-MM-DD)

8.2 OCLC Standards

Founded in 1967, OCLC Online Computer Library Center is a nonprofit, membership, computer library service and research organization dedicated to the public purposes of furthering access to the world's information and reducing information costs. More than 50,540 libraries in 84 countries and territories around the world use OCLC services to locate, acquire, catalog, lend and preserve library materials.

8.3 Information Retrieval Standard

This standard was processed and approved for submittal to ANSI by the National Information Standards Organization. It was balloted by the NISO Voting Members March 29, 2002 - May 13, 2002. It will next be reviewed in 2007. Suggestions for improving this standard are welcome. They should be sent to the National Information Standards Organization, 4733 Bethesda Avenue, Suite 300, Bethesda, MD 20814. NISO approval of this standard does not necessarily imply that all Voting Members voted for its approval.

Metadata draft standards (based on Dublin core) for DLI (Digital Library of India) Initiative

Field	**Details**
Language	[Ass/Ben/Eng /Guj/Hin/Kan/ Mal/Mar/Ori/Pun/Tam/Sans/Tel/Urd...]
Title	_____
Creator/Author	_____
Keyword Description	_____
Subject	[General/Philosophy, Psychology/ Relogion, Theology/Social Sciences/ Natural Sciences...]
Publisher	_____
Contributor	_____
Date	_____
Document Type	[Art Objects/Fabrics/Floppies/Glass/ Magnetic Tapes/Microfilm/Palm Leaf/Paper/wood...]
Format	[TIFF…]
Identifier	_____
Source	_____
Relation	_____
Coverage	_____
Rights	[Copyright Permitted/In Public Domain/Not Available]
Copyright Date	_____
Scanning Centre	[IISc.B/Central Library.Hyd/SASTRA/MIDC/IIIT Allahabad/SV Digital Library Tirupati/ CDAC.N/…]
Scanner Number	_____
Digital Republisher	[Digital Library of India]
Digital Publication Date	_____

9. Digital Library Framework for Developing Nations

International Conference on Digital Libraries (ICDL) held in February 2004 concluded with recommendations concerning *Content, Technology, Users* and *Policy & management* issues. There is rich heritage knowledge that may be put to web. Technologies for scanning, indexing, security, access & delivery in multilingual environment need to be developed. Metadata and delivery standards need to be evolved and finalised. Types of users and their requirements need to be identified. Sub-group on Policy & Management - comprising of Michael Seadle, OmVikas, Harsha Parekh - deliberated issues concerning duration of copyright, online registry, copy left provision, compulsory licensing, and ethics in digital world.

India needs practical, affordable, and immediate access to scholarly and research information in order to bridge the digital divide that separate rich and poor countries, and the rich and the poor within countries. The quantity of all forms of information, scholarly as well as commercial, is increasing rapidly. Existing copyright laws within member countries of the Berne Convention lock that information for the life of the author plus a number of years (60 in India), and make no distinction between the information type and intent.

An optional end to copyright protection after 5 or even 10 years would free a large amount of academic scholarship without affecting the rights of commercially valuable works. India could implement copyright in it's own laws.

The principle of "copy left" is that the rights holder should have the right to choose not to continue copyright protection in a standard, legally binding, and recently registered way.

Automatic licensing does not end protection for a copyrighted work, but enables its widespread use through predictable, low cost-per-use charges. Automatic licensing would open decades worth of past works to safe and affordable public use. Automatic licensing should be implemented for scholarly and research materials at home.

Translations are derivative works that require permission from the original right holder. For multilingual searching and multilingual societies, translation is an important enabling tool and should provide automatic licensing for translations. India may recommend to Berne that non-commercial translation involving minimal human effort or creativity is exempt from copyright protection.

While compulsory deposit for paper materials is well established, digital depository requirements exist in only a few countries and are not systematically enforced. Registration and deposit for any materials - digital or analog - will get long term copyright protection. This would: a) ensure that materials would be available to national libraries; b) assist in establishing the authenticity of copies by comparison with a trusted repository; c) could provide information about the ownership of a work, and d) assist in attempting to preserve intellectual property for future generations by freeing them at least from the burden of finding copies.

The ICDL 2004 recommendations on Policy and Management may be summarized as below:

1. Online Registry - Every digital material produced in this country should be registered with the Digital-Object- Identifier (DOI).

2. A Depository should be created for our heritage as well. This depository would provide authenticity and ownership as well as will enable preservation of intellectual property rights.

3. Copy-left Provision This provision will enable a copyright holder to give up the rights before or after a certain period. Provision for this should be kept open.

4. Copyrights lock-in period be reduced to 25

5. Compulsory Licensing has been well tested for cable TV applications. Similarly, compulsory licensing is also recommended for the Digital India

6. Moral Rights and ethics in the Digital World It is, necessary to make sure that those who deal with digital documents are of elite character with strong ethics so as to ensure that nobody can manipulate the information and claim the credibility for the work as his/or her.

10. Conclusion

ICT enables access to digital information to anyone, anywhere, anytime, any device. Knowledge resources of various communities are available in various forms – print, manuscripts, sculptures, drawing etc. on various media. Creativity of people enriches civilisation with innovative products, services and solutions to real life problems as well as arts and culture.

Digital Library transforms creative material in electronic form that is virtual copy that can be automatically searched and retrieved, anywhere, anytime by anyone with some constraints. It would otherwise have been impossible for many to physically see such a piece of creative work at some far off place. New creative works may also be added under the class that is yet to be reviewed and undergo editorial quality control. Feedback on this work will further stimulate imagination of the creator. This process must spread from one to many to all. *Universalisation of creativity* will make all the communities vibrant with innovative aptitude and ability to adjust to change. They will retain their traditional values and participate in the new culture of cooperation: rise, raise & race.

Basic DL technologies should be made available to developing countries at affordable price. These countries may adopt some underdeveloped countries to bring them up under the umbrella of some UN agency. Village Knowledge Centers (Gaon Gyan Kendras) may be set up to bring up rural masses. Open Access initiatives need to be encouraged. Advanced nations should focus on developing futuristic knowledge networking technologies, and assisting in spreading connectivity and organizing training programs.

11. References

1. ACM-IEEE Joint Conference on Digital Libraries, Rice University, Houston, Texas 27-31 May, 2003.
2. Proceedings of International Conference on Digital Libraries, ICDL 2004. The Energy & Resources Institute (TERI), New Delhi, 24-27 February, 2004
3. Michael Seadle, Om Vikas & Harsha Parekh, Report of the ICDL' 2004 subgroup on Policy and Management, February 24-27, 2004, New Delhi
4. R K Mishra, "The Dublin Core Metadata Set for HTML 4.0: a format to map web resources", International Conference on Digital Libraries: ICDL-2004, February 24-27, 2004, New Delhi
5. B. Schatz & H. Chen, "Digital Libraries: Technological Advances and Social Impacts", IEEE Computer, February 1999 pp 45-50.
6. B.Schalz, et.al., "Federated Search of Scientific Literature", IEEE Computer, February 1999, pp 51-58.

7. S W Thomas, K Alexander & K Guthrie, “Technology Choices for the JSTOR Online Archive”, IEEE Computer, February 1999, pp 60-65.

8. H D Wactlar, M G Christel, Y Gong & A G Hauptmann, “Lessons Learned from Building a Terra-byte Digital Video Library”, IEEE Computer, February 1999, pp66-73.

9. I H Witten, R J Mc Nab, S Jones, M Apperley, D Bainbridge & S J Cunningham, “Managing Complexity in a Distributed Digital Library”, IEEE Computer, February 1999, pp74-79.

10. Gregory B Newby, “Digital library Models and Prospects” ASIS Mid year 1996 meeting

11. Raj Reddy, “Information Technology and Digital Libraries”, Meeting on Universal Digital Library (UDL) project, at CMU, 26-30 May, 2002 (also reprinted in VishwaBharat@tdil, July 2002 (ISSN No.0972-6454), pp8-13).

12. Development Dialogue, 1999: 1-2, Dag Hamarskjold Center

13. World Culture Diversity, UNESCO, 1995

14. World Culture Report – Culture, Creativity and Markets, UNESCO, 1998, published by Department of Information Technology, Government of India, New Delhi

15. VishwaBharat@tdil, Language Technology Flash (Quarterly), Year 2000, 2001, 2002, 2003, 2004.

16. TDIL Website: http://tdil.mit.gov.in

17. DLI Website: http://www.dli.ernet.in

18. Metadata standards : http://dublincore.org/documents/1999/07/02/dces, http://www.loc.gov/z3950/agency/document.html

About Author



Dr. Om Vikas is the Senior Director and Head of the Human Centered Computing Division in the Ministry of Communications & Information Technology, Government of India. He holds B.Tech.(EE), M.Tech.(EE) and Ph.D from IIT, Kanpur. He has vast experience of R&D, Teaching, Projects planning and International cooperation in industry, academia and government. He has been active member of several regional and international conference committees such as processing of Asian Languages, Object Oriented Languages and Systems, Thesaurus Modeled Sanskrit Database (Univ. of Texas), High Performance Computing, Speech & Language Technology, NLP & KBCS. He is on several inter-ministerial committees. Dr Vikas is also national coordinator of the mission program on Technology Development for Indian Languages (TDIL) as well as the Digital Library of India initiative. He represented India and actively participated in the UNESCO Experts meetings on Multi-lingualism and Universal Access to Cyberspace, in Paris. He is Senior member of IEEE and Fellow of IETE. Fellow of Russian Academy of Informatization of Education, Senior Member of Computer Society of India, and IE (India), and also member of IEEE_Computer Society & IEEE_Engineering Management Society. He has several research papers, articles in conferences, and techno-economic analysis reports as well as a patent on encrypt & decryption. He is editor of the quarterly publication - VishwaBharat@tdil - on language technology in India for last four years. For his outstanding contributions in the field of ICT for masses, he received several awards such as “Vishisht Padak”, “Indira Gandhi Rajbhasha”, “Atmaram” & “Vigyan Bhushan”, and recently “VASVIK Industrial Research” Awards. His current research interests include Computer architecture, Data Design, Natural Language Processing, Knowledge Management and Informatics curriculum development.

E-mail : omvikas@mit.gov.in

Mining of Confidence-Closed Correlated Patterns Efficiently

R Hemalatha A Krishnan C Senthamarai R Hemamalini

Abstract

Correlated pattern mining has become increasingly important recently as an alternative or an augmentation of association rule mining. Though correlated pattern mining discloses the correlation relationships among data objects and reduces significantly the number of patterns produced by the association mining, it still generates quite a large number of patterns. This paper proposes closed correlated pattern mining to reduce the number of the correlated patterns produced without information loss. A new notion of the confidence-closed correlated patterns is proposed first, and then an efficient algorithm is present, called CCMine, for mining those patterns. Confidence closed pattern mining reduces the number of patterns by at least an order of magnitude. It also shows that CCMine outperforms a simple method making use of the traditional closed pattern miner. Confidence-closed pattern mining is a valuable approach to condensing correlated patterns.

Keywords : Data Mining, CC Mine, Database Systems.

0. Introduction

Association mining often generates a huge number of rules, but a majority of them either are redundant or do not reflect the true correlation relationship among data objects. To overcome this difficulty, interesting pattern mining has become increasingly important recently and many alternative interestingness measures have been proposed [1,2,3,415,16,17,18]. While there is still no universally accepted best measure for judging interesting patterns, all confidence is emerging as a measure that can disclose true correlation relationships among data objects [5,6,7,8].

One of important properties of all confidence is that it is not influenced by the co-absence of object pairs in the transactions—such an important property is called null-invariance [8]. The co-absence of a set of objects, which is normal in large databases, may have unexpected impact on the computation of many correlation measures. All confidence can disclose genuine correlation relationships without being influenced by object co-absence in a database while many other measures cannot. In addition, all confidence mining can be performed efficiently using its downward closure property [5].

Although the all confidence measure reduces significantly the number of patterns mined, it still generates quite a large number of patterns, some of which are redundant. This is because mining a long pattern may generate an exponential number of sub-patterns due to the downward closure property of the measure. For frequent itemset mining, there have been several studies proposed to reduce the number of items mined, including mining closed [9], maximal [10], and compressed (approximate) [11] itemsets. Among them, the closed itemset mining, which mines only those frequent itemsets having no proper superset with the same support, limits the number of patterns produced without information loss. It has been shown in [12] that the closed itemset mining generates orders of magnitude smaller result set than frequent itemset mining.

This paper introduces the concept of confidence closed correlated pattern, which plays the role of reducing the number of the correlated patterns produced without information loss. All confidence is used here is

correlation measure. However, the result can be easily extended to several other correlation measures, such as coherence [6]. First, propose the notion of the confidence-closed correlated pattern.

Previously used concept is support-closed pattern, i.e., the closed pattern based on the notion of support. However, support-closed pattern mining fails to distinguish the patterns with different confidence values. In order to overcome this difficulty, this paper introduces confidence-closed correlated pattern, which encompasses both confidence and support. Then an efficient algorithm is proposed, called CCMine, for mining confidence-closed patterns. The experimental and performance study shows that confidence-closed pattern mining reduces the number of patterns by at least an order of magnitude. It also shows the superiority of the proposed algorithm over a simple method that mines the confidence-closed patterns using the patterns generated by the support-closed pattern miner.

1. Background

Let I = {i_1, i_2, . . . , i_m} be a set of items, and DB be a database that consists of a set of transactions. Each transaction T consists of a set of items such that T Í I. Each transaction is associated with an identifier, called TID. Let A be a set of items, referred to as an itemset. An itemset that contains k items is a k-itemset. A transaction T is said to contain A if and only if A Í T. The support of an itemset X in DB, denoted as sup(X), is the number of transactions in DB containing X. An itemset X is frequent if it occurs no less frequent than a user-defined minimum support threshold[15,16,118,17).

Generally in data mining, only the frequent itemsets are considered as significant and will be mined. The all confidence of an itemset X is the minimal confidence among the set of association rules i_j®X ® i_j, where i_j ® X. Its formal definition is given as follows. Here, the max_item_sup of an itemset X means the maximum (single) item support in DB of all the items in X.

Definition 1: All-confidence of an itemset

Given an itemset X = {i1, i2, . . . , ik}, the all confidence of X is defined as,

$$\text{Max_item_sup}(X) = \max\{\text{sup}(i_j) | "i_j \hat{I} X\} \quad (1)$$

$$\text{All_conf}(X) = \frac{\text{sup}(X)}{\text{max item sup}(X)} \quad (2)$$

Given a transaction database DB, a minimum support threshold min-_sup and a minimum all_confidence threshold min_a, a frequent itemset X is all_confidence(x) or correlated if all conf(X) ³ min_a and sup(X) ³ min_sup.

2. Confidence Closed Correlated Patterns

It is well known that closed pattern mining has served as an effective method to reduce the number of patterns produced without information loss in frequent itemset mining. Motivated by such practice, extend the notion of closed pattern so that it can be used in the domain of correlated pattern mining. The formal definitions of the original and extended ones are in Definitions 2 and 3, respectively. The former is called as support-closed and the latter is called as confidence-closed.

Definition 2. Support-Closed Itemset : An itemset Y is a support-closed (correlated) itemset if it is frequent and correlated and there exists no proper superset Y' É Y such that sup(Y') = sup(Y).

Since the support-closed itemset is based on support, it cannot retain the confidence information—notice that the confidence means the value of all confidence. In other words, support-closed causes information loss.

Example 1

Let itemset ABCDE be a correlated pattern with support 30% and confidence 30% and itemset CDE be one with support 30 and confidence 80%. How to get a set of non-redundant correlated patterns when min_sup = 20 and min_a = 20% ? Support-closed pattern mining generates ABCDE only eliminating CDE since ABCDE is superset of CDE with the same support. Thus lose the pattern CDE. However, CDE might be more interesting than ABCDE since the former has higher confidence that the latter. Thus extend the support-closed itemset to encompass the confidence so that it can retain the confidence information as well as support information.

Definition 3. Confidence-Closed Itemset : An itemset Y is a confidence-closed itemset if it is correlated and there exists no proper superset Y'ÉY such that sup(Y') = sup(Y) and all_conf(Y') = all_conf(Y).

By applying mining of confidence-closed itemsets to Example 1, obtain not only itemset ABCDE but also CDE as confidence-closed itemsets since they have different confidence values and therefore no information loss occurs. So, call the support-closed pattern as SCP and the confidence closed pattern as CCP, respectively.

3. Mining Confidence - Closed Correlated Patterns

In this section, two algorithms for mining CCPs named CCFilter and CCMine are introduced. CCFilter is a simple algorithm that makes use of the existing support-closed pattern generator. CCFilter consists of the following two steps:

First, get the complete set of SCPs using the previous proposed algorithms [13].

Second, check each itemset and its all-possible subsets in the resulting set whether it is confidence-closed or not.

If its confidence satisfies min_a and it has no proper superset with the same confidence, it is generated as a confidence-closed itemset. CCFilter is used as a baseline algorithm for comparison in Section 5. CCFilter has a shortcoming: It generates SCPs with less confidence than min_a during the mining process. At the end, these patterns are removed. In order to solve this problem, CCMine integrates the two steps of CCFilter into one. Since all confidence has the downward closure property, push down the confidence condition into the process of the confidence-closed pattern mining. CCMine adopts a pattern-growth methodology proposed in [14].

The CLOSET+ [13] and CHARM [15] for mining SCPs, two search space-pruning techniques, item merging and sub-itemset merging, have been mainly used. However, if these techniques are applied directly into confidence-closed pattern mining, a complete set of CCPs cannot be obtained. This is because if there exists a pattern, these techniques remove all of its sub-patterns with the same support without considering confidence. Modify these optimization techniques so that they can be used in confidence-closed pattern mining.

Lemma 1 : Confidence-closed item merging

Let X is a correlated itemset. If every transaction containing itemset X also contains itemset Y but not any proper superset of Y , and all_conf(XY) = all_conf(X), then XY forms a confidence closed itemset and there is no need to search any itemset containing X but no Y .

Lemma 2 : Confidence-closed sub-itemset pruning

Let X is a correlated itemset currently under construction. If X is a proper subset of an already found confidence-closed itemset Y and all_conf(X) = all_conf(Y) then X and all of X's descendants in the set enumeration tree cannot be confidence-closed itemsets and thus can be pruned.

Lemma 1 means, the X-conditional database and the XY -conditional database separately have to mine if all_conf(X) ¹ all_conf(XY). However, though all_conf(X) and all_conf(XY) are different, the X and XY conditional databases are exactly the same if sup(X) = sup(XY). Using this property, avoid the overhead of building conditional databases for the prefix itemsets with the same support but different confidence. Maintain a list candidateList of the items that have the same support with the size of the X conditional database but are not included in the item merging because of their confidence. The list is constructed as follows. For X-conditional database, let Y be the set of items in f_list such that they appear in every transaction.

Do the following: Check that for each item Y_i in Y , if sup(Y_i)£ max_item_sup(X), X = XÈY_i; otherwise insert Y_i to candidateList. Check whether an itemset Z containing X(Z É X) is confidence-closed, also check whether the itemset Z È(Y' = Y1 . . .Yk, Y_i Î CandidateList) could be confidence-closed. Using this method, compute CCPs without generating the two conditional databases of X and of XY when all conf(X) > all conf(XY) and sup(X) = sup(XY).

Algorithm 4 shows the CCMine algorithm, which is based on the extension of CLOSET+ [13] and integrates the above discussions into the CLOSET+. Among a lot of studies for support-closed pattern mining, CLOSET+ is the fastest algorithm for a wide range of applications. CCMine uses another optimization technique to reduce the search space by taking advantage of the property of the all confidence measure. Lemma 3 describes the pruning rule.

Lemma 3 : Counting space pruning rule

Let a = i1i2 . . . ik. In the a-conditional database, for item x to be included in an all_confident pattern, the support of x should be less than sup(a)/min_a. Proof. In order for ax to be an all_confident pattern, max_item_sup(ax)sup(ax)/min_a.

Moreover, |sup(a)| ³ |sup(ax)|. Thus, max_item_sup(ax) ³sup(a)/min_a. Hence the lemma. With this pruning rule, reduce the set of items I_b to be counted and, thus, reduce the number of nodes visited when we traverse the FP-tree to count each item in I_b.

Example 2

Let us illustrate the confidence-closed mining process using an example. Figure 1 shows the running example of the transaction database DB. Let min sup = 2 and min_a = 40%. Scan DB once. Find and sort the list of frequent items in support descending order. This leads to f_list = (a:9, b:7, c:6, e:6, g:5, f:4, d:3, i:3, k:3, j:2, h:1). Figure 2 shows the global FP-tree. For lack of space, two representative cases: mining for prefix j:2 and eg:5. are shown after building the FP-tree mine the confidence-closed patterns with prefix j:2.

Computing counts :

Compute the counts for items a, c, e, f, and i to be included in the j-projected database by traversing the FP-tree shown in Fig. 2. First, use Lemma 3 to reduce items to be counted. The support of item z(z Î{a, c, e, f, i}) should be less than or equal to sup(j)/min_a = 2/0.4 = 5. With this pruning, items a, c and e are eliminated.

Now, compute counts of items f and i and construct j-projected database. They are 2 and 1, respectively. Pruning: We conduct pruning based on min sup and min_a. Item i is pruned since its support is less than min_sup. Item f is not pruned since and its confidence(2/4) is not less than min_a. Since f is the only item in j-conditional database, no need to build the corresponding FP-tree. And fj:2 is a CCP.

Algorithm CCMine: Mining confidence-closed correlated patterns

Input : A transaction database DB; a support threshold min sup a minimum all confidence threshold min_a

Output : The complete set of confidence-closed correlated patterns.

Method :

1. Let CCP be the set of confidence-closed patterns. Initialize CCP ¬ F
2. Scan DB once to find frequent items and compute frequent list f_list(=(f0, f1, . . .)).
3. Call CCMine(F, DB, f_list, CCP,F).

4. Procedure CCMine(á, CDB, f list, CCP, candidate List)

1. For each item Y in f_list such that it appears in every transaction of CDB, delete Y from f_list and set á ¬ YÈa if all_conf(Ya)³all_conf(á), otherwise insert Y into candidateList in the support increasing order; {confidence-closed item merging}
2. call GenerateCCP(á, candidate List, CCP);
3. build FP-tree for CDB using f list, which excludes all the items Y s in the previous step;
4. for each a_i in f_list (in reverse descending support order) do
5. set â = á È a_i;
6. call Generate CCP(â, candidate List, CCP);
7. get a set I_b of items to be included in â-projected database; {counting space pruning rule}
8. for each item in I_b, compute its count in â-projected database;
9. for each bj in I_b do
10. if sup(âb_j)< min_sup, delete b_j from I_b; {pruning based on min sup}
11. if all conf(âb_j)< min_ao, delete b_j from I_b;{pruning based on min a}
12. end for
13. call FP-mine(â, CDB, f list, CCP, candidate List);
14. delete the items that was inserted in step 1 from candidate List;
15. end for

5. Procedure Generate CCP(á, candidate List, CCP)

for k-itemset Y = Y1 . . . Yk(Y¡Î candidate List) **do**

add áÈY into CCP if all_conf(á ÈY) ¡Ý min a if áÈY is not a subset of X (in CCP) with the same support and confidence; {confidence-closed sub-itemset pruning}

end for

2. After building conditional FP-tree for prefix g:5 and we mine g:5-conditional FPtree with f_list = (a:5, e:5, b:4, c:3).

6. Confidence Item Merging

Try confidence-closed item merging of a and e. Delete a and e from f_list. Since all_conf(ag) < all_conf(g), insert a into candidateList. Then, extend the prefix from g to eg by the confidence-closed item merging. Generate CCP: generate eg:5 as a CCP. In addition, also generate aeg:5, in which item a comes from candidateList. Now, in f_list, only two items :4 and c:3 are left. Mine the CCPs with prefix ceg:3. First, we generate ceg as a CCP. However, we cannot generate aceg as CCP since all_conf(aceg) < min_ a. Since item b is the only item in f_list, bceg is a CCP. Again, abceg cannot be CCP, since it also does not satisfy min_a. In this way, mine the beg:4- conditional database and generate beg and abeg as a CCP. After returning mining beg:4-conditonal FP-tree, item a is removed from candidateList.

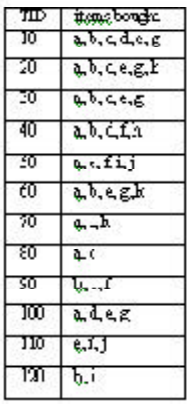


Fig. 1. A Transaction Database DB

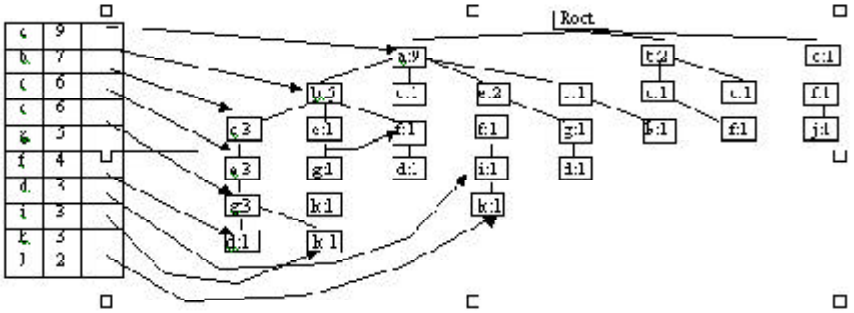


Fig. 2. FP-Tree for the Transaction Database DB

7. Experiments

In this section, we report out experimental results on the performance of CCMine in comparison with CCFilter algorithm. The result shows that CCMine always outperforms CCFilter especially at low min_sup. Experiments were performed on a 2.2GHz Pentium IV PC with 512MB of memory, running Windows 2000. Algorithms were coded with Visual C++.

The experiments were performed on two real datasets, as shown in Table 1. Pumsb dataset contains census data for population and housing and is obtained from http://www.almaden.ibm.com/software/

quest .Gazelle, a transactional data set comes from click-stream data from Gazelle.com. In the table, ATL/MTL represents average/maximum transaction length. The gazelle dataset is rather sparse in comparison with pumsb dataset, which is very dense so that it produces many long frequent itemsets even for very high values of support.

Table 1. Characteristics of Real Datasets.

Dataset	#Tuples	#Items	ATL/MTL
Gazelle	59602	497	2.5/267
Pumb	49046	2113	74/74

First show that the complete set of CCPs is much smaller in comparison with both that of correlated patterns and that of SCPs. Figure.3 shows the number of CCPs, correlated patterns, and SCPs generated from the gazelle data set. In this figure, the number of patterns is plotted on a log scale.

Figure 3(a) shows the number of patterns generated when min_sup varies and min_a is fixed while Figure 3(b) shows those when min_a varies and min_sup is fixed. First describe how many , can reduce the number of correlated patterns with the notion of CCPs. Figures 3(a) and 3(b) show that CCP mining generates a much smaller set than that of correlated patterns as the support threshold or the confidence threshold decreases, respectively. It is a desirable phenomenon since the number of correlated patterns increases dramatically as either of the thresholds decreases.

These figures also show that the number of SCPs is quite bigger than that of CCPs over the entire range of the support and confidence threshold. These results indicate that CCP mining generates quite a smaller set of patterns even at the low minimum support threshold and low minimum confidence threshold.

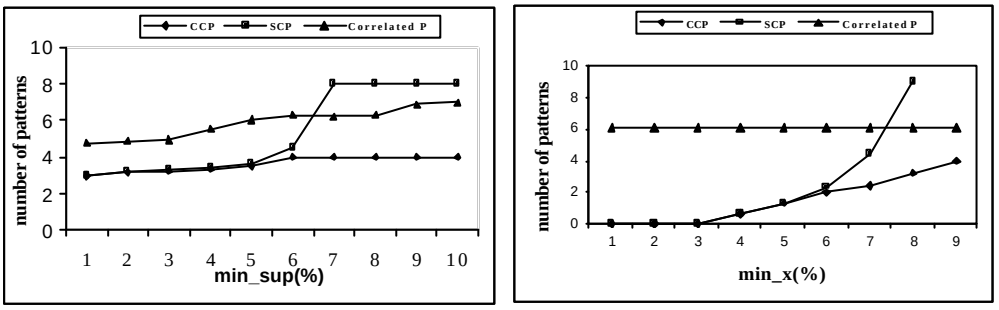


Fig. 3. Number of patterns generated from the gazelle data set.

Let us then compare the relative efficiency and effectiveness of the CCMine and CCFilter methods. Figure 4 (a) shows the execution time of the two methods on the gazelle dataset using different minimum support threshold while min_a is fixed at 25%. Figure 4(a) shows that CCMine always outperforms CCFilter over the entire supports of experiments. When the support threshold is low, CCMine is faster more than 100 times compared with CCFilter, e.g., with min sup 0.05%, CCFilter uses 20 seconds to finish while CCMine only uses 0.2. The reason why CCMine is superior to CCFilter is that CCFilter has to find all of the support closed patterns although many of them do not satisfy the minimum confidence threshold and the number of these patterns increases a lot as the minimum support threshold decreases. Figure 4(b) shows the performance on the gazelle dataset when min sup is fixed at 0.01% and min_a

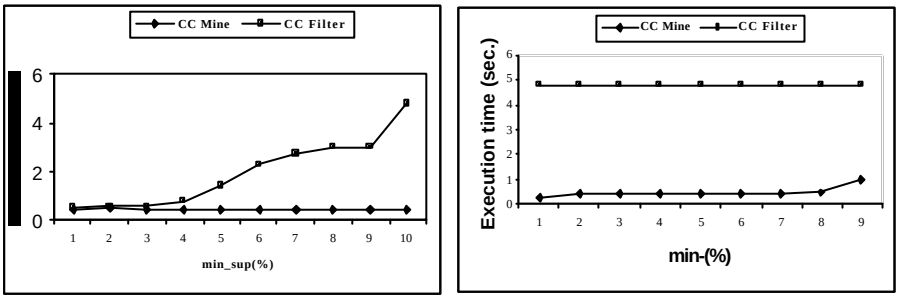
varies. As shown in the figure, CCMine always outperforms CCFilter and the execution times of CCMine increases very slowly while min_a decreases.

CCFilter almost does not change while min_a varies, which means it does not take any advantage from min_a. This is because it spends most of processing time on mining SCP. Now, conduct the experiments on the pumsb dataset, which is a dense dataset. Figure 5(a) shows the execution time on the pumsb dataset when min_a varies while min_sup is fixed at 60%. Figure 5(a) shows that CCMine method outperforms CCFilter method when min_sup is less than 60%. When min_sup becomes less then 50%, CCFilter run out of memory and cannot finish. Figure 5(b) shows that CCMine method always outperforms CCFilter method over entire range of min_a. In summary, experimental results show that the number of confidence closed correlated patterns are quite small in comparison with that of the support-closed patterns. The CCMine method outperforms CCFilter especially when the support threshold is low or the confidence threshold is high.

8. Conclusions

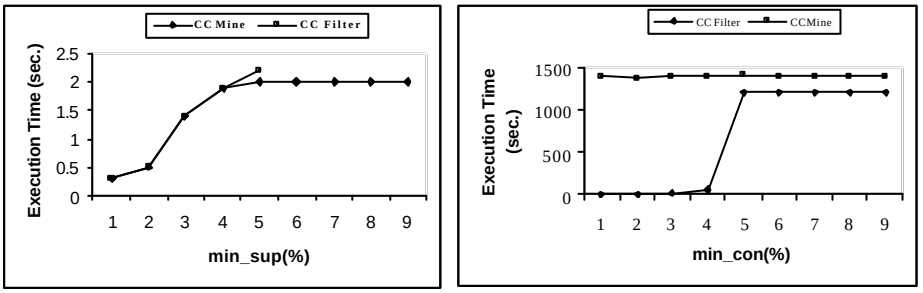
This paper presented an approach that can effectively reduce the number of correlated patterns to be mined without information loss. A new notion of confidence-closed correlated patterns is proposed. Confidence-closed correlated patterns are those that have no proper superset with the same support and the same confidence. For efficient mining of those patterns, we presented the CCMine algorithm. Several pruning methods have been developed that reduce the search space.

The performance study shows that confidence-closed, correlated pattern mining reduces the number of patterns by at least an order of magnitude in comparison with correlated (non-closed) pattern mining. It also shows that CCMine outperforms CCFilter in terms of runtime and scalability. Overall, it indicates that confidence-closed pattern mining is a valuable approach to condensing correlated patterns.



(a) *min á* = 25% (b) *min sup* = 0.01%.

Fig. 4. Execution time on gazelle data set



(a) when min á = 60% *(b) when min sup = 50%.*

Fig. 5. Execution time on the pumsb dataset.

All confidence is one of several favorable correlation measures, with null in variance property. Based on the examination, CCMine can be easily extended to mining some correlation measures, such as coherence or bond [6, 5, 8]. It is an interesting research issue to systematically develop other mining methodologies, such as constraint-based mining, approximate pattern mining, etc. under the framework of mining confidence-closed correlated patterns.

9. References

1. C. C. Aggarwal and P. S. Yu. A new framework for itemset generation. In Proc. 1998 ACM Symp. Principles of Database Systems (PODS'98), pages 18–24, Seattle, WA, June 1999

2. S. Brin, R. Motwani, and C. Silverstein. Beyond market basket: Generalizing association rules to correlations. In Proc. 1997 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'97), pages 265–276, Tucson, Arizona, May 1997.

3. S. Morishita and J. Sese. Traversing itemset lattice with statistical metric pruning. In Proc. 2000 ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems (PODS'00), pages 226–236, Dallas, TX, May 2001.

4. P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In Proc. 2002 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'02), pages 32–41, Edmonton, Canada, July 2002.

5. E. Omiecinski. Alternative interest measures for mining associations. IEEE Trans. Knowledge and Data Engineering, 15:57–69, 2003.

6. Y.-K. Lee, W.-Y. Kim, D. Cai, and J. Han. CofiMine: Efficient Mining of Correlated Patterns. In Proc. 2003 Int. Conf. Data Mining (ICDM'03), pages 581–584, Melbourne, FL, Nov. 2003.

7. S. Ma and J. L. Hellerstein. Mining mutually dependent patterns. In Proc. 2001 Int. Conf. Data Mining (ICDM'01), pages 409–416, San Jose, CA, Nov. 2001.

8. H. Xiong, P.-N. Tan, and V. Kumar. Mining Strong Affinity Association Patterns in Data Sets with Skewed Support Distribution. In Proc. 2003 Int. Conf. Data Mining (ICDM'03), pages 387–394, Melbourne, FL, Nov. 2003.

9. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In Proc. 7th Int. Conf. Database Theory (ICDT'99), pages 398–416, Jerusalem, Israel, Jan. 1999.

10. R. J. Bayardo. Efficiently mining long patterns from databases. In Proc. 1998 ACMSIGMOD Int. Conf. Management of Data (SIGMOD'98), pages 85–93, Seattle,WA, June 1998.

11. J. Pei, G. Dong, W. Zou, and J. Han. On computing condensed frequent pattern bases. In Proc. 2002 Int. Conf. on Data Mining (ICDM'02), pages 378–385, Maebashi, Japan, Dec. 2002.

12. M. Zaki. Generating Non-redundant Association Rules. In Proc. 2000 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'00), Aug. 2000.

13. J. Wang, J. Han, and J. Pei. Closet+: Searching for the best strategies for mining frequent closed itemsets. In Proc. 2003 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'03), Washington, D.C., Aug. 2003.

14. J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'00), pages 1–12, Dallas, TX, May 2000.

15. M. Zaki and C. Hsiao. CHARM: An Efficient Algorithm for Closed Itemset Mining, In Proc. SDM'02, Apr. 2002.

16. A. Krishnan, R. Hemalatha, C. Senthamari, "Efficient Mining of Association Rules using Partition Algorithm", National Conference on Mathematical and Computational Models (NCMCM 2003), December 2003

17. A. Krishnan, R. Hemalatha, C. Senthamarai, "Mining of Association Rules in Distributed Database Using Partition Algorithm", International Conference on Systemic, Cybernetics & Informatics (ICSCI 2004), February, 2004

18. A. Krishnan, R. Hemalatha, "Parallel Association Rule Mining – Finding Frequent Patterns Without Candidate Generation", National Level Conference, Tech Fete 2004 on "Intelligence Techniques", February 2004

19. A. Krishnan, R. Hemalatha, R. Hemamalini, C. Senthamarai, "An Efficient Data Mining Approach to Extract High Frequency Rules from Different Data Sources", National Conference on Recent Trends in Computational Mathematics, March 2004

About Authors

Mrs. R Hemalatha is a Lecturer in K. S. R. College of Technology, Tiruchengode. Namakkal Dt., Tamil Nadu. She holds M.Sc. (I.T).
E-mail : hemaa_msc@yahoo.com

Dr. A Krishnan is a Principal in R. R. Engineering College, Tiruchengode. Namakkal Dt., Tamil Nadu.
E-mail : a_krishnan26@hotmail.com

Mrs. C Senthamarai is heading the Computer Science Department in K.S.R.College of Technology, Tiruchengode. Namakkal Dt., Tamil Nadu. She holds MCA.
E-mail : senthukumaran20023@yahoo.com

Ms. R. Hemamalini is a Lecturer in Computer Science Department in R. R. Engineering. College, Tiruchengode. Namakkal Dt., Tamil Nadu. She holds MCA.
E-mail : rk_hema2000@yahoo.com

Mining Frequent Item Sets More Efficiently Using ITL Mining

R Hemalatha A Krishnan R Hemamathi

Abstract

Correlated The discovery of association rules is an important problem in data mining. It is a two-step process consisting of finding the frequent itemsets and generating association rules from them. Most of the research attention is focused on efficient methods of finding frequent itemsets because it is computationally the most expensive step. This paper presents a new data structure and a more efficient algorithm for mining frequent itemsets from typical data sets. The improvement is achieved by scanning the database just once and by reducing item traversals within transactions. The performance comparisons of the algorithm against the fastest Apriori implementation and the recently developed H-Mine algorithm are given here. These results show that the algorithm outperforms both Apriori and H-mine on several widely used test data sets.

Keywords : Data Mining, Data Structure.

0. Introduction

Association Rules are used to identify relationships among sets of items. They are relevant to several domains such as the analysis of market basket transactions in retail stores, target marketing, fraud detection, finding patterns in telecommunication alarms, etc. In retail stores for example, this information is useful to increase the effectiveness of advertising, marketing, inventory control, and stock location on the shop floor. Since the introduction of association rules a decade ago [1], a large number of increasingly efficient algorithms have been proposed [2,3,4,5,6,7].

The process of mining association rules consists of two steps :

1) Find the frequent itemsets that have *minimum support*;

2) Use the frequent itemsets to generate association rules that meet the *confidence threshold*.

Between these two steps, step 1 is the most expensive since the number of itemsets grows exponentially with the number of items. The strategies developed to speed up this process can be divided into two categories. The first is the candidate generation-and-test approach. Algorithms in this category include Apriori and its several variations [1,2,13,14,15,16]. They use the Apriori property also known as antimonotone property that any subset of a frequent item set must be a frequent item set. In this approach, a set of candidate item sets of length $n + 1$ is generated from the set of item sets of length n and then each candidate item set is checked to see if it meets the *support threshold*. The second approach of pattern-growth has been proposed more recently.

It also uses the Apriori property, but instead of generating candidate item sets, it recursively mines patterns in the database counting the support for each pattern. Algorithms in this category include TreeProjection [7], FP-Growth [3,14,15] and H-Mine [4]. Algorithms based on candidate generation and test such as Apriori runs very slowly on long pattern data sets because of the huge number of candidate itemsets it has to generate and test.

Testing the candidate itemsets for minimum support requires scanning the whole database many times, although for some small to moderate size databases, the scan can be made faster by storing the database in main memory. The problem of generating candidates can be avoided by using the pattern-growth approach.

For most data sets, these algorithms perform better than Apriori. Among the existing pattern-growth algorithms, H-Mine runs faster than TreeProjection and FP-Growth on several commonly used test data sets. H-mine scans the transaction database twice. It performs repeated horizontal traversals of the transactions in memory while generating frequent itemsets. H-Mine also needs to continually re-adjust the links between transactions in its H-struct data structure during the mining process. If these costs are reduced, the mining process will be improved further.

A more efficient algorithm based on the pattern-growth approach is introduced here. It reduces the cost by scanning the database only once, by significantly reducing the horizontal traversals of the transactions in memory and keeping the links between transactions in memory unchanged during the mining process. To achieve these reductions in cost, a new data structure called Item-Trans Link (ITL) and the new algorithm called ITL-Mine is presented.

The performance of ITL-Mine is compared with Apriori and H-Mine algorithms and found that ITL-Mine performs better on a range of widely used test data sets. H-mine was chosen for performance comparisons because it is an improvement over FP-Growth and TreeProjection algorithms. The H-Mine algorithm is based on its description in [4]. The Apriori program used is from [11], which is generally acknowledged as the fastest Apriori implementation available.

Tid	Items
1	3 4 5 6 7 9
2	1 3 4 5 13
3	1 2 4 5 7 11
4	1 3 4 8

Tid	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	0	1	1	1	1	1	0	1	0	0	0	0
2	1	0	1	1	1	0	0	0	0	0	0	0	1
3	1	1	0	1	1	0	1	0	0	0	1	0	0
4	1	0	1	1	0	0	0	1	0	0	0	0	0

Fig. 1. The Transaction Database

1. Definitions and ITL Data Structure

This section defines the terms used for describing association rule mining. The conceptual basis for design of the data representation is presented followed by a description of the ITL data structure.

1.1 Definition of Terms

The basic terms needed for describing association rules using the formalism of [1,13,14,15]. Let I={i1, i2, … , in} be a set of items, and *D* be a set of transactions, where a transaction *T* is a subset of *I* (*T* Í *I*). Each transaction is identified by a *TID*. An association rule is an expression of the form *X* Þ *Y*, where *X* Ì *I*, *Y* Ì *I* and *X* Ç *Y* = Æ. Note that each of X and Y is a set of one or more items and the quantity of each item is not considered. *X* is referred to as the *body* of the rule and *Y* as the *head*.

An example of association rule is the statement that 80% of transactions that purchase A also purchase B and 10% of all transactions contain both of them. Here, 10% is the *support* of the itemset {A, B} and 80% is the *confidence* of the rule *A* Þ *B*. An itemset is called a *frequent itemset* if its *support* is greater than or equal to a *support threshold* specified by the user, otherwise the itemset is *not frequent*. A *k-frequent itemset* is a frequent itemset that contains *k* items.

1.2 Binary Representation of Transactions

The transactions in the database could be represented as a binary table [1] as shown in Figure 1. Counting the support for an item can be considered as counting the number of 1's for that item in all the transactions. In most datasets, the number of items in each transaction is much smaller than the total number of items, and therefore the binary table representation will not allow efficient use of memory. Therefore, use a more efficient representation scheme.

1.3 Item-Trans Link (ITL) Data Structure

Researchers have proposed various data representation schemes for association rule mining. They can be broadly classified as horizontal data layout, vertical data layout, and a combination of the two. Most candidate generation and test algorithms (e.g. Apriori) use the horizontal data layout and most pattern-growth algorithms like FP-Growth and H-Mine use a combination of vertical and horizontal data layouts. A data structure called **Item-Trans Link (ITL)** that combines the vertical and horizontal data layouts is proposed (see Figure 2).

The data representation used by the algorithm is based on the following observations :

1. Item identifiers may be mapped to a range of integers.
2. Transaction identifiers can be ignored provided the items of each transaction are linked together.

ITL consists of an item table (ItemTable) and the transactions linked to it (TransLink) as follows:

1. ItemTable: It contains all the items and the support of each item. It also has a link to the first occurrence of each item in the transactions of TransLink described below.
2. TransLink: It represents the items of every transaction for all the transactions in the database. The items of a transaction are arranged in sorted order.

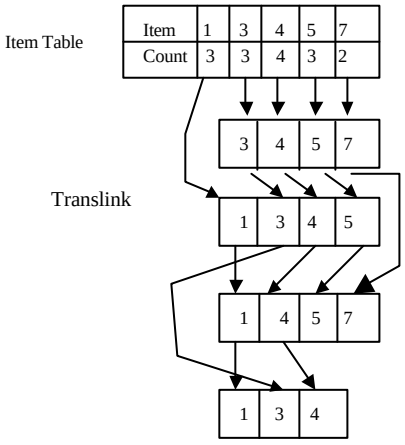


Fig. 2. The Item-Trans Link (ITL) Data Structure

For each item in a transaction, it contains a link to the next occurrence of that item in another transaction. In other words, this link will represent all the 1's for each item so that the counting can be done quickly. For example, in Figure 2, to check the occurrences of item 7, go to the cell of 7 in tid 1 in the TransLink and then directly to the next occurrence of 7 in tid 3 without traversing tid 2. Since ITL has features of both horizontal and vertical data layouts, it is general and flexible enough to be used by algorithms that need horizontal, vertical or combined data layout. It also makes it possible to combine existing ideas for efficient mining based on both layouts.

ITL is similar to H-struct proposed in [4], except for the vertical links between the occurrences of each item in the transactions. In H-struct, the links always point to the first item of the transaction, and therefore to get a certain item, traverse the transaction from the beginning. ITL points directly to the occurrence of the item which makes it faster to traverse all occurrences of an item.

2. ITL-Mine Algorithm

This section describes the ITL-Mine algorithm, and provides a running example. ITL-Mine assumes that the ItemTable and TransLink will fit into main memory. With the availability of increasingly larger sizes of main memory that currently approach gigabytes, many small to moderate databases will fit in the main memory. However, the extension of this algorithm to mine very large databases is currently in progress.

There are three steps in the ITL-Mine algorithm as follows:

- Construct ItemTable and TransLink: In this step, the ItemTable and TransLink are constructed by a single scan of the transaction database. At the end of this step, the 1-frequent itemsets will be identified in the ItemTable by the support count.
- Prune: Using the anti-monotone property, the infrequent items are pruned or deleted from the TransLink since infrequent items will not be useful in the next step.
- Mine Frequent Itemsets: In this step, all the frequent itemsets of two or more items are mined using a recursive function as described further in this section.

Example 1 : The ITL-Mine algorithm is illustrated by this example. Let Table 1 be the transaction database and suppose the user wants to get the Frequent Itemsets with minimum support = 50% (minimum 2 transactions). In Step 1, all transactions in the database are read in a single scan to construct the ItemTable and TransLink. For each item in a transaction, the existence of the item in the ItemTable is checked. If the item is not present in the ItemTable, it is entered with an initial count of 1, otherwise the count of the item is incremented. After that, the item is entered in the proper location in the TransLink.

3. Procedure Construct-ITL

For all transactions in the DB

For all items in transaction

If item in ItemTable

Increment count of item

Else

Insert item with count = 1

End If
Insert Item into TransLink
Connect link of previous
occurrence to this item
End For
End For

Procedure Prune-ITL
For all xÎItemTable where count(x)< min_sup
Delete x from ItemTable and TransLink
End For

Procedure Mine-FI
For all xÎItemTable
Add x to the set of Frequent Itemsets
Prepare and fill tempList for x
For all yÎtempList where count(y)³min_sup
Add xy to the set of Frequent Itemsets
For all zÎtempList after y
where count(z) ³min_sup
RecMine (xy, z)
End For
End For
End For

Procedure RecMine(prefix, test_item)
tlp:= tid-list of prefix
tli:= tid-list of test_item
tl_current = Intersect(tlp,tli)
If size(tl_current) ³?min_sup
new_prefix:= prefix + test_item
Add new_prefix to the set of Frequent
Itemsets

For all zÎ tempList after test_item

Where count (z) ³ ?min_sup

RecMine(new_prefix, z)

End For

End If

Fig. 4. Algorithms for Construct, Prune and Mine-FI and the links are made.

Prefix	**TempList (count)**	**Freq-Itemsets (count)**
1	3 (2), 4 (3), 5 (2), 7 (1)	1 (3), 1 3 (2), 1 4 (3), 1 5 (2)
1 3	4 (2), 5 (1)	1 3 4 (2)
1 4	5(2), 7(1)	1 4 5 (2)
3	4 (3), 5 (2), 7 (1)	3(3), 3 4 (3), 3 5 (2)
3 4	5 (2), 7(1)	3 4 5 (2)
4	5 (3), 7(2)	4 (4), 4 5 (3), 4 7 (2)
4 5	7 (2)	4 5 7 (2)
5	7(2)	5 (3), 5 7 (2)
7	None	7(2)

Fig. 3. Mining Frequent Itemsets Recursively Support of frequent itemsets shown in brackets)

On completing the database scan, the 1-frequent itemsets can be identified in the ItemTable as {1, 3, 4, 5, 7}. In Step 2, all items in the ItemTable are traversed to prune the infrequent items. If the support count of an item in the ItemTable is below the minimum, it is removed from both the ItemTable and the TransLink. The pruning is done by following the link from the ItemTable to traverse all the occurrences of that item in the TransLink. After the pruning, the TransLink is as shown in Figure 2.

In the last Step, each item in the ItemTable will be used as a starting point to mine all longer frequent itemsets for which it is a prefix. As an example, starting with item 1, follow the link to get all other items that occur together with item 1. Items that occur together with item 1 will be registered in a simple table that called TempList together with their support count and list of tids. As in Figure 3, for prefix 1, items {3, 4, 5} that are frequent (their support ³ 2). Generating the frequent patterns for this step involves simply concatenating the prefix with each frequent-item.

As an example, the frequent itemsets for this step are 1 3 (2), 1 4 (3) and 1 5 (2). After generating the 2-frequent-itemsets for prefix 1, since we have got the tid-list of each item in the TempList, we can recursively use the tid intersection scheme to generate the subsequent frequent itemsets. For example, we can use the tid-list of 3 and intersect with tid list of 5 to generate frequent itemsets 1 3 5. At the end of recursive calls with prefix item 1, all frequent itemsets that contains item 1 will be generated: 1 (3), 1 3 (2), 1 4 (3), 1 5 (2), 1 3 4 (2), 1 4 5 (2). In the next sequence, item 3 will be used to generate all frequent itemsets that contain item 3 but does not contain item 1. Then item 5 will be used to generate all frequent itemsets that contain item 5 but does not contain items 1 and 3. The algorithm of ITL-Mine is shown in Figure 4.

3. Performance Study

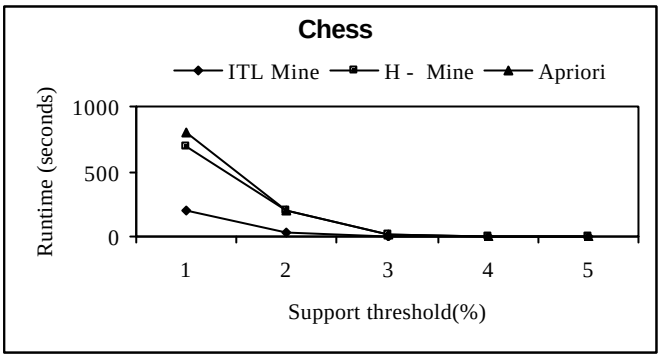


Fig. 5. Performance comparison of Apriori, H-Mine and ITL-Mine on chess datasets

In this section, the performance evaluation of ITL-Mine is presented. All the tests were performed on an 866MHz Pentium III PC, with 128 MB RAM and 30 GB HD running Microsoft Windows 2000. ITL-Mine is written in Microsoft Visual C++ 6.0. In this paper, the runtime includes both CPU time and I/O time.

The chess dataset has been used to test the performance Chess (3196 trans, max 37 items per trans). The Chess dataset is derived from the steps of Chess games. Chess is dense dataset since it produces many long patterns of frequent itemsets for very high values of support. It is downloaded from [9]. Comparison of ITL-Mine with the fastest available Apriori implementation [12] and the implementation of H-Mine. Performance comparison of Chess dataset is shown in Figure 5. The result shows that ITL-Mine outperforms Apriori and H-Mine on these datasets for the given parameters.

4. Discussion

The algorithm follows the pattern growth approach without candidate generation, so it is compared with H-Mine, the most recently proposed algorithm of this class. H-Mine is also considered to be more efficient than TreeProjection and FP-growth algorithms [4]. The better performance of the algorithm is:

- This algorithm traverses the database only once and performs the rest of the mining process using ITL. H-mine performs two scans of the database to build the H-struct data structure.
- After the ITL data structure is constructed and pruned to remove infrequent 1-itemsets, it remains unchanged while mining all of the frequent patterns. In H-Mine, the pointers in the H-struct need to be continually readjusted during the extraction of frequent patterns and so needs additional computation.
- ITL-Mine uses a simple temporary table called TempList during the recursive extraction of frequent patterns. ITL-Mine stores the tid-list of each item in the TempList and uses tid intersection scheme to generate frequent patterns. The TempList stores only the information for the current recursive call and the space will be reused for the subsequent recursive calls. H-Mine builds a series of header tables linked to the H-struct and it needs to change pointers to create or re-arrange queues for each recursive call.

- The recursive calls in H-Mine also involve repeated traversals of relevant parts of the H-struct. ITL-Mine avoids these repeated traversals by using the tid intersections. So the additional computation required by H-Mine to extract the frequent itemsets from Hstruct are more than for ITL-Mine. It supports more efficient interactive mining, where the user may experiment with different values of minimum support levels. In this case the pruning step would be skipped. Using the constructed ItemTable and TransLink in the memory, if the user wants to change the value of support threshold, there is no need to re-read the transaction database.

5. Conclusion

In this paper, a generic data structure called Item-Trans Link (ITL) and a new algorithm called ITL-Mine for discovering frequent itemsets is presented. The algorithm needs to scan the transaction database only once. The performance of ITL-Mine is compared against Apriori and H-Mine on chess dataset and the result show that ITL-Mine outperforms both Apriori and H-Mine on the data set for the given ranges of support levels. Assume that the ItemTable and TransLink will fit into main memory.

However, this assumption will not apply for huge databases. The extension of ITL-Mine for very large databases is currently underway. Several researchers have investigated ways to reduce the size of frequent itemsets to manageable levels for users and to allow greater user focus in the mining process [8]. User-specified constraints on the mining process represent a promising approach to this problem.

6. References

1. R. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules between Sets of Items in Large Databases”, Proc. of the ACM SIGMOD Conf., Washington DC, 1993.
2. R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules”, Proc. of the 20th Int. Conf. on VLDB, Santiago, Chile, 1994.
3. J. Han, J. Pei, and Y. Yin, “Mining Frequent Patterns without Candidate Generation”, Proc. of the ACM SIGMOD Conf., Dallas, TX, 2000.
4. J. Pei, J. Han, H. Lu, S. Nishio, S. Tang, and D. Yang, “H-Mine: Hyper-Structure Mining of Frequent Patterns in Large Databases,” Proc. of the 2001 IEEE ICDM, San Jose, California, 2001.
5. M. J. Zaki, “Scalable Algorithms for Association Mining,” IEEE Transactions on Knowledge and Data Engineering, vol. 12, pp. 372-390, May/June 2000.
6. P. Shenoy, J. R. Haritsa, S. Sudarshan, G. Bhalotia, M. Bawa, and D. Shah, “Turbo-charging Vertical Mining of Large Databases,” Proc. of the ACM SIGMOD Conf., Dallas, TX USA, 2000.
7. R. Agarwal, C. Aggarwal, and V. V. V. Prasad, “A Tree Projection Algorithm for Generation of Frequent Itemsets”, Journal of Parallel and Distributed Computing (Special Issue on High Performance Data Mining), 2000.
8. J. Pei, J. Han, and L. V. S. Lakshmanan, “Mining Frequent Itemsets with Convertible Constraints”, Proc. of 17th ICDE, Heidelberg, Germany, 2001.
9. Irvine Machine Learning Database Repository: http://www.ics.uci.edu/~mlearn/ MLRepository.html
10. http://www.ecn.purdue.edu/KDDCUP
11. http://www.cs.sfu.ca/ ~peijian/personal/publications
12. Apriori version 4.01, available at http://fuzzy.cs.unimagdeburg.de/~borgelt/).

13. A. Krishnan, R. Hemalatha, C. Senthamarai, “Mining of Association Rules in Distributed Database Using Partition Algorithm”, International Conference on Systemic, Cybernetics & Informatics (ICSCI 2004), February, 2004

14. A. Krishnan, R. Hemalatha, “Parallel Association Rule Mining – Finding Frequent Patterns Without Candidate Generation”, National Level Conference, Tech Fete 2004 on “Intelligence Techniques”, February 2004.

15. A. Krishnan, R. Hemalatha, R. Hemamalini, “Mining Frequent Patterns Without Candidate Generation in Distributed Databases”, National Conference on Distributed Database and Computing, March 2004.

16. A. Krishnan, R. Hemalatha, C. Senthamarai, “Association Rule mining with the Pattern Repository”, National Conference on Data Mining, December 2004.

About Authors

Mrs. R Hemalatha is a Lecturer in Department of Computer Science in K. S. R. College of Technology, Tiruchengode. Namakkal Dt., Tamil Nadu. She holds M.Sc. (I.T.)
E-mail : hemaa_msc@yahoo.com

Dr. A Krishnan is a Principal in the R. R. Engineering College, Tiruchengode. Namakkal Dt., Tamil Nadu.
E-mail : a_krishnan26@hotmail.com

Ms. R Hemamathi is a MCA Final Year Student at M. Kumarasamy College of Engineeing, Thalavappalayam, Karur Dist., Tamil Nadu.
E-mail : hemamathi_mca@rediffmail.com

Temporal Association Rule Using Without Candidate Generation

Keshri Verma O P Vyas

Abstract

Associationship is an important component of data mining. In real world, the knowledge used for mining rule is almost time varying. The items have the dynamic characteristic in terms of transaction, which have seasonal selling rate and it holds time-based associationship with another item. In database, some items which are infrequent in whole dataset may be frequent in a particular time period. If these items are ignored then associationship result will no longer be accurate. To restrict the time based associationship, calendar based pattern can be used [5]. Calendar units such as months and days, clock units, such as hours and seconds & specialized units , such as business days and academic years, play a major role in a wide range of information system applications.[11] Our focus is to find effective time sensitive algorithm for mining associationship by extending frequent pattern tree approach [3]. This algorithm reduces the time complexity of existing technique[5]. It also uses the advantages of divide & conquer method to decompose the mining task into a smaller tasks for database.

Keywords : Data Mining, Temporal Data Mining, Temporal Association Rule, Frequent Pattern Approach.

0. Introduction

The Associationship is an important component of data mining. It indicates the co-relationship of one item with another. For example Eggà coffee (support 3%, confidence 80%)

It means that 3% of all transaction contain both egg & coffee, and 80% of transaction that have egg also have coffee in them.

One important extension to association rule is to include a temporal dimension. For example egg and coffee may be ordered together primarily between 7 to 11 AM in this interval the support & confidence is 40% but at another interval, support is as low .005% in other transaction [5]. In market, various items have a seasonal selling rate.. It is also importance because many items are introduced or removed form the database, that is items lifespan[ZMTW] which means that item is valid on specific time interval. To discover such temporal intervals (with calendar information) together with the association rules that hold during the time interval may lead to useful knowledge. If calendar schema is applied it is called calendar based temporal association rule

A hierarchy of calendar concepts determines a calendar schema. A calendar unit such as months and days, clock units, such as hours and seconds & specialized units, such as business days and academic years, play a major role in a wide range of information system applications.[11]. A calendar schema defines a set of simple calendar – based patterns. Each calendar pattern defines a set of time intervals.

Our data mining problem is to discover all temporal association rules w.r.t. calendar schema from a set of time stamped transactions. This paper ,improve an existing frequent pattern tree approach to discover temporal association rule to increase the optimization over existing one[5].

The organization of the paper as follows :

we develop a data structure called Temporal FP-tree.

we introduce a novel algorithm for mining frequent pattern using Temporal FP-tree.

The rest of the paper is organized in five section In Section 2, we discuss some related works. In section 3 we define temporal association rule in term of calendar schema. In Section 4 elaborate the extended algorithm of frequent pattern approach ,section 5 shows conclusion & future works and section 6 provides application of above investigation.

1. Related Work

The concept of association rule was introduced as Apriori algorithm [1]. Its performance was improved by deploying frequent pattern tree approach [3]. In paper [7] the omission of the time dimension in association rule was very clearly mentioned. A temporal aspect of association rule was given by Juan [2]. The transaction in the database are time stamped and time interval is specified by the user to divide the data into disjoint segments, like month , days & years. Further The cyclic association rule was introduced by Ozden [7] with minimum support & high confidence. Using the definition of cyclic association rule, It may not have high support & confidence for the entire transactional database.

A nice bibliography of temporal data mining can be found in the Roddick literature [8].

Rainsford & Roddick presented extension to association rules to accommodate temporal semantics. According to his logic the technique first search the associationship than it is used to incorporate temporal semantics. It can be used in point based & interval based model of time simultaneously[10]. A Frequent pattern approach for mining the time sensitive data was introduced in[9] Here the pattern frequency history under a tilted-time window framework in order to answer time-sensitive queries. A collection of item patterns along with their frequency histories are compressed and stored using a tree structure similar to FP-tree and updated incrementally with incoming transactions [9].

2. Problem Definition

2.1 Association Rule

The concept of association rule, which was motivated by market basket analysis and originally presented by Agrawal. [1]. Given a set of T of transaction, an association rule of the form Xà Y is a relationship between the two disjoint itemsets X & Y. An association rule satisfies some user-given requirements. The support of an itemset by the set of transaction is the fraction of transaction that contain the itemset. An itemset is said to be large if its support exceeds a user-given threshold minimum support. The confidence Xà Y over T is a transaction containing X and also containing Y. Agrawal first introduced Apriori algorithm in1994 for mining associationship in market basket data [1]. Due to complex candidate generation in the data set Jiewai Han invented a new technique of FP-growth method for mining frequent pattern without candidate generation [3]. Efficiency of this mining technique is better than all mos all algorithm like Apriori, aprioriTid, Apriori Hybridm because (1). a large dataset is compressed into a condensed ,smaller data structure, which avoids costly & repeated data scan , (2). FP-tree-based mining adopts a pattern-fragment growth method too avoid the costly generation of a large number of candidate generation sets and,(3). A partitioning-based , divide- and-conquer method is used to decompose the mining task into a set of similar tasks for conditional database, which dramatically reduce the search space.

In our opinion this mining technique will be become more useful if we include the time factor in to it.

2.2 Temporal association rule

Definition 1 : The frequency of and itemset over a time period T is the number of transactions in which it occurs divided by total number, of transaction over a time period. In the same way , confidence of a item with another item is the transaction of both items over the period divided by first item of that period.

Support(A) = Frequency of occurrences of A in specified time interval / Total no of Tuples in specified time interval

Confidence(A => B[Ts,Te]) = Support_count(A Ç B) over Interval / occurrence of A in interval

2.3 Simple calendar based Pattern :

When temporal information is applied in terms of date, month , year & week form the term calendar schema. It is introduced in temporal data mining. A calendar schema is a relational schema (in the sense of relational databases) R = (f_n : D_n, F_{n-1} : D_{n-1},.....F_1 :d_1) together with a valid constraint. A calendar schema (year : {1995,1996,1997.....} , month : {1,2,3,4,.....12}, day : {1,2,3.....31} with the constraint is valid if that evaluates (yy, mm, dd) to True only if the combination gives a valid date. For example <195,1,3> is a valid date while ,<1996,2,31> is not.

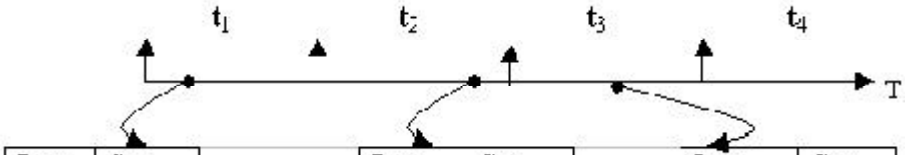
In calendar pattern , the branch e cover e' in the same calendar schema if the time interval e' is the subset of e and they all follow the same pattern. If a calendar pattern <d_n,d_{n-1},d_{n-2}.....d_1> covers another pattern <d'_n, $d'_{n-1;}$, d'_{n-2} .....d_1> if and only if for each I, 1<=i<=n or $d_i = d'_i$.

Now Our task is to mine frequent pattern over arbitrary time interval in terms of calendar pattern schema.

3. Proposed Work

The support of dataset in the data warehouse can be maintained by dividing it in different intervals. The support of a item in interval t1 can not be the same in interval t2. A infrequent or less support item in interval t1 can be frequent item in interval t2.

The calendar schema is implemented by applying apriori algorithm [5]. It follows the candidate generation approach in order to mine the frequent item. We assist here that divide & conquer approach is more efficient than apriori approach. It construct a tree & each branch indicate the association ship of item. It reduces the size of dataset and increases the performance & efficiency of algorithm. It can solve following queries (1) What are the frequent set over the interval t_1 and t_2 ? (2) what are the period when (a,b) item are frequent ? (3) Item which are dramatically change from t4 to t1.



Item	Sup.
a	100
b	92
c	80
ab	78
ac	75

Item	Sup.
b	120
a	85
c	76
ab	76
ac	53

Item	Sup.
a	48
b	50
c	45
ab	37
ac	29

Figure 1 Frequent pattern in different interval

Lemma 1 : During transaction in database the association of a item over the support x can be obtained by projection of the branch of FP-tree.

Rationale : Based on the TFP-tree construction process its frequent item can be projected into a single branch.

For a path a_1, a_2 $a_{3.....}$ a_k from the root to a node in a FP-tree. Suppose a_{ak} be the count at the node labeled $a_{k\ and}$ c'_{ak} be the sum of the count of the branch of the node.

Table 1. A Transaction database in running example

Tid	Item bought	Date	Calendar pattern	Item in Desceing order of Frequency
100	f, a,c,d,g,i,m,p	01/01/2004	<*,01,04>	f,c,a,m
200	a,b,c,f,l,m,o	01/01/2004	<*,01,04>	f,c,a,b,m,o
300	b,f,h,j,o	02/01/2004	<*,01,04>	f,c,b,o
500	a,f,c,,e,l,p,m,n	03/06/2004	<*,06,04>	f,c,a,e,m,l
600	f,a,c,d,e	04/06/2004	<*,06,04>	f,c,a,e
700	b,f,h,j,m,l,o	06/06/2004	<*,06,04>	f,m,l,

Definition (Temporal FP-Tree) **–** A Temporal frequent pattern (FP) is tree structure defined below.

1. It consists of root labeled as “null”.
2. It consist of a set of item-prefix subtrees as the children of the root, and a frequent item header table.
3. Each node in the item prefix subtree consists of four fields : (a) Item name - Item name represents the name of item which is registers on that node (b) count - count registers the number of transactions represented by the portion of the path reaching this node (c) node link - node-link links to the next node of temporal FP-tree (d) calendar pattern time calendar pattern represent the time in which the item transaction appeared.
4. Each entry in the Frequent –item header table consists of two fields (1) Item name (2) Head of node link

Algorithm : (FP- Tree construction)

Input : A transaction database DB and a minimum support threshold

Output : FP Tree, Fp Tree, Frequent item

Method : The FP tree is constructed ad follows :

1. Scan the database DB once. Collect f, the set of frequent item and support of each item. Sort F from support in descending order as Flist, the list of frequent items
2. Create the root of Temporal FP tree and label it as “Null”. For each transaction in DB and do the following

   Select the frequent items in Trans and sort them in descending order of Flist. Let the sorted frequent –item list in the Trans be p[P] where p is first element and P is the remaining list. Cal insert_tree(p[P],T).

```
Procedure insert_tree(p[P],T).
{

Step(1)  If  T has a child N such that
Step(2)  if (N.time = P.time) then
Step(3) if (N.itemname = P. itemname) then
   Step(a) N.count = N.count +1  // Increment the count by 1
Step(4) else create a new node // Node created on the same branch
Step(5) Link to its parent P.count = 1  // Initialize the counter by 1.
Step(6)  else create a new Branch link from the root.
Call insert_tree(P,N)  recursively

} // End of Function
```

Temporal Frequent pattern Tree : Design & Construction

Let I = {a_1, $a_{2,}$ $a_{3..}$ a_m} be a set of items , and a transaction database DB {T_1, $T_{2,}$ $T_{3.....}$ T_n } where T_i {i Î [1..n] } is a transaction which contains a set of items in I.

3.1 Temporal Frequent- Pattern Tree

To design the Temporal FP-tree for frequent pattern mining, let's first examine example from table1.

1. Since time is the most important feature of real world data set, arrange the item in according to time and define the calendar pattern or interval in calendar unit form.
2. In calendar pattern <*> is used to define any day or month. For example <dd,mm,yy> calendar pattern <*,01,04> represents any day of Month January & year 2004.
3. Since only the frequent item will play a role in the frequent pattern mining, so first scan is used to identify the set of frequent items
4. If the set of frequent items of each transaction can be stored in some compact data structure , it may be possible to avoid repeatedly scanning the original transaction database.
5. If multiple transaction share a set of frequent items, it may be possible to merge the shared sets with the number of occurrences registered as count. It is easy to check whether two sets are identical, if the frequent items in all of the transaction are listed according to a fixed order.
6. If two transactions share a common prefix , according to some sorted order of frequent items, the shared part can be merged into one prefix structure as long as the count is registered properly.

With the above observation , a Temporal frequent pattern tree can be constructed as follows :

First , a scan of DB drives a frequent list of items in schema <*,01,04> are {(f:3),(C:2),(b:2), (a:2),(m:2} same for calendar pattern <*,06,04> frequent items are {(f:3),(c:2),(a:2),(e:2),(m:2)(l:2)} and remaining items are infrequent, so skip those item .

Second, the root of the tree is created and labeled with "null". The FP- tree is constructed as follows by scanning the transaction database DB in second time.

1. The scan of the first transaction leads to construction of the first branch of the tree {(f:1),(c:1),(a:1),(m:1),(p:1)}& it follows the calendar pattern <*,01,04>.
2. For the second transaction, its temporal period is same, so it follows the same branch & the frequent item list {f,c,a,b,m,o}, shares a common prefix <f,c,a> and the count of each node along the prefix is incremented by 1. And a new node (b:1) is created and linked to child of (a:2), another new one (m:1) is created and linked to as the child of (b:1) and another new one (o:1) is created and linked to as the child of (m:1),
3. The third transaction, its time period being same as previous the transaction <f,c,b,o>, shares common prefix <f,c>. So f & c's count is incremented by 1, and a new node b is created although b is already existing but it does not go to that branch as it is not common prefix of the node. Node b is linked as a child of (c:2) and a new node o is created initializing the count because o is first time introduced on Temporal FP-tree and linked as a child of node <b:1>
4. The scan of forth transaction leader constructs another branch because its time period <*,06,04>does not match with existing branch's node time period. New nodes are created with <(f:1),(c:1),(a:1),(e:1),(m:1),(l:1)>
5. The scan of fifth transaction follows the time interval of forth transaction and it follows the same branch if the item prefix matches, It can share the common prefix <f,c,a,e> with existing path <f,c,a,e,m,l> and the count of each node is incremented by 1.
6. For the last transaction, <f,m,l> its time interval matches with second branch so it follow the second branch of FP-tree. Here it shares common prefix f, its count is incremented by 1 , a new node is created for item m , & it is linked to node f by initializing the counter value to1. For next item l again a new node will be created by initializing its counter value & it is linked as a child of node m.

3.3 Mining the frequent item from FP-tree

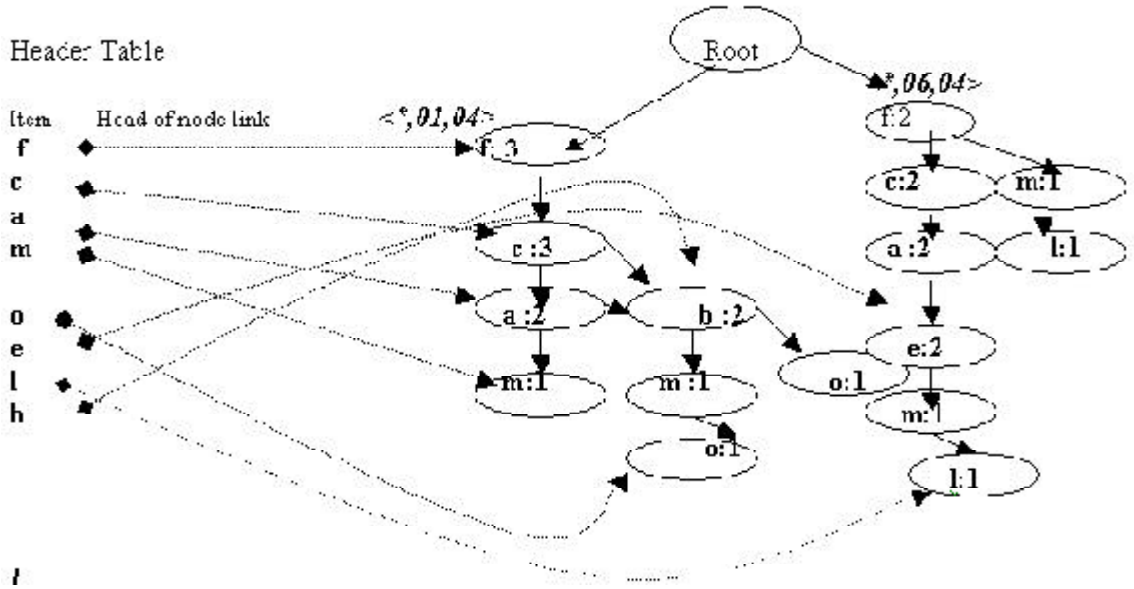


Figure 2 FP tree with different time interval

In Step(2) if (N.time = P.time) * then, it means that when a new node p appears in the FP-tree we check the time of transaction. If it is inside the time of transaction of N item, defined as per table below, then it follows the same branch otherwise a new branch will be created in FP-tree.

Table 2. Shows the path depend on the time of transaction in itemset.

Item name	***N.Time***	***P. Time***	***Branch***
F	<01,01,04>		First
C	<01,01,04>	<01,01,04>	First
A	<01,01,04>	<01,01,04>	First
E	<01,06,04>		Second
M	<01,06,04>	<01,06,04>	Second

Property [3] Node Link property . For any frequent item a_i, all the possible patterns contining only frequent items and a_i can be obtained by following a_i's node link's , starting from a_i 's head in the FP-tree header.

Example 2 : Mining process from the constructed Temporal FP-tree is shown in figure2. We examine the mining process by starting from the bottom of node-link header table.

For calendar pattern <*,01,04> for node m ,its intermediate frequent pattern is (m:3), and its path is <f:2,c:2,a:1,m:1,p:1> , <f:1,c:1,b:1,m:1> and <f:1, m:1>. Thus to study which appears together with m at time period <*,01,04> only m prefix {(fca:1) ,(fcb:1),(f:1) }, m's sub -pattern base , which is called m's pattern conditional pattern base.(which is called m's conditional Fp tree) leads to two different branches (fc:3) & (f:2).

For node a , its immediate frequent pattern is (a:4)and it is in two different path one for<*,01,04> & second for <*,06,04>. Calendar pattern <*,01,04> consist of <f:3,c:3 a:2> and <*,06,04> consists of <f:2,c:2>

Table 3 Mining Temporal frequent patterns by creating conditional (sub) pattern base

Item	**Time Interval**	**Conditional Pattern base**	**Conditional FP-tree**
m	<*,01,04>	{(fca:1) ,(fcb:1),(f:1) }	{f:2 c:2|m)
	<*,06,04>	{(fcae:1),(f:1)}	(f:2|m)
a	<*,01,04>	{(fc:2)}	(f:2|a)
	<*,06,04>	{(fc:2)}	(f:2|a)
0	<*,01,04>	{(fcabmo:1),(fcbo:1)}	(f:2,c:2,b:2|o)
	<*,06,04>	Nil	Nil
l	<01,*,04>	Nil	Nil
	<*,06,04>	{(fcaeml:1),(fml:1)}	(f:2|l)
e	<01,*,04>	Nil	
	<*,06,04>	{(fca:2)}	(f:2,c:2a:2|e)

c	<01,*,04>	{(fc:3)}	(f:2|c)
	<*,06,04>	{(fc:2)}	(f:2|c)
b	<01,*,04>	{(fcb:2)	(f:2,c:2|b)
	<*,06,04>	Nil	Nil
f	<01,*,04>	f	f
	<*,06,04>	f	f

From the Temporal FP-tree we get the conditional frequent pattern tree , which provides frequent pattern items by calling the procedure FP-growth[3]

4. Conclusion & Future work

We conclude that proposed algorithm gives an efficient time sensitive approach for mining frequent item in the dataset. Discovered rule is easier to understand. It uses divide & conquer technique for construction & traversing of tree which is used to decompose the mining task into a set of smaller task for mining confined pattern in conditional database which dramatically reduce the search space. is better than candidate generation. In fat Data mining concepts are applied where there are huge set of data available in data warehouse. It requires more scanning & processing time. Hence, after applying our logic of the scanning, this valid time & processing time can be decreased. It is very useful for retailer to create its own market strategy as per the requirement of time.

The work can be further extended by designing the tree that is hierarchy of time interval, if the retailer requires the associationship, it can be projected for that period of time.

5. Applications

- *Business Application :* This technique is most useful in Business mining. Most of the real world data have the time varying features. So the retails change their business policy with time to time for maximize the output, Time is here most important feature because some model of vehicle are not available from 1980s , suppose is currently appears in the market.Its history indicate no associationship but the fact is that product is not available on that period , so its associationship is started from the interval where it was valid.
- *Web Mining :* The concept can be applicable in web mining , In WWW the site which is no longer so its associationship also be no longer.
- *Clustering Problem :* This approach can be useful to solve the clustering problem, the cluster can be designed on the basis of period of data., that will reduce the size of data & processing time also.

6. References

1. R. Agrawal & R. Srikant (1994) “Fast algorithm for mining association rule.” In VLDB’94 Chile , Sept ,pp –487-499.
2. Juan M .Ale , Gustavo H. Rossi (2002) “ An approach to discovering temporal association rules”, ACM SIGDD March 1..21.

3. Jian Pei, Jiawei Han, Yiwen Yin and Running Mao (2002) "Mining Frequent Pattern without Candidate Generation", Kluwer online Academy.

4. Jiawei Han, Micheline Kamber , Book (2001) : "Data Mining Concept & Technique".

5. Yingjiu Li, Peng Ning, X. Sean Wang , Sushil Jajodia (2003) "Discovering calendar- based temporal association rules", Data & Knowledge Engineering volume 4 pp– 193-214 ,2003

6. Geraldo Zimbrao, Jano Moreeira de Souza, Victor Teieira de Almeida, Wanderson Araujo da Silva (2000) "An algorithm to Discover Calendar –based Temporal Association Rules with Item's Lifespan Restrictions"

7. Banu Ozden , Sridhar Ramaswamy , Avi Silberschatz (1998) "Cyclic Association Rule" ,In Proc. Of forteenth International conference on Data Engineering pp 412-425

8. John F. Roddick, Kathleen Hornsby, Myra Spiliopoulou (2000) "An Updated Bibliography of Temporal, Spatial, and Spatio-temporal Data Mining" Research. TSDM 2000: pp147-164.

9. *Chris Giannella_, Jiawei Hany, Jian Peiz, Xifeng Yany, Philip S. Yu* (2003) "Mining Frequent Patterns in Data Streams at Multiple Time Granularities", pp 191 – 210, H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha (eds.), Next Generation Data Mining.

10. *Chris P. Rainsford, John F. Roddick (1999)* " Adding Temporal semantics to association rule", *3rd International conference KSS Springer 1999, pp 504-509*

11. *Claudio Bettini, X. Sean Wang (2000)*" Time Granularies in databases , Data Mining , and Temporal reasoning*" , pp 230, ISBN 3-540-66997-3.*

About Authors



Ms. Keshri Verma is a lecturer in School of Studies in Computer Science, Pt. Ravishankar Shukla University, Raipur, Chhattisgarh.
E-mail : seema2000_2001@yahoo.com

Mr. O P Vyas is working in School of Studies in Computer Science, Pt. Ravishankar Shukla University, Raipur, Chhattisgarh.
E-mail : opvyas@rediffmail.com

Platform Independent Terminology Interchange Using MARTIF & OLIF

M Ramshirish

Abstract

In the present society, information is treated as one of the most valuable resources and its importance has increased manifold in the networked digital environment. No doubt, the developments in the field of IT have facilitated the sharing of information from one knowledge domain to another, but the difference in terminologies acts as communication barrier and impedes the free flow of information. Various terminological and lexical databases have been developed to help the professionals; however, the problem of data interchange/ exchange across different platforms still persists. This paper gives an overview of MARTIF and OLIF, the tools for platform independent terminology exchange/ interchange.

Keywords : Terminology Database, Localization, Terminology Interchange, MARTIF, OLIF

0. Introduction

Wide spread growth of human knowledge and the growth in the field of Information Technology has led to the preparation of many databases of terms and lexicons. Development of term databases and lexicons is time consuming and a costly affair. Moreover, differences in software, hardware, and methodology further complicate the interchange of common terms across databases. To find a solution to these problems and to make the web of chaos “a semantic web”, the idea of exchange/interchange of terminology and lexicons came up.

Terminology is a set of terms used by the subject specialists and experts in their respective specialized areas. It provides base from which original technical texts and translations are prepared. Technical terminology is useful not only for translators and technical writers as document producers, but also for others like teachers who must be able to guide their students in acquiring the specialist knowledge.

MARTIF (Machine Readable Terminology Interchange Format) and OLIF (Open Lexicon Interchange Format) have been developed to solve the problems of platform-independent terminology interchange between different databases.

1. What is Terminology

In any subject, the subject specialists use technical terms and other terms that are related to that particular subject, these specialized terms constitute the terminology.

Terminology is involved in one or the other way, whenever and wherever, specialized information and knowledge is created, communicated, recorded, processed, stored, transformed, re-used.

Terminology can be defined as “a structured set of concepts and their relations in a particular subject field. It can also be considered as the infrastructure of specialized knowledge”. [1]

1.1 Why Terminology

Due to increase in the R&D activities, rapid growth of Internet, literature, general people as well as researchers have seamless access to a variety of information from different sources. But, if people fail to understand the information and data, then this may lead to chaos.

To solve this problem, need was felt to develop terminologies related to a particular subject area where the terms will help one in understanding the concept and relations between different facets of a subject.

1.2 Terminology Databases and their Significance

In any subject field, new terms are created and also some terms become obsolete. The conceptual meaning of terms also may change with time and developments that take place in a subject field. Some terms have very precise meaning in some situation, and after many years they are used in a different meaning associated with old usage [2]. E.g.: - The term Ontology was used earlier in philosophy to describe the nature of existence, but now it means construction of knowledge models with specific concepts or objects, their attributes, and inter-relationships.

So, terminology databases were developed to overcome the above problems; the databases attempted to rationalize the use of terms in technical language. In them, confusing terms such as homonyms were avoided to facilitate exclusive usage of terms in a very particular subject area and to achieve vocabulary control as a result.

It is impossible to think of technical writing and technical documentation without the knowledge of proper terminology. Terminologies used in a specialized subject area contain not only terms but also formulae, symbols, even drawings where communicating potential of the document lies in the terminology used. Adequate use of terminological resources increases quality and productivity for documentation. Further, we need suitable terms to get exact meaning from the original document for translation of any document from one language to another.

1.3 Users of Terminology Databases

Some standard format is needed to ensure reliable and qualitative exchange and for facilitating interoperability. Different groups of professionals use different terminology databases for various purposes.

They can be divided to following broad categories [1]: -

Translators

Technical writers

Information managers

Professionals dealing with controlled vocabularies tools like thesauri, classification Systems, etc. for documentation and information retrieval

They use the terminology databases for [3]: -

- ✍ Data capture and presentation – i.e., enter, store and review the concepts. Information integration, indexing, retrieval, decision support, linking records. Messaging between software systems i.e., linking different information systems.
- ✍ Reporting.

1.4 Need of Terminology Interchange [4]

In order to exchange expert information and prevent duplication of efforts, users of terminological databases need interchange formats, as preparation of terminological database is both time and money consuming.

The terminology databases are implemented in different formats and they are prepared to run on different operating systems, requiring sometimes-different file format and sign conventions. So, a standard format is needed for terminology interchange for preparation of data interchange tools.

A common format has to be defined for terminology export and import to provide a channel for terminology interchange. Possibly terminology interchange could be managed without any structure, i.e. an unstructured text file, but this leads to the problem of reformatting the unstructured data manually; a time and cost intensive undertaking.

2. Need of Standards for Terminology and Lexicon Interchange

The exchange of database records needs to be done carefully because structure of terminological records varies considerably from one database to another. In addition to this, even the designs and user needs vary, so, here a universal interchange format is essential to make interchange easier.

But the language-processing tools are not well integrated and interoperable. Terminology databases, translation memory systems, controlled English systems, machine translation systems lack seamless integration. The desired integration, once achieved, would increase productivity among translators, terminologists and other workers. So, the tools for terminology extraction, terminology consistency checking and translation workflow are needed. [5]

To solve these problems, various standard formats (including MARTIF and OLIF) evolved for modeling and representing the terminological data.

3. MARTIF & OLIF

3.1 What is MARTIF

MARTIF stands for Machine Readable Terminology Interchange Format. It is a SGML based format to facilitate the interchange of terminological data among Terminology Management systems [4].

Its origin can be traced back to the development of a powerful tool for terminology interchange, done in cooperation with Text Encoding Initiative (TEI) and Localisation Industry Standards Association (LISA) [LISA is an association of companies and institutions working on the translation and adaptation of software into different languages and TEI is an initiative for encoding texts with a relevant structure and semantic information]. The goal of this cooperation was to produce a format that would be a platform independent and publicly available format. The resulting format was MARTIF, which is also known as ISO (FDIS i.e., Final Draft International Standard) 12200 (which is in turn based on ISO 12620 [ISO 12620 is designed to promote consistency in the storage and interchange of terminological data through the use of a standard set of data categories for term entries]).

3.2 Evolution of MARTIF [6]

Modern developments in Terminology exchange field can be seen from an earlier format called MATER (ISO 6156) (Magnetic Tape Exchange for Terminological lexicographic Records (1987), which was published only after magnetic tape become obsolete. MATER developed to MicroMATER (designed to meet the needs of the then new PC architectures). The various versions, which came up in due course are MARTIF part 1 (Negotiated Interchange) and MARTIF part 2 (Blind Interchange).

In the first approach, two partners use a common framework for interchange and negotiate details within the framework of the intermediate format to allow writing export and import routines that preserve as much information as possible. Presently work is going on in the area of second approach i.e. MARTIF part 2 (1998) (Blind Interchange). In this, the details are predefined, so, export and import routines can be written without knowing who the other interchange partners will be (i.e., one need not 'see' the interchange partners).

3.3 Purpose of MARTIF [7]

To provide a universally applicable format for the negotiated interchange of structured terminological data among various applications, system environments, and hardware platforms.

To be used with terminological data that can be stored, read, retrieved, and manipulated by a computer (ISO DIS 12200.2:1).

3.4 Features of MARTIF [8]

MARTIF is concept-based approach rather than non-concept oriented approaches to terminology e.g., lexicographic and NLP approaches.

It allows directional change in using the database or in importing the data and readily allows for the interchange of data to or from other data models where different languages are given preference.

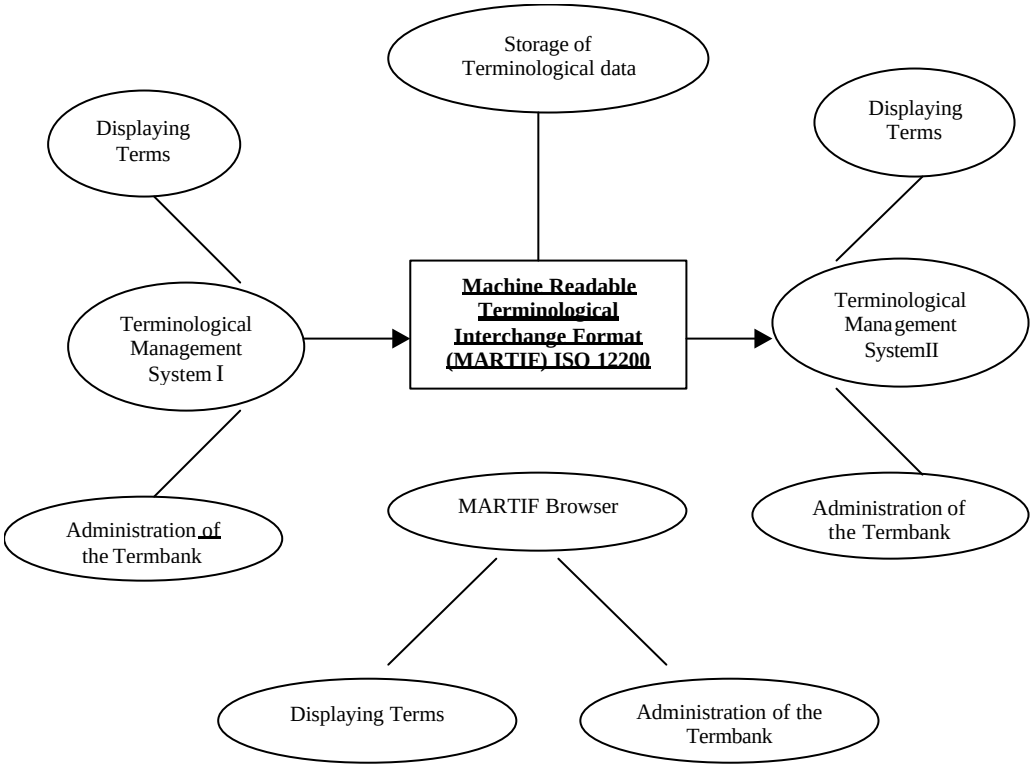
3.5 MARTIF Specifications

Categories for MARTIF are divided into ten sections, which are grouped into four classes. The four classes with their sections are [4]:

1. Term: Section(1)- consists of the data category term;
2. Term-related information: Section (2)- consists of term-related information (such as POS, etymology, term type, ...) and section (3)- on the degree of equivalence (how close two terms are related);
3. Descriptive information: Section (4)- relation to the domain, section (5)- descriptions of the concept (i.e. definitions, examples), section (6)- relations between concept entries, section (7)- data categories that relate a concept entry to its position in the concept system, section (8)- general notes;
4. Administrative information: section (9)- consists of data categories relating a concept entry to a node in a thesaurus or to other forms of documentation, section (10)- has data categories which contain administrative information (update information, author, etc.).

3.6 Usage of MARTIF

The usage of MARTIF can be depicted in following diagrammatic manner



4. What is OLIF?

It can be defined as Open Lexicon Interchange Format, it is an XML compliant standard [9].

A lexicon can be defined as

1. The collection of words in a language
2. It is a special list of terms and related terms used for subject searches.
3. It is a linguistic tool with information about the morphological variations and grammatical usage of words.

4.1 Puspose of OLIF

OLIF allows the transfer of terminological and lexical data between or from different translation tools, including NLP systems such as Machine Translation as well [10].

4.2 Origin of OLIF

OLIF has its origin in the Open Translation Environment for Localization (OTELO) project, which was funded by European commission sharing lexical resources [10]. Members of OLIF Consortium designed it and are the maintaining organization.

The OLIF consortium and SALT group (Standard based Access Service to multilingual Lexicons and Terminologies) are collaborating with each other. Here, SALT aims to create a lexicon exchange format and SALT focuses on the terminological side of the format (in the tradition of MARTIF), while OLIF focuses on the lexical side.

OLIF and SALT's XML based formats for Lexicon and Terminologies (XLT) define a common set of data categories to enable integration between OLIF and XLT. It is an organization of major NLP technology suppliers, corporate users of NLP and research institutions, Systran, Logos, Sail Labs, IBM/Lotus, Lingua Tec, Pa Trans, Trados, Xerox, German Research Centre for Artificial Intelligence, IAI and Others etc.

4.3 Features of OLIF [10]

OLIF is more practical and tries to accommodate existing lexical resources while other lexicon projects are more research oriented, focusing on elaborated lexical descriptions.

OLIF concentrates on Lexical exchange rather than terminology, and leans more towards pragmatic than theoretical or research based projects.

The basic idea of OLIF is to facilitate the exchange of primarily the pivotal information in lexical entries. It also provides the option of deeper lexical representation included in the OLIF format

It offers the user a mechanism for encoding the information in a general way that allows portability.

4.4 OLIF Specifications

The basic unit in OLIF is uniquely defined by a set of key data categories: Canonical form, parts of-speech, language code, subject area, and, in the case of homonyms a semantic reading etc [11].

5. Conclusion

MARTIF & OLIF are not the only formats which facilitate the terminological and Lexical Interchange, other formats also exist, but they are not powerful enough to handle the complexity of the problem. MARTIF uses SGML structures in order to preserve and communicate the information and functionalities present in many terminological databases whereas OLIF uses XML. The standards are no doubt complex for the end users to understand but the complexity is required to facilitate terminology exchange among different databases.

6. References

1. Galinski, Christian & Budin, Gerhard.Terminology. http://cslu.cse.ogi.edu/HLTsurvey/ch12node7.html (accessed on 15/1/04)
2. Electronic Dictionaries DicoBase. http://www.linga.fr/LingEn/DicoOffreEn.htm. (accessed on 25/10/2004)

3. Rector, Alan. Why do we need Medical Terminologies? http://www.cs.man.ac.uk/mig/links/RCSEd/terminology-why.htm (accessed on 09/12/2003)

4. Trippel, Thorsten. Terminology interchange: Facing multiple requirements. http://coral.lili.uni-bielefeld.de/~ttrippel/terminology/node82.html#SECTION008400000000000000000 (accessed on 26/09/2004)

5. Warburton, Kara.Results of the LISA Terminology survey. http://www.lisa.org/2001/termsurveyresults.html (accessed on 08/11/2004)

6. Robin Bonthrone, Fry & Partnerschaft, Bonthrone. MARTIF Lite: User-driven Terminology Interchange. http://www.lisa.org/archive_domain/newsletters/1998/1/bonthrone.html (accessed on 30/09/2004)

7. The CLS Framework:Negotiated Sharing. Introduction to ISO 12200 (negotiated MARTIF) http://www.ttt.org/clsframe/negotiated.html (accessed on 26/10/2004)

8. MARTIF putting complexity in perspective. http://www.ttt.org/clsframe/termnet1.html (accessed on 30/09/2004)

9. www.olif.net (accessed on 19/10/2004)

10. Lieske, Christian.Cormick, Susan Mc & Thurmair, Gregor. The Open Lexicon Interchange Format (OLIF) comes of Age. http://www.eamt.org/summitVIII/papers/lieske.pdf (accessed on 29/09/2004)

11. OLIF. http://www.w3.org/2002/02/01-i18n-workshop/OLIFExample.xml (accessed on 08/11/2004)

12. MARTIF. http://www.creativyst.com/cgi-bin/M/Glos/st/GetTerm.pl?fsGetTerm=802.1b (accessed on 07/10/2004)

13. Terminology Interchanges. http://korterm.org/research3-e.htm (accessed on 07/10/2004)

About Author

Mr. M. Ramshirish is working as Librarian at RRG E-Media, Ramoji Film City, Hyderabad. He has done ADIS from DRTC, Indian Statistical Institute, Bangalore and BLISc. from Osmania University, Hyderabad.

E-mail : ramshirish@yahoo.co.in or mramshirish@yahoo.co.in

Digitization : Basic Concepts

B Mini Devi

Abstract

The introduction of digital libraries is changing not only the face but whole body of the libraries around the world. In a global village the concept of digital library is of great importance. Hence the process of digitization. The author discusses the steps involved in the process of digitization, types of materials to be digitized and the problems that the libraries are facing on this respect.

Keywords : Digitization, Digital Libraries, Content Management.

0. Introduction

The rumbling change that shook each and every field of library and information science is the emergence of digital libraries. In a country like India, rich in cultural, spiritual heritage indigenous research and development in science and technology, humanities, the preservation of her ancestral information sources becomes the duty of librarians and information scientists. The growth of internet, communication channels, information technology and digitization of documents are revolutionizing the traditional concept of library. Any user from any part of the world can access the information he wants in a digital environment.

1. Digital Library

According to the Berkeley Digital Library Project, University of California[1] , the digital library is be a collection of distributed information sources, producers of information make it available, and consumers find it perhaps through the help of automated agents"

The Information Infrastructure Technology and Applications (IITA) working Group considers "digital libraries as systems providing users with coherent access to a very large, organised repository of information and knowledge"[2]. According to Association of Research Libraries 'digital libraries are not a single entity, require technology to link the resources of many, linkages are transparent to the users, permit universal access, not limited to document surrogates but extended to digital artifacts'. Some consider digital library as 'library without books' or 'library without walls'. The synonyms to digital library are 'virtual library' or 'electronic library'.

The purposes of digital library are[3]

- Collect, store and organise information and knowledge in digital form.
- Promote economic and efficient delivery of information.
- Put considerable investments in computing/communication infrastructure
- Strengthen communication and collaboration between research, business, government and educational communities.
- Contribute for lifelong learning opportunities

2. Materials to be digitized

The priority of materials to be digitized depends upon the type of library, category of users, finance, infrastructure, etc. However the following materials are to be given priority.

The scientists create a large quantity of information that are documented in the form of research reports. They are the authentic primary scientific data which is of great value to the coming generations. These reports form the basis of further research work.

2.1 Research Reports

As research publications are the result of original scientific investigations, the archiving of them needs utmost priority.

2.2 Journals

The articles published in journals in different subjects are to be traced and brought under one roof for archiving. In the field of Library and Information Science publishers from Asia work together as a group to publish their journals on major Website databases. As a result these journals get a wide publicity and usage. Recently Ulrich International Periodical Directory had selected 29 journals from India for digitalizing by taking into account of the standard and coverage of articles published by them.

2.2 Cultural heritage documents of the country

The preservation and archiving of the cultural heritage documents of our country are of great value to our future generations. Now-a-days this is one of the crucial issues which librarians of our country are facing. This may be due to the physical conditions of this type of documents, which are damaged or decayed due to various factors such as age, quality of paper, binding, insects, climate, handling etc.

2.3 Theses and dissertation

Theses and dissertations form a rare and valuable collection of many academic libraries. The digitization of them results in wide publicity and duplication of research can be avoided.

3. Digitizing

Digital libraries offer access to contents over computer and communication networks. In present days due to the escalating price, journals are out of reach to ordinary people. The paper used for printing these journals becomes brittle as time passes. In the case of books also the above situation exists. Moreover, the torn out, brittle and dusty documents create problems in maintenance and the health of librarians as well as the users. Here comes the importance of digitization which not only enhances the life of these documents but also provides easy access to wide audience with exhaustive search engines, and effective bibliographic control.

4. Methodology

- Content Searching and Selection
- Scanning
- PDF Creation & OCRing
- Content Indexing and Metadata
- Information Retrieval Procedure

4.1 Content Searching and Selection

The first step is to identify which materials are to be digitized and which are not to. The utmost priority should be given to the following factors

- policy of the library
- needs of the user
- type of document

4.1.1 Policy of the library

Now-a-days libraries give priority in digitization of their documents. In the case of a Science and Technology library, the digitization of research reports and journals should be done first because they are the result of original scientific investigation. On the other side in the case of a Social Science library, documents of historical importance should be given first preference. But in an academic library, research reports, theses, journals are to be selected first.

4.1.2 Needs of users

As users come from different cross section of the society, their needs are also varied. In a S & T library, scientists form major user community and their priority should be on research reports and journals. In a Social Science library, the user community constitutes mostly historians, literary writers and their importance will be on historical documents and literature books. But in an academic library, text books, journals, theses and dissertations are searched first by students, teachers and research scholars.

4.1.3 Type of document

Costly and rare books are to be protected from damage, multiliation and loss. Documents of old, cultural and historical materials are to be given prime importance- as they form part of history of our country.

4.2 Scanning

The fundamental conversion technique is scanning of the document. This can be done by sampling the image of the document on a grid of points. Each point is represented by a brightness code ie in black/ white colour. As in photographic work, a very high resolution is not needed in this case. Very good images can be created with a resolution of 300 dots per inch. The scanners take care of quality of paper of the document and spot-marks.

4.3 PDF Creation and OCRing

Portable Document Format (PDF) created by Adobe is a better format for storing page images in a portable format. PDF is the most popular page description language used today. A PDF document consists of pages, made of text, graphics and images, and supporting data. PDF can supports hyperlinks, searching etc. PDF can store bit-mapped images and Adobe provides optical – character-recognition software for the creation of PDF files.

The way to generate a PDF file is to divert a stream of data, then the file can be converted from post script or from another format, stored, transmitted over a network, displayed on a screen and then printed. The PDF thus created are batch optimised so that maximum files are occupied in minimum space and hence searching become quicker.

Optical Character Recognition is the technique of converting scanned images of characters to their equivalent characters. Here a computer program separates out the individual characters and then compares each character to mathematical templates.

4.4 Content Indexing and Metadata

Indexing should be done for easy retrieval of the information. The Dublin core is usually used for the purpose. From 1995, an International Group led by Stuart Weibel of OCLC has been working to device a set of simple metadata elements that can be applied to a wide variety of digital Library materials. The set of elements developed by the group is known as the Dublin core.

The fifteen elements constitute the metadata set of Dublin Core are title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage and rights. There are two options in Dublin core, (i) minimalist option to meet the original criterion of being usable by people and (ii) structured option which is more complex requiring special guidelines. Storing the metadata and data together is convenient for long term achievement since computer programs have access to the data and to the metadata at the same time. For an HTML page, the attaching of the metadata is done in the page by using special HTML tag<meta>, which comes from an HTML description of the Dublin Core element set. In file types other than HTML, Resource Description Framework (RDF) are developed by World Wide Web consortium.

4.5 Information Retrieval Procedure

Information can be retrieved from the file by asking a query, a search term. The query may be a single search term, a string of terms, a phrase in natural language etc. Full text-search – facility is also there. A Boolean query consisting of two or more search terms related by AND, OR, NOT can also be used.

5. Problems Facing Digitization

Even though libraries and librarians all over the world are marching towards digitization, there exist some constraints in the process and their maintenance. The problems facing digitization are

5.1 Longevity of Storage media

Many of the storage media praised by people all over the world may become less useful only long after they become unreadable. Thus documents digitized and stored in such media become useless and their maintenance will be more difficult than print media. The digital archival media today used are magnetic tapes, CD-ROM discs and DVDs. From the scene magnetic tapes disappeared because of their short life due to demagnetization, material decay and oxidation.

During 1980's CD-ROMS emerged into the field and boasted of a longer life span of 30-100 years. Now a days most of the CD's go to the way of 51/4 diskettes. DVD having several standards pushed CD's behind the screen. The changes and improvements of storage medium put serious questions about the future of digitized materials and their alteration.

5.2. Technology obsolescence

The technology behind digitization is undergoing drastic changes continuously. The computer hardware, software, storage media etc are undergoing great revolution. The digitized materials become unreadable if the background devices become obsolete as time passes by which ultimately results in the loss of

data. Like print media, digital media is also affected by light, heat, moisture, insects, acid content and air pollution. Digital storage media are always under the threat of above factors. While selecting the storage medium, technological obsolescence should be taken into consideration.

5.3 Migration

The periodic change of digital systems from one configuration to another to overcome the problems caused by technological obsolescence is termed as migration. Migration to a new storage system is more expensive and this will ultimately result in the loss of data.

5.4 Selection of Documents

In an age of information explosion and information pollution, librarians are in a dilemma about 'what type of records are to be digitized' and 'what type of records not to be digitized'. The documents in high demand today may become obsolete even tomorrow because of the vast developments in the subject and printing and publishing industry. A digitized document deselected from the collection is lost for ever. To overcome the problem, librarians should seek the advice of subject experts in each field and users of the library about the importance of each and every record and from this list selection of records for digitization can be done.

5.5 Copyrights

The issues regarding copyright raise serious matters before librarians in digitization. Research scholars usually include graphs, data from books and journals without prior permission of the author. In a digital library users are always demanding back issues of journals and rare historical archives for which the library has no copyright. This may lead to serious dissatisfaction about digitization among users. As a final solution to this matter, librarians must be given permission to digitize copyright works in connection with digitization.

6. Use of Digitized Materials

The typographic standards ie titles, headings, subheadings, typefaces, paragraphing and folios, followed by printed sources are lacking in digital information. The users give utmost importance to the above factors. The reading of a book from online is time consuming, laborious task, causes several problems to the health of the user, requires more money for downloading and printing it. Eventhough writers and reading community are charging the society that 'reading is dying', in reality this can only be taken as an allegation in the digital environment. Actually, serious readers always prefer printed version and digital information only as a supplement and not as a substitute. The concept of 'digital divide' is of special mention at this juncture.

7. Conclusion

The digitization of collection of a library opens its doors to the world so that local collections get a wider exposition. In the field of Science and Technology, there is emergence of interdisciplinary and multidisciplinary subjects and research reports. Articles are being published in science journals in a huge amount than in the parts. The escalating price of the journals are not affordable to each library. The emergence of E-journals and digitization of journals abstracts and indices reduce the burden of their procurement and save storage of space. Although there are drastic changes in digital technology, finance, staff training, manpower, infrastructure etc are serious problems to be tackled before libraries attempt for digitization. In a country like India having great history in traditional medicine, ancient art, culture, architecture, etc, the information that our great ancestors gave us through inscriptions, archives, and through rare books is to be digitized for our future generation.

8. References

1. Communications of the ACM, 38 (H), 1995, P. 59 – 60.

2. Lynch, Clifford and Gracia – Molina, Hector. (Eds.) Inter-operability scaling and the digital libraries research agenda, 12 August 1995. A report on the May 18-19, 1995 IITA Digital Libraries Workshop. (http://ww-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html).

3. http://sunsite.berkeley.edu/ARL/definition.html (accessed on 8.4.2003).

About Author



Ms. B Mini Devi is currently working as Technical Assistant in the University Library, University of Kerala. She holds MSc (Botany) MLIS and has over 10 years professional experience. She has also qualified the UGC NET for JRF & Lectureship and is doing research in the field of Informetrics.

Features in the Web Search Interface: How Effective Are They ?

Deepak P Sandeep Parameswaran

Abstract

With web search getting to be more and more popular, and predictions that they would be even more popular in the coming years, given the exorbitant growth of the web in recent years, search engines, in their quest to be branded the best, have regularly been providing additional features. The users of web search interfaces are typically diverse and have wide ranging interests. This is in contrast to other application interfaces which cater to a specialized group of people. In this study, we examine the influence the interface manifestation of such features has, in their usability and effectiveness.

Keywords : Web Search, Web Search Interface, Search Engines.

0. Introduction

It has been acknowledged [1] that web application interfaces are fundamentally different from common GUI based applications and hence have to be designed taking a lot of additional factors in mind. Further, web searches are directed towards fulfillment of widely varying requirements [2] by users of widely varying backgrounds that using assumptions on the awareness and the knowledge level of the typical user to optimize the interface may often turn out to be harmful. With the competition among search engines for popularity escalating by the minute, more and more search engines provide value-added features [3] that aid the web user. It has been opined that feature addition to web search (interface) is hard [4], given that added features are often seen by the hasty user as more of a nuisance than an aid. In this paper, we propose to evaluate how effective the add-on features in web search engines are (in usability), by means of an objective evaluation and a survey on a pool of users who use the Google [5] search engine, which is, by far the most popular search engine.

Section 2 reviews the more relevant works in web search interfaces. Section 3 is a study on the features that are present in contemporary search engines, focusing almost entirely on their interface manifestation and usability. It further enumerates the way the features can be incorporated in the search engine interface, how they could be placed to aid the user in the search, and a list of the more common features present in contemporary search engines. Section 4 presents a survey conducted to evaluate the usability and effectiveness of the features in the web search interface. Section 5 lists the conclusions derived from the study and generalizations from the opinions gathered in the survey. The list of references forms Section 6.

1. Web Search Interfaces

Web search interfaces have undergone a drastic change through the years. The initial text-box only interface, has almost completely returned back after being replaced unsuccessfully, temporarily and partly by a graphical interface [4]. In a web application like the search engine, the user typically does not devote his entire attention at a single application for long periods of time. Rather, he switches back and forth between applications and web sites. Thus, response time and quality of results are very critical issues for a search engine, as they decide as to whether the user would return to it. Although not as important as the above two factors, the interface design is also a major issue for the search engine. A search engine that provides a load of features, must ensure that they do not clutter up the screen real estate, making the search engine difficult to use. One among the major guiding principles is that the

single text box, the place where the user has to enter the search query should be center of attention in a web search interface. Web application interface designers need to understand, accommodate and support the extensive freedom that a user enjoys while using a web application [1]. Sometimes, we can force users through set paths and prevent people from linking to certain pages, but sites that do so feel harsh and dominating.

Web search engine users are casual users, who due to some reason or intention, approach the application to get a URL to a web page that they want to visit. Thus, the web search engine is a means to the end. The importance of the search engine interface has been acknowledged even in very early days [6]. A study conducted way back in 1995 [6] evaluating the interface of the then search engines, branded OpenText as the best citing its “powerful search interface”. Further, they opined that WebCrawler had the easiest-to-use interface for novices. Another study [7] concludes that the best search engine would be the one that can produce “accurate results from easy-to-use interfaces”. It has been postulated [8] that “interfaces and documentation” are the two most important factors that affect users efforts in learning and using a search engine.

The usability of computer systems are often classified based on five components [9]: learnability, efficiency, memorability, errors (error rate) and satisfaction. Of these, the most important factors that concern search engine interfaces are learnability and memorability. Web search interfaces have to be instantly learnable as the users are mostly casual users. Interfaces that are easiest to learn usually turn out to be the most memorable as well. The same study [10] concludes that among the specialized interfaces for web search engines, the “connector menu interface”, which consists of a series of text boxes with the option of connecting the words by boolean operators is a very usable and preferred one.

The main usability questions to ask when reviewing a search vendor or evaluating the user experience of your own search solution are, “Can you customize the results” and “Can you simplify the search interface”. These two things matter for ease of use and value to the user. Search results are as important as the interface [11]. These arguments, postulated keeping site search in mind, also apply equally well to web search.

A review classifies search interfaces into 4 classes [12]. A simple search button is regarded as a “passive” interface. A “standard” interface has an edit box, a search button and a link to a more specialized kind of search. Surfacing interfaces, more prevalent in site search, includes a combo box, which contains the different filters or categories for search. A qualifying interface, which also includes the combo box contains non-category filters like time or location filters for narrowing down the search. Many contemporary search engines use the language filter to direct searches towards documents in a specified language.

2. Features in Search Engines

As already opined, the ever-escalating competition among search engines to attract user attention and to draw more audience, has led to search engines providing much more [13] than a single text edit box in their interfaces. Most of these features are aimed at refining the search to provide the user with better relevant results. The web search is typically described as a process which includes stages such as query formulation, deployment, a systematic review of the search results followed by possible query reformulation and re-deployment (perhaps, multiple times) [14]. The usual features that search engines provide are those that aid searching and speed-up convergence to the relevant results. But, of late, search engines have ventured into providing features, which are unrelated to the searching process (e.g., the calculator, web page translator and stock qoute features of Google [15]).

2.1 How do search engines provide additional features on the interface ?

Incorporating new features to an application interface is often a tough and challenging task. We might have to undo certain careful decisions taken earlier regarding the structure of the interface and perhaps, might even have to redo the entire interface. As already mentioned, add-on features to the search engine can either be totally unrelated to the search process or can aid the search process. Whatever be the type, there are commonly two options to incorporate it in the interface

- Include it as a link, explicitly in the interface
- Introduce a new operator for the purpose and include documentation for the same in the 'Help' page

The first option is often unwelcome as it uses up screen real estate. Such incorporation of features often aid in cluttering up the page rather than aiding the user [4]. Yet certain important features such as the introduction of the service in a vernacular language are often presented this way [16]. The second option is perhaps the most widely used option. The most widely used illustration of this option is the filetype filter, whereby a user who requires results of type pdf only can append a "filetype:pdf" string to the search query. But users are often unaware of such features, as the presence of such a feature is not evident on the main search page. Flash alerts on the main page just after the introduction of a feature would however, aid in spreading the news to the regular users.

2.2 Where (in the search process) do search engines provide additional features ?

This section is concerned only with features that aid the searching process. Features may aid the searching process during its various phases

- Query Formulation/Query Creation: The surfacing and qualifying interfaces for search engines [12] as discussed in Section 2 aid in imposing filters before the search query is actually presented to the system
- Search Result Review/Selection of a Search Result: The search results are usually taken to be the set of URLs returned by the system. But seldom do search engines provide such plain results. Almost all the search engines of today provide short descriptions of each search result along with the search result and some even cluster search results [17].
- Refinement of the Search: The help in this phase is usually provided by means of queries like, "do you want to see the omitted results", "similar pages" etc [5].
- Reformulation of the query: The help here is provided by means of features such as the spell checker of Google [5] (the "did you mean" feature)
- Examining Selected Item: Help can also be provided when the user is examining a result. Google provides "search query highlighting" in the cached copies of results
- Other situations: Google automatically provides an option to view the (old) cached copy of a page in cases when the result page is no longer alive. Such a condition is in fact, a failure on the part of the search engine to provide recent results.

The additional features can be provided on

- The main search interface OR
- Results pages

Evidently, only features that aid query creation or formulation can appear in the main search interface. They are often duplicated in the results pages as well, given that the phase of reformulation of the query has to be given at least as much support as the initial query creation.

2.3 Features in the interfaces of contemporary search engines

Here we would like to enumerate a list of the more common features in contemporary search engines and examine how they are represented in the interface. We restrict the investigation to features that aid the search process directly. This is, by now means, an exhaustive list, but ventures to provide a taste of the common functionalities of the current search engines.

2.3.1 Features that aid query formulation

- Surfacing interfaces: A combo box is provided to narrow down the search to a particular category. Yahoo [18] provides a combo box by the side of the search box with options such as searching in yellow pages, images, products etc. Google [5] also provides a category specific narrow down option to categories such as images, groups and news, but by means of specific links to separate sites. Other filters are provided in the advanced search interface as well.
- Qualifying interfaces: These aid narrow down by time and location filters. A lot of contemporary search engines have these filters in their advanced query interfaces. e.g., site search, inward link search etc.
- Other filters: Filters can be provided by means of specialized operators as well. There are often operators to restrict the search to results of a particular type (filetype:), to documents that have the query in the title (intitle:) or URL (inurl:) etc. These are advantageous from the interface designers point of view in that they do not consume real estate in the screen (refer Section 3.1)
- Google has a feature [15] (“I’m feeling lucky”) which directs the user to a single result page which presents itself as a separate button in the main interface.

2.3.2 Features that aid (or complement) search result review/selection

- Descriptions: Almost all search engines provide descriptions for each search result, which may either be derived from the pages themselves or from pages with inward links to the page in question.
- Context Clustering: Vivisimo [17] provides result clustering on the fly, so that the user is provided with groups of search results. The visual appeal of such clustered results can often be inviting. Algorithms for such context clustering are available in literature [19].
- Resources: Teoma [20] provides a collection of relevant links from “experts and enthusiasts” along with the search results. It is provided in a separate column, and hence does not clutter up the results page too much.
- News headlines: Of late, Google [5] provides any relevant news headlines along with the search results. Although this does not aid the search process, it does complement it. The news headlines take up less space compared to the results themselves, given that they do not have descriptions.

2.3.3 Features that aid search refinement

- Similar pages: Google [5] provides a similar pages link with almost every result page, to refine the search to pages that are similar to it (represented by an explicit link with each search result)

- ✍ See Omitted Results: Google automatically removes similar pages from a site to avoid redundancy in search results. But it provides an option to view the omitted pages in order to refine the search towards the direction of a specific site (implicitly).
- ✍ Context Clustering: This also acts as an implicit search refinement towards different clusters of results.
- ✍ Refine: The refine feature of Teoma is best discussed under the next section.

2.3.4 Features that aid query reformulation

- ✍ Refine: Teoma [20] provides a listing of possible query refinements or reformulations which could be used to narrow down the search towards a specific direction.
- ✍ "Did you mean": The did you mean feature of Google lists possible reformulations of the query to alleviate spelling mistakes and typo errors.

2.3.5 Miscellaneous Features

- ✍ Cached copy: As the current search methodology searches on indexed collections rather than the current state of the web, the results can be stale. Such imperfections are usually masked by providing a cached copy.
- ✍ Search Query Highlighting: A further advantage of providing the cached copy is that the search engine can provide features such as search query highlighting in the cached copy of the results (as they are held entirely by the search engine)
- ✍ Web Page translation: The user who requires results in a particular language can be satisfied by means of translating the results originally in different languages, to the language in question. Yet, this is often regarded as something totally unrelated to the search process.

As can be seen from the above, most of the additional features introduce one or more links in the interface. Thus the usability of such features, given that screen real estate is hot property in search engine interfaces is a matter of interest. The following sections describe a survey on the usability of additional features in a search engine.

3. A Survey on the Usability of Additional Features in Web Search Engines

3.1 Choice of the search engine

We, in our attempt to restrict ourselves to questions such as "how do you find that feature", chose to restrict ourselves to a particular popular search engine. The obvious choice was Google, which is by far the most popular search engine, at least for the less tech-savvy population in developing countries like India. Google [5] is reported to be searching 4k million web pages. Further, it provides vernacular interfaces which makes it popular in a multilingual population like India.

3.2 The audience of the survey

As the evaluation of the usability and effectiveness of addition of features in the web search interface of Google was the subject of the survey, we chose people who use the Google search engine frequently. We took care to include as much variety in the audience. People with ages ranging from 15-50 were chosen for the purpose. With an almost equal balance between men and women, the population was chosen

from different walks of life, ranging from non-tech-savvy people like shopkeepers who are just casual computer users to academicians in the discipline of computer science. Each member of the population was held at an equal weighting.

3.3 The features to be evaluated

The following six features of the Google search engine were chosen for evaluation on the survey. The trade-off was between including features of as much variety, at the same time, keeping the questionnaire to a reasonable size.

- Cached Links: A per-result link which leads to the cached copy of the result
- File Types: A pre-search filter which enables searching for only specific types of files. This feature, which can be deployed using the filetype: operator does not manifest itself as a link
- "I am feeling lucky": A prominent button in the main search interface, which automatically leads the user to the most prominent result of the search query, without having to go through the search engine results pages
- Similar Pages: A per-result link which refines the search to the pages similar to the specific result
- Site Search: A pre-search filter like the implemented by the operator site: which directs the search to pages on a specific site
- Boolean operators: Boolean operators can be used between search query words either by explicit mention in the search query or by using the fields in the advanced search option.

The options given against each feature measures how effectively the user has deployed the feature. It also tries to measure why it has not been deployed, if at all the user has not used the feature. A remarks column per feature allows the user to enter data likely to aid the survey.

3.4 The Questionnaire

The table representing the questionnaire is given as below

Feature	Alternatives	Tick	Remarks
CACHED LINKS	Frequent User Infrequent User Never use it, infact never felt the need Used to use it Who needs to see an old version What is caching ✍		
FILE TYPES	Frequent User Infrequent User Used to use it I need only web pages Never knew about the feature Never knew about the feature, used to try alternative methods Cant understand the meaning		

I AM FEELING LUCKY	Frequent User Infrequent User Used to use it Never knew about the feature Cant understand the meaning		
SIMILAR PAGES	Frequent User Infrequent User Used to use it Never knew about the feature Cant understand its meaning		
SITE SEARCH	Frequent User Infrequent User Used to use it Never knew about the feature Never knew about the feature, used to try alternative methods Cant understand the meaning		
BOOLEAN OPS	Frequent User Infrequent User Used to use it Never knew about the feature Cant understand the meaning		

3.5 Results of the survey

The table representing the results of the survey with each the percentage of people who chose each option recorded against each option is given as below.

Feature	**Alternatives**	**Responses %**
CACHED LINKS	Frequent User Infrequent User Never use it, infact never felt the need Used to use it Who needs to see an old version What is caching L	10 13 28 00 20 29
FILE TYPES	Frequent User Infrequent User Used to use it I need only web pages Never knew about the feature Never knew about the feature, used to try alternative methods Cant understand the meaning	26 04 00 27 11 27 05
I AM FEELING LUCKY	Frequent User Infrequent User Used to use it Never knew about the feature Cant understand the meaning	07 09 00 05 79

SIMILAR PAGES	Frequent User Infrequent User Used to use it Never knew about the feature Cant understand its meaning	40 25 10 15 10
SITE SEARCH	Frequent User Infrequent User Used to use it Never knew about the feature Never knew about the feature, used to try alternative methods Cant understand the meaning	18 10 02 40 21 09
BOOLEAN OPS	Frequent User Infrequent User Used to use it Never knew about the feature Cant understand the meaning	74 12 00 05 09

3.6 Brief Per-Feature observations

3.6.1 Cached links:

The feature manifests itself as a link with anchor text “cached” with each search result. 30% of the respondents did not understand the feature. This might just be a special case, as understandability of technical jargons cannot be expected in an Indian population. It can be seen that only 10% of the population actually use the feature frequently. This is a feature whose presence is enforced by the imperfections of the current search methodology, whereby the search engine cannot guarantee that every link provided by the search engine would be alive. Thus this feature is not effective and usable, given that many people do not understand it, and further, people who understand what the link stands for, do not seem to comprehend the necessity of such a link perhaps due to unawareness of the current web search methodology.

3.6.2 File Types :

As discussed in section 3.2, absence of an explicit link to represent the feature does hinder usability. Users are unaware of the feature unless they see the explicit representation. Many opined in the remarks column that they in fact had to narrow down the search to files of a specific type more than once, but that they resorted to appending the type of the file in the search query (e.g., “anatomy pdf” as a query). This shows that absence of an explicit filter in the main interface does hinder usability a lot. This feature, however has an explicit manifestation in the advanced search interface.

3.6.3 “I am feeling lucky” :

Although this is an explicit button in the main search interface, about 80% of the users could not understand the meaning. Many opined that the name of the feature was too cryptic. Many frequent users thought that the name ought to be changed to something like “best result”. But users who once read about the feature have resorted to using it frequently.

3.6.4 Similar Pages :

A large volume of the users in fact do use the similar pages feature. This is a link that appears with each result with the same font and prominence as “cached”. A simple expressive phrase like “similar pages” does the trick here.

3.6.5 Site Search :

This is implemented by means of an operator. Akin to the filetype feature, people who came to know of the feature through the survey immediately acknowledged the usefulness of the feature. Further, people confessed having used a lot of other alternatives (just as in Section 4.6.2) to achieve the same result. This feature has an explicit manifestation in the advanced search interface.

3.6.6 Boolean Operators :

This feature, is a much celebrated feature in search engines. It is extensively used, that it has spread without the need for an explicit manifestation in the main page (it does have a representation in the advanced search interface). Boolean operators are widely regarded as a “cannot-do-without” feature.

As a general observation, very few people ticked the “used to use it” option. It can be concluded that the features were in fact very relevant and useful that a person who once used it did use it whenever demand arose. It is a further indication that the learn-ability of the features is good. Further, many opined that a link to the features page [15] from the main search page would be of extreme help in understanding the features. Although Boolean operators, file types and site search have explicit manifestations in the advanced search interface, the others are not as popular as the boolean operators. Boolean operators occur as the first entries in the advanced search interface. This points to the need to ensure that a lot of features should not clutter up the advanced search interface as well. Each one of the search interface is to be designed so as to include only the absolutely necessary features. Including more would hinder the usability of all the features. Further, we can also conclude that understandable descriptions are more inviting than eye-catching ones like “I am feeling lucky”.

4. Conclusion

The generalizable conclusions, mostly based on the survey and partly based on the evaluation of the interfaces, show that web search users prefer the easiest interface to the most powerful interface. Learnability and memorability form two very important criteria for web search interfaces, primarily due to the fact that most of the web search users are casual users who come to the search engine interface with widely varying intentions [2]. Further, from the remarks of the people who attended the survey, it can be generally opined that most users who come to the search engine, prefer to actually type in the search query as soon as possible (and refine the results later on, if needed) to sitting down and using the pre-query formulation features to generate a query more probable of getting them accurate results. Further generalizations are presented as below.

Initial query formulation features aid more in cluttering up the search interface. Reducing the main search interface to include only very useful and essential features (or filters) would be a good choice. Advanced search interface is typically not used extensively. Giving a useful feature an explicit and expressive manifestation in the advanced search interface in addition to an operator based access, would perhaps be the best choice. People seldom like going back to the advanced search interface each time, and prefer to use the operator once they have understood the feature. Current representations of useful

features are in-fact very much learnable. This is testified by the observation that people who have used the feature continue to use it. Understandable short descriptions are preferable to eye-catchy ones to represent features. The survey remarks showed that people generally preferred the features that aid refinement of search and reformulation of the query. If the validity of such a hypothesis can be proved, more and more research has to go into finding how initial query formulation features like time and location filter can be adapted to useful features to aid subsequent phases based on information gathered from the search process.

5. References

1. Jakob Nielsen, "The difference between Web design and GUI design", May 1997, http://www.useit.com/alertbox/9705a.html
2. Rose & Levinson, "Understanding user goals in web search", Proceedings of the 13th World Wide Web Conference, NY, pp.13-20
3. "Search Engine Features Chart", April 2004, http://www.searchengineshowdown.com/features/
4. Krishna Bharat & Bay-Wei Chang, "Web Search Engines: Algorithms and User Interfaces", CHI 2003 Tutorial, http://www.chi2003.org/docs/t20.pdf
5. Google Search Engine, http://www.google.com/
6. Courtis et. al., "Cool Tools for searching the web: A Performance Evaluation", Online, 19(6), November 1995, pp. 14-32
7. Scoville, Richard, "Find it on the Net!", PC World, January 1996
8. Chu and Rosenthal, "Search engines for the world wide web: A comparitive study and evaluation methodology", ASIS 1996 Annual Conference Proceedings, October 1996
9. Nielsen, "Usability Engineering", Cambridge MA, 1993, Academic Press
10. Peterson, Michael, "Designing World Wide Web search engine interfaces that Aid users in implementing boolean modified search queries", Term Paper for a course on Human-Computer Interaction, Indiana University at Bloomington, http://www.monroe.lib.in.us/~bpmchi/scholarship/peterson/peterson.html
11. Frank, "Site search can be flattened by usability", http://experiencedynamics.blogs.com/site_search_usability/2004/01/the_power_of_si.html
12. "Search interfaces", Article on bobulate.com, http://www.bobulate.com/popups_full/search_p1.html
13. Notess, "Search Engine Features Chart", April 2004, http://www.searchengineshowdown.com/features/
14. Quesenbery, Whitney, "Designing usable search", Article on WQusability.com, 2002
15. Google Web Search Features, http://www.google.com/help/features.html
16. Google India, http://www.google.co.in
17. Vivisimo Search Engine, http://www.vivisimo.com/
18. Yahoo! Search Engine, http://www.yahoo.com/
19. Deepak P, Sandeep Parameswaran, "Context Disambiguation in Web Search Results", Journal of the Indian Institute of Science, Sept-Dec 2003, pp. 93-102
20. Teoma Search Engine, http://www.teoma.com

About Authors

Deepak P is working in Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai.
E-mail : deepakswallet@yahoo.com

Sandeep Parameswaran is working in IBM Global Services (India) Pvt. Ltd., Embassy Golf Links Campus, Bangalore.
E-mail : sandeep_potty@yahoo.com

Mutual Authentication Protocol For Safer Communication

S P Shantharajah K Duraiswamy

Abstract

With a network communication requires cryptography authentication for secure transmission. This paper takes two challenge-response protocols by which entities may authenticate their identities to one another in a mutual communication way. Here the two entities, each holding a share of a decryption exponent, collaborate to compute a signature under the corresponding public key. These may be used during session initiation, and at any other time that entity authentication is necessary. The authentication of an entity depends on the verification of the claimant's binding with its key pair and the verification of the claimant's digital signature on the random number challenge. The network defined mutually authenticated protocols are for entity authentication based on public key cryptography, which uses digital signatures and random number challenges. Authentication based on public key cryptography has an advantage over many other authentication schemes because no secret information has to be shared by the entities involved in the exchange. This paper analyzes the protocols, which minimizes cost, very simple and suggests security for mutually signing the signature schemes and provides proofs of security for the safer communication. This paper specifies the way of mutual communication and the method of processing that conversation between entities.

Keywords : Authentication, Computer Security, Cryptographic Modules, Cryptography, Digital Signatures, Proofs of Security.

0. Introduction

Authentication based on public key cryptography has an advantage over many other authentication schemes because no secret information can be shared by the entities involved in the exchange. A user (claimant or Initiator or A) attempting to authenticate oneself must use a private key to digitally sign a random number challenge issued by the verifying entity. This random number is a time variant parameter, which is unique to the authentication exchange. If the verifier can successfully verify the signed response using the claimant's public key, then the claimant has been successfully authenticated.

Here, consider a signature scheme of the "hash-then-decrypt" variety, meaning the public key is N, e, the secret key is d, and the signature of message M is $H(M)^d \bmod N$, where H is a public hash function, N is an RSA modulus, e is an encryption exponent, and d is the corresponding decryption exponent. However, instead of there being a single signer, the public key is associated to a pair of entities (even that termed to be a initiator and a responder or a client and a server). The decryption exponent *d* is not held by any individual party, but rather is split into shares d_c and d_s, and these are held by the two entities i.e. the initiator and the responder respectively. A collaborative computation, or signing protocol, is used to produce a signature for the receiver which leads to authentication.

1. Computation of Collaborative Signature

RSA, due to its algebraic properties, lends itself naturally to collaborative signature computation. The decryption exponent is split multiplicatively, meaning

$d_c d_s \; aH \; d \pmod{ö(N)}$

Collaborative signature computation is then based on the equation

$H(M)^{d} \; aH \; H(M)^{dcds} \bmod N$

This paper considers a natural and simple signature schemes based on direct exploitation of the second Equation (above). We ***divide them into two classes***. In the *common-message* class of schemes, the message *M* to be signed is a common input to an initiator and responder i.e. client and server. Within this class we consider two protocols:

MCS: Client sends $x_c = H(M)^{dc} \bmod N$ to the server;
Server computes signature $x = X_c^{ds} \bmod N$,
Verifies it, and returns it to client.

MSC: Server sends $x_s = H(M)^{ds} \bmod N$ to client;
Client computes signature $x = x_s^{dc} \bmod N$. 1

(In the terms MCS, MSC states à M-,Message from, C-Client and S-server)

The leading "M" in the common-message protocols is the "Message" that both entities know. In the client-message class of schemes, the message M to be signed is input to the client but not to the server. Within this class we again consider two protocols for computing a "partial" signature:

HCS : Client sends y, x_c to the server,
Where $y = H(M)$ and $x_c = y^{dc} \bmod N$;
Server computes signature $x = x_c^{ds} \bmod N$,
Verifies that $x^{e} \; a" \; y \pmod{N}$ and returns x to client.

HSC : Client sends y to the server, where $y = H(M)$;
Server sends $x_s = y^{ds} \bmod N$ to client;
Client computes signature $x = x^{dc}_{s} \bmod N$.

(Similarly, in the terms HCS, HSC states à H-, Hash from, C-Client and S-server)

Here the leading "H" in the names of the client-message protocols stands for the "Hash" that the client flows to the server. The other letters reflect the order in which the Client and Server use their shares of the decryption exponent in the protocol.

In this paper we analyze the security of the above four protocols with regard to meeting well-defined modern cryptographic goals in provable ways. We find that the security goals, and the assumptions on the underlying primitives required to prove security, vary from protocol to protocol in a perhaps surprising way.

2. Security Analyses

The first security goal that comes in consideration is to prevent forgery by a third party. We suggest however that this goal is too weak, and instead ask that forgery be hard even for an adversarial client who

is in possession of the correct share d_c and is allowed to engage in interactions with the d_s-holding server. Security against third parties is implied by security against client adversaries, so consideration of the latter only strengthens the security results. This is appropriate because we view the server is a "co-signer" of the client. A verifier who accepts a client signature does so under the belief that the server "*endorses*" it. (A client who succeeds in creating a signature that the server has not endorsed should be viewed as having been successful in forgery.)

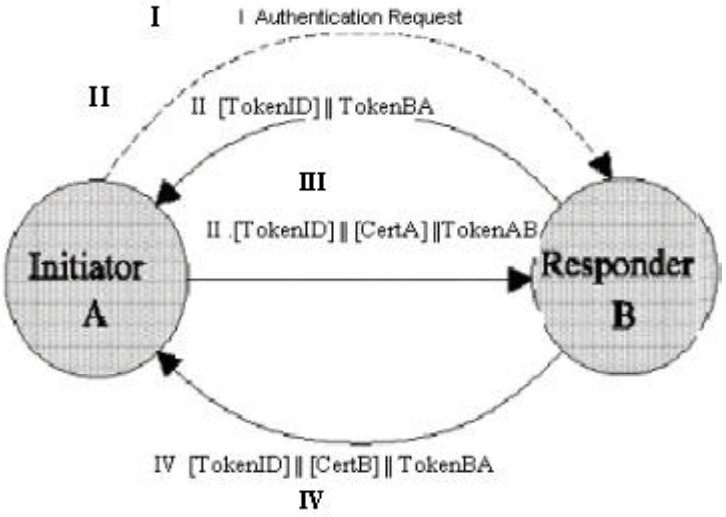


Diagram –1. Authentication process

2.1. Mutual authentication protocol

The mutual entity authentication protocol said to be "Three pass authentication" is given in the above diagram(Diagram –1). Certain authentication token fields and protocol steps are specified in greater detail in this paper. Either entity may choose to terminate the authentication exchange at any time. The mutual authentication protocol refers to two entities (say "initiator" - A and "responder" - B). Here each entity acts as *both* a claimant and a verifier in the protocol (i.e. an initiator and a responder). It is important to note that the success of an entity's authentication, according to this standard, is not dependent on the information contained in the text fields.

The authentication of an entity depends on two things:

1. The verification of the claimant's binding with its key pair,

2. The verification of the claimant's digital signature on the random number challenge.

Authentication occurs as follows:

I. The initiator, A, selects the responder, B, with which it will mutually authenticate, and makes an authentication request to B

II. The responder, B, determines if it will continue, initiate, or terminate the authentication exchange. If it attempts to authenticate the initiator, the responder then

i. Generates a random number challenge, which is the value for the RB field in Token BA.

ii. Generates and/or selects other data which is to be included in the TokenBA.

The responder creates a challenge token of the following form:

Enc [TokenBA = RB || Text]

where 'Enc' **refers** Encryption.

RB – Random Number of B

Entity B sends a message consisting of TokenBA and an optional TokenID to the initiator.

The message from the responder to the initiator is of the form:

[TokenID] || TokenBA

III. Upon receiving the message including TokenBA , the initiator, A, Uses TokenID to determine which token is being received.

Retrieves information from the Text1 field, using it in a manner, which is outside the scope of this standard.

iii. Generates a random number challenge which is the value for the R field in TokenAB

iv. Selects an identifier for the responder, and includes that in the B field of TokenAB.

The initiator creates an authentication token, TokenAB, by concatenating data and generating a digital signature:

TokenAB = Enc [RA || [RB] || [B] || [Text] || S (RA || RB || [B] || [Text])]

The signed data are present only when their corresponding values are present in the unsigned part of TokenAB, although RB does not have to be in the unsigned data of TokenAB.

In addition to containing TokenAB, the message may include a token identifier, TokenID, and the initiator's certificate CertA. i.e. The message from the initiator to the responder is of the form:

[TokenID] || [CertA] || TokenAB

Upon receiving the TokenAB transmission, the responder, B,

i. Uses TokenID to determine which token is being received.

ii. Verifies that the value of RB which is present in the unsigned part of TokenAB.

iii. Verifies the initiator's certificate and verifies the initiator's signature in TokenAB.

Successful completion of this process is that the initiator, A, has authenticated itself to the responder,

B. If any of the verifications fail, then the authentication exchange is terminated.

The responder, B,

(i) Selects an identifier for the initiator, and includes that in the A field of TokenBA

(ii) Generates and selects other data which is to be included in the Text fields. In TokenBA, Text is a subset of the Text field.

The responder creates an authentication token, TokenBA, by concatenating data and generating a digital signature:

TokenBA = Enc[[RA] || [RB] || [A] || [Text5] || S (RA || RB || [A] || [Text4])]

The responder to the initiator is of the form:

[TokenID] || [CertB] || TokenBA

Upon receiving the message including TokenBA, the initiator, A,

i. Uses TokenID to determine which token is being received.

ii. Verifies that the value of R

iii. Verifies that the identifier for the responder has been obtained in CertB, TokenBA.

iv. Verifies the responder's certificate

v. Verifies the responder's signature on TokenBA.

Successful completion is that the responder, B, has authenticated itself to the initiator, A, and thus the entities have successfully mutually authenticated.

4. Conclusion

From the points discussed above, we can conclude with security issues which state that, by assuming the base signature as secure one, we make the mutual authentication more secure one. A responder (B) can identify a initiator (A) in a safer way so that nobody can impersonate any one. The additional security is that no hackers can intrude into their communication area. During communication the verification is done every time by both the entities and leads to secure transmission against hackers attack.

Here both the entities are identified by each other by performing cryptographic computations, validations and verification. Finally, the mutually authentication protocols provides authentication between the entities.

5. References

1. M. Bellare, C. Namprempre, D. Pointcheval and M. Semanko ,(2001), The one-more-RSA-Inversion problems and the security of Chaum's blind signature scheme. Lecture Notes in Computer Science Vol.2330, P. Syverson ed., Springer-Verlag.

2. A. Yao, (1986), How to Generate and Exchange Secrets, Proceedings of the 27th Symposium on Foundations of Computer Science, IEEE.

3. S. Goldwasser, S. Micali and R. Rivest, (April 1988) A digital signature scheme secure against adaptive chosen-message attacks, SIAM Journal of Computing, 17(2); 281-308.

4. D. Chaum, (1983), Blind Signatures for untraccable payments. In Advances in Cryptology – CRYPTO'82, Plenum Press.

5. ITU-T Rec. X.509 | ISO/IEC 9594-8, (1993), Information Technology - Open Systems Interconnection - The Directory: Authentication Framework, Editor's DRAFT.

About Authors

S. P. Shantharajah is a Lecturer in Department of Computer Science in K.S.R. College of Technology, Tiruchengode, Namakkal Dt., Tamilnadu.
E-mail : spshantha_raj@yahoo.com

Dr. K. Duraiswamy is a Principal in K.S.R. College of Technology, Tiruchengode, Namakkal Dt., Tamilnadu.

Effectiveness of Name Searching in Web OPAC: From Authority Control to Access Control

Veerankutty Chelatayakkot V Jalaja

Abstract

Since the advent of Unicode, electronic representation of regional language script is not as issue. However, transliteration is inevitable in a database designed to meet the needs of people who does not know that language. By transliteration, a unique name in one language or culture may have variant spelling in another language. The cataloguer maintainsname authority file to solve this problem. This article analyse the optionsfor retrieving transliterated Arabic names in major OPACs.

Keywords : OPAC, Web OPAC, Information Retrieval.

0. Information Retrieval and Name Searching

Information Retrieval is the process of finding some desired information from a store of information or database. It implies the concept of selectivity. Information recovery is not the same as information retrieval unless there has been a selection process. Name searching is one of the major activities in information retrieval. The name searching may be either from free-text or from a database.

Database name matching technology has long been used in criminal investigations, counter-terrorism efforts, and in a wide variety of government processes such as Visa processing. In this technology, a name is compared to names contained in one or more databases to determine whether there is a match. Sometimes this matching operation may be straightforward exact match, but often the process is more complicated. Two names may not match exactly for a wide variety of reasons and yet still refer to the same individual. The challenges of name matching are greatly increased by: a) structure of names: people uses different structures in rendering their names i.e., the surname come first, before the given name, b) variant spelling of a unique name, c) pseudonyms, d) name with suffix / prefixes and e) transliterated names from native script (e.g. Arabic, Chinese, Cyrillic, Hindi).[1] As against manual searching, a slight difference in input may affect the effectiveness of Information Retrieval in the context of database name searching. In library context, maintaining name authority files solves these problems.

1. Transliterated Names and Authority Control

The name authority file links variant forms of heading to the preferred form of heading to help preparation of cross-references. An author may have different names/pseudonyms or variant spellings may be occurred due to transliteration of name. Transliteration is the process of formulating a representation of word in one language using the alphabet of another language. The aim of transliteration is to represent the script of a source language by using the letters or symbols of another script, usually in accordance with the orthographical conventions of the target language. By the result, a unique name in one language or culture may have variant spelling in another language. Further, owners of these names may take certain liberties with spelling of their names. This is the same condition, where an author uses different names or pseudonyms. The cataloguer should create cross-references from variant spellings (different name) to the heading.

Since the advent of Unicode, electronic representation of non-Roman characters is not an issue. As a universal character code set, Unicode provides a unique number to every character used in modern scripts thought out the world. The Unicode Standard is the universal character-encoding standard used for representation of text for computer processing. Versions of the Unicode Standard are fully compatible and synchronized with the corresponding versions of International Standard ISO/IEC 10646. For example, Unicode 4.0 contains all the same characters and encoding points as ISO/IEC 10646:2003. The Unicode Standard provides additional information about the characters and their use.

The application of Unicode helps to reduce the relevance of transliteration up to an extent. However, transliteration is inevitable in a database, which is designed to meet the needs of people all over the world. In a database having Arabic names entered in Arabic using Unicode can't be of use to such people who have no knowledge of Arabic language. It means that Unicode can solve the problem of electronic representation of regional languages, but such a database can be of use only to a limited people who know that language. So transliteration of personal names from regional languages is essential to meet the need of international users.

2. Information Retrieval and Authority Control

The quality of online databases is very much related to the quality of data in the database. There will be variant spelling, omission, deletion or typing errors for a unique entity. By the acceleration of electronic publishing, there is increase in number of online databases. The challenge is to get correct information from the databases. One aspect of improving data quality is detection of variant name for a unique entity and link them to improve searching. This is similar to authority work. The technique of authority work has vide implication in the effectiveness of Information Retrieval in online databases.[2]

When a user keys in a known form of a heading, the system follows internal linkage and displays the requested item even though the preferred form of heading might be quiet different from the form entered. There should be direct linkage between the form of heading entered and record displayed. The authority control mechanism works as invisible mechanism so far as a user is concerned.

In the case of transliterated names authority file, the cross-references are mostly variant spellings of personal names. While transliterating names in to Roman script, by many reasons, variant spellings denote a unique name. The usage of variant spellings compels the cataloguer to create cross-references to the selected heading.

3. Options for Retrieval of Transliterated Arabic Names in Major OPAC

A review of options for retrieval of transliterated Arabic names in major OPACs will help to understand how the problems in name searching are solved in OPACs. The review is based on Library of Congress Name Authority File (LCNAF), which is the largest authority file in the world. The headings available in both libraries i.e., Library of Congress and reviewed library, were taken. Each heading and its see references in LCNAF were searched in reviewed OPAC and total hits were analyzed. Fifteen OPACs were compared with LCNAF. The major options for name searching in OPAC are:

1. *See References* (4xx in MARC). Some OPAC automatically links to variant forms of heading by using See Reference as provided in the LCNAF. For example,

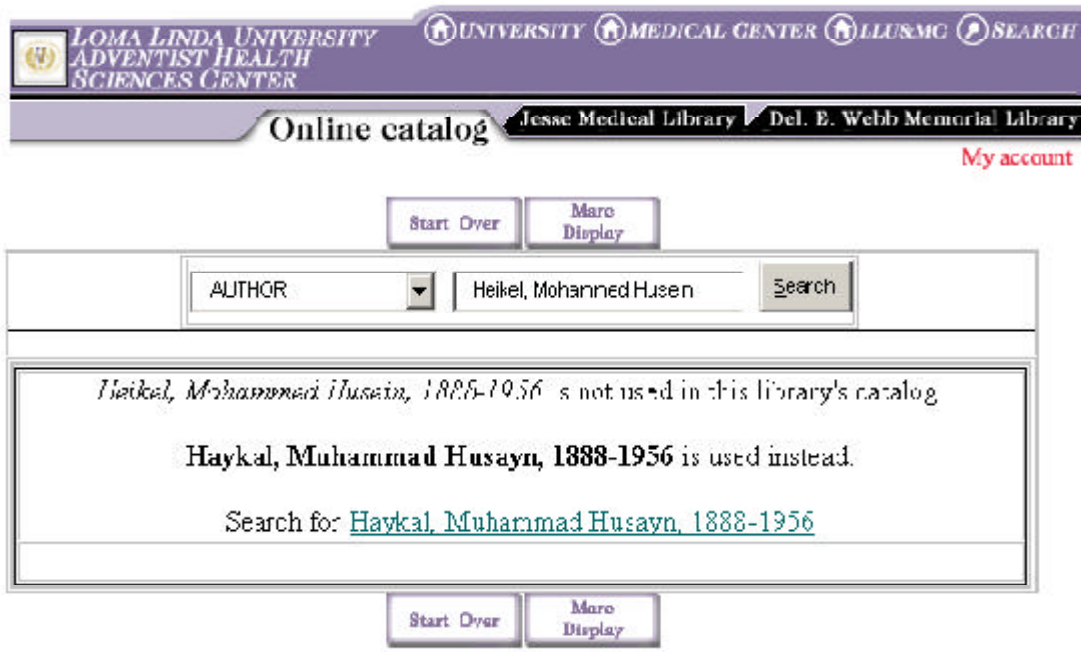


Fig.1 OPAC of Lomia Linda University http://catalog.llu.edu

2. *Inverted Heading*: Inverted heading are used to solve the problems in name searching occurred by the structure of name. A personal name may consist various elements denoting their family name, house name, place name, honorific titles etc. The structure and rendering of names have no relevance in a automated system, where the system automatically inverts the headings. For example,

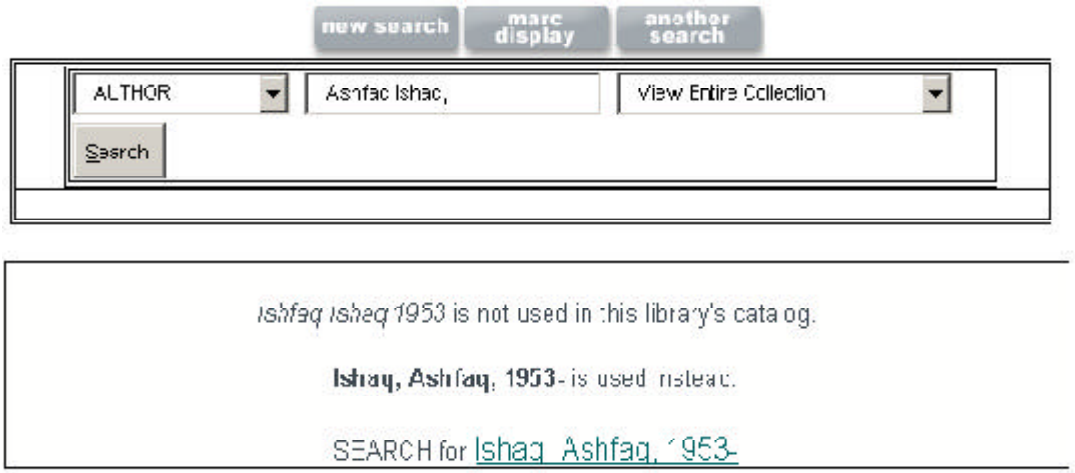


Fig.2 OPAC of University of South Africa http://oasis.unisa.ac.za

3. *Nearby Authors*: Some OPACs provide the option to display nearby authors in order to solve the problem of spelling variations in input. For example,

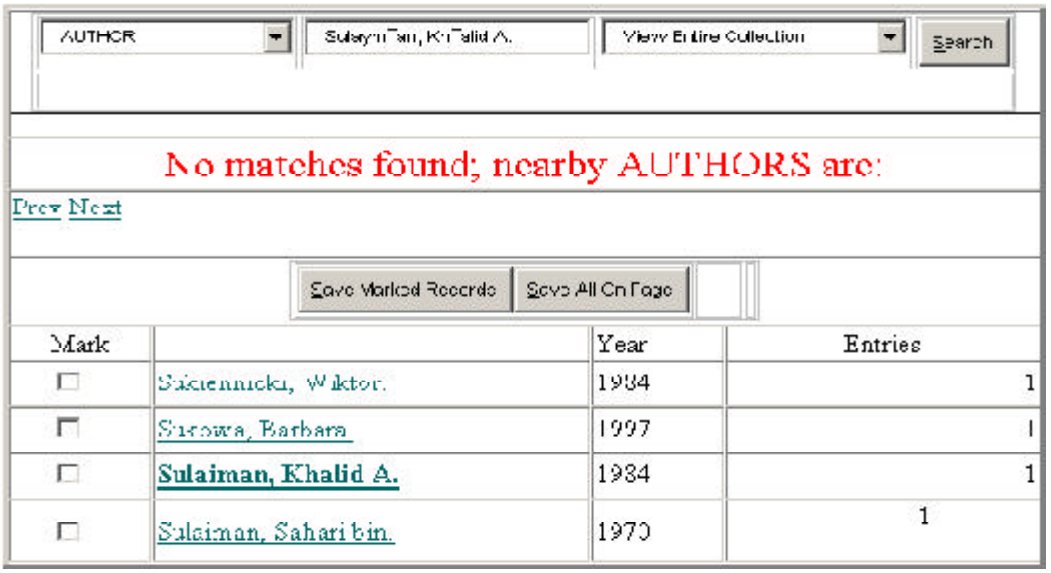


Fig.3 OPAC of Colorado State University http://catalog.library.colostate.edu/

Analysis :

No.	Name of the University/OPAC	Heading as in LCNAF	Inverted Reference	See References	Nearby Author's
1	COPAC	Yes	Yes	No	No
2	Duke	Yes	Yes	No	Yes
3	Edith Cowan	Yes	Yes	No	No
4	Lona Linda	Yes	No	Yes	Yes
5	Ottawa	Yes	No	No	Yes
6	Wales	Yes	No	Yes	No
7	St.Andrew	Yes	No	Yes	Yes
8	South Africa	Yes	Yes	Yes	Yes
9	Colorado	Yes	No	Yes	No
10	Essex	Yes	Yes	No	Yes
11	Ohio	Yes	Yes	Yes	No
12	Oxford	Yes	Yes	Yes	No
13	Trinity	Yes	No	No	No
14	Australia	Yes	Yes	Yes	Yes
15	Nevada	Yes	Yes	Yes	No

The analysis of above 15 OPAC shows following findings:

1. All headings are similar to Library of Congress headings
2. Most OPACs have the option of Inverted Reference (i.e. Inverted Heading)
3. Most OPACs provide See References
4. Only 6 OPACs have option of Nearby Authors.

4. Authority Control Versus Access Control

The access control record is the next generation of the authority record. It may be called as "super authority record" because of the potential it contains for enriched information for indexing.[3] Access control records can be linked both to bibliographic records, to collocate all manifestations of a work, and to other related access control records, to collocate related works. The basic concept behind the access control record is removing both the label and notion of "authority". While authority control record declare a heading as "authorized" form, access control record links all variations without declaring one heading as authorized form. Access control records allow users to choose their preferred form or name, or to have displayed a default heading. It allows for more flexibility in display. The concept of access control entirely contradicts the whole second part of AACR-2, which is devoted to painstaking rules for how to construct authorized forms of names and titles.

The idea behind the access control is that an entity can be known by more than one name. An individual is an entity but may be called by different names by different people at different times in life. In the international realm living persons have name representations in many languages and script. The possible solution is that instead of selecting one form of name as heading, give chance to select the one of the heading as default headings according to their interest (local convenience) and give non-default as reference. We can hope good news from technocrats

5. References

1. Patman, Frankie., and Paul Thompson (2003)Names: A Frontier in Text Mining*in Lecture Notes in Computer Science*, pp.27-38,edited by H Chen, Springer-Verlag Heidelberg .
2. Tillet, Barbara B (2001) " Authority Control on the Web" Conference on Bibliographic Control in the New Millenium USA, November 11-15,2000, organized by Library of Congress URL: http:// www.loc.gov/catdir/bibcontrol/tillett-_paper.html (Accessed on 23/11/2004)
3. Barnhart, Linda (1996)"Access Control Records : Precepts and Challenges" Authority Control in 21st Century :An invitational Conference Ohio, USA, March 31-April1,organized by OCLC Online Computer Library Center Inc. URL http://www.lib.byu.edu/dept/catalog/authority (Accessed on 23/ 11/2004)

6. Appendix

Details of Data Analysis

1. COPAC (http://www.copac.ac.uk/copac)

Entry in LC Authority Record	Nature of Entry in LC	Hits in COPAC	Nature of Entry In COPAC
Haykal, Muhammad Husayn, 1888-1956	Heading	26	Heading
Haikal, Muhammad Husain, 1888-1956	See Ref.	0	No. Ref.
Heikel, Mohammed Husein, 1888-1956	See Ref.	0	No. Ref.
Heikel, Mohamed H 1888-1956	See Ref.	0	No. Ref.
Basheer, Vaikom Muhammad, 1910	Heading	21	Heading
Vaikom Muhammad Basheer, 1910	See Ref.	19	Inverted
Muhammad Bas¯ir, Vaikkam, 1910	See Ref.	0	No. Ref.
Vaikkam Muhammad Bas¯ir, 1910	See Ref.	0	No. Ref.

2. Duke University , USA (http://catalog.library.duke.edu/)

Entry in LC Authority Record	Nature of Entry in LC	Hits in DUKE Uni.	Nature of Entry in DUKE Uni.
R¯ashid, Rushd¯I	Heading	2	Heading
Rushd¯i R¯ashid	See. Ref	0	No Reference
Rashed, Roshdi	See. Ref	0	No Reference
Husain, ‘¯Amir Liy¯aqat	Heading	1	Heading
Amir Liy¯aqat Husain	See. Ref	0	Inverted
Hussain, Aamer Liaquat	See. Ref	0	No Reference

3. Edith Cowan University , Australia (http://library.ecu.edu.au/)

Entry in LC Authority Record	Nature of Entry in LC	Hits in Edith Cowan Uni.	Nature of Entry in Edith Cowan Uni.
Mohamed Yusoff Ismail	Heading	1	Headings
Ismail, Mohamed Yusoff	See. Ref	0	No Reference
Ahmad Mansoor	Heading	1	Heading
Mansoor Ahmad	See. Ref	11	Inverted

4. Loma Linda University , USA (http://catalog.llu.edu/)

Entry in LC Authority Record	Nature of Entry in LC	Hits in Loma Linda Uni. OPAC	Nature of Entry in Loma Linda Uni. OPAC
Ahmed, M. Samir	Heading	1	Near by Authors
Ahmed, Samir	See. Ref	1	See Ref.
Ahmed, Mahmoud Samir	See .Ref	1	Heading
Haykal, Muhammad Husayn, 1888-1956	Heading	1	Heading
Haikal, Muhammad Husain, **|d** 1888-1956	See .Ref	1	See. Ref
Heikel, Mohammed Husein, **|d** 1888-1956	See .Ref	1	See. Ref

5. University of Ottawa . (http://www.biblio.uottawa.ca/orbis-e.php)

Entry in LC Authority Record	Nature of Entry in LC	Hits in Uni. Ottawa	Nature of Entry in Ottawa Uni. OPAC
Hassan Ibrahim	Heading	1	Nearby authors
Ibrahim Hassan	See. Ref	1	Heading
Ibrahim, Datuk Haji Hassan	See .Ref	111	Related Heading
Yousef, Yousef A.	Heading	1	Heading
Y¯usuf, Y¯usuf A.	See. Ref	0	No reference
Yousef A. Yousef	See .Ref	1	Nearby Aauthors

6. University of Wales., UK (http://library.bangor.ac.uk/)

Entry in LC Authority Record	Nature of Entry in LC	Hits in Uni. of Wales	Nature of Entry in Uni. of Wales OPAC
Jinnah, Mahomed Ali, 1876-1948	Heading	1	Heading
Muhammad ‘Al¯i Jinnah, 1876-1948	See. Ref	1	See. Ref
Quaid-i-Azam	See .Ref	1	See. Ref
Faruqee, Rashid, 1938-	Heading	1	Heading
Rashid Faruqee	See .Ref	1	See. Ref
Faruqee, R	See. Ref	1	See. Ref

7. University of St.Andrews (http://138.251.116.3/)

Entry in LC Authority Record	Nature of Entry in LC	Hits in SAULCAT	Nature of Entry in SAULCAT
Ali, Mohamed	Heading	1	See. Ref
Ali, Muhammad	See. Ref	1	Heading
Mohammad Ali, **|c** Maulana,	See. Ref	0	No. Ref.
Sulaiman, Khalid A.	Heading	1	Heading
Sulaym¯an, Kh¯alid A.	See. Ref	1	Nearby auth.

8. University of South Africa (http://oasis.unisa.ac.za/)

Entry in LC Authority Record	Nature of Entry in LC	Hits in Nat. Lib. of S. Africa	Nature of Entry in Nat. Lib. of S.Africa
Ishaq Ashfaq	Heading	1	Heading
Ashfaq Ishaq,	See. Ref	1	Inverted heading
Rasheed, Sadig	Heading	1	Heading
Rash¯id, Sad¯iq	See. Ref	1	See. Ref
Rasheed, Sadiq	See. Ref	1	See. Ref

9. Colorado University USA (http://catalog.library.colostate.edu/)

Entry in LC Authority Record	Nature of Entry in LC	Hits in Colorado State Uni. Lib.	Nature of Entry in Colorado State Uni.
Salam, Abdus, 1926	Heading	1	Heading
Muhammad ‘Abd al-Sal¯am, **|d** 1926-	See. Ref	1	See. Ref
Salam, Muhammad Abdus, **|d** 1926-	See. Ref	1	See. Ref
Sulaiman, Khalid A.	Heading	1	Heading
Khalid A. Sulaiman	See. Ref	1	See. Ref
Kh¯alid A. Sulaym¯an	See. Ref	1	See. Ref

10. University of Essex, UK (http://serlib0.essex.ac.uk/)

Entry in LC Authority Record	Nature of Entry in LC	Hits in Essex Nature	Nature of Entry in Essex OPAC
Basheer, Tahseen	Heading	1	Heading
Bashir, Tahseen	See. Ref	0	Not found
Tahseen Basheer	See. Ref	1	Inverted reference
Ahmad Ibrahim	Heading	1	Nearby Authors
Ahmad Mohamed Ibrahim	See. Ref	0	Not found
Ahmad bin Mohamed Ibrahim	See. Ref	101	Heading

11. Ohio University, Athens (http://www.library.ohiou.edu/)

Entry in LC Authority Record	Nature of Entry in LC	Hits in Essex Ohio	Nature of Entry in Ohio
Bin Laden, Osama	Heading	1	Heading
Usama bin Laden	See. Ref	1	See. Ref.
Ibn L¯adin, Us¯amah, **|**	See. Ref	1	Inverted reference
Zakaria bin Haji Ahmad	Heading	1	Heading
Ahmad, Zakaria bin Haji	See. Ref	1	See. Ref
Zakaria Haji Ahmad	See. Ref	1	See. Ref

12. Oxford Library and Information System. (http://www.lib.ox.ac.uk/olis/)

Entry in LC Authority Record	Nature of Entry in LC	Hits in OLIS	Nature of Entry in OLIS
B¯ar¯ud¯i, ‘Abd All¯ah ‘Umar	Heading	1	Heading
Abd All¯ah ‘Umar al-B¯ar¯ud¯I	See. Ref	1	See. Ref
Husayn¯i, ‘Abd All¯ah ibn ‘Umar al-B¯ar¯ud¯i	See. Ref	0	Not Found
Bashier, Zakaria	Heading	1	Heading
Zakaria Bashier	See. Ref	1	Inverted Ref.
Bash¯ir, Zakar¯iy¯a	See. Ref	0	Not found

13. Australian National University, Canberra (http://library.anu.edu.au/search~S1/)

Entry in LC Authority Record	Nature of Entry in LC	Hits in ANU	Nature of Entry in ANU
Abdesselem, Mohamed	Heading	1	Heading
Mohamed Abdesselem	See. Ref	1	See. Ref
Muhammad ‘Abd al-Sal¯am	See. Ref	1	See. Ref
Ish¯aq, Muhammad Qamar, 1961-	Heading	1	Heading
Ishaque, Mohammad Qamar,1961-	See. Ref	1	See. Ref
Mohammad Qamar Ishaque, 1961	See. Ref	1	See. Ref

14. Trinity Theological College, Australia (http://www.trinity.qld.edu.au/)

Entry in LC Authority Record	Nature of Entry in LC	Hits in Triniti OPAC	Nature of Entry in Triniti Theo. College
Muhammad Abul Quasem	Heading	1	Heading
M. A. Quasem	See. Ref	0	Not Found
Abul Quasem	See. Ref	1	Heading
Ahmad, Bashiruddin Mahmud	Heading	1	Heading
Bashiruddin Mahmud Ahmad	See. Ref	1	Heading
Mahmood Ahmad, Bashir-ud-Din	See. Ref	0	Not found

15. University of Nevada, Rino(http://www.library.unr.edu/)

Entry in LC Authority Record	Nature of Entry in LC	Hits in i UNLOPAC	Nature of Entry in Uni. of Nevada Lib.
Badaw¯i, Muhammad Mustafá	Heading	1	Heading
Badaw¯i, Mustafá	See. Ref	1	See. Ref
Badawi, M. M.	See. Ref	1	See. Ref
Hussain, Asaf	Heading	1	Heading
Asaf Hussain	See. Ref	1	Inverted Heading
Hüseyin, Asaf	See. Ref	0	Not found

About Authors



Veerankutty Chelatayakkot, Junior Librarian, Cochin University of Science and Technology, Kochi-22
E-mail : veerankutty@cusat.ac.in



Dr. V. Jalaja, Presently she is working as Reader, Dept. of Library and Information Science, University of Calicut. She has 15 years experience in teaching. A number of Ph.D has produced under her guideship. She has a number of publications in her credit. Her area of interest is Bibliomining.

Digital Preservation of Art, Architectural and Sculptural Heritage of Malwa (Madhya Pradesh)

S Kumar Mukesh Kumar Shah Leena Shah

Abstract

Digitization is also an outcome of development of technologies. The distinct form of art and culture requires preservation for the study of mankind and its progress. Paper stresses need to capture them in digitized form for future interest of the researchers. It will also give an account of art and architecture and sculpture of Malwa region of Madhya Pradesh (India). It had one of the ancient civilizations in India and World. History of Malwa region of Madhya Pradesh (India) dates to prehistoric period. The archeological excavation and heritage structures of Malwa are of no less importance than any other part of India. Much is being done on digitization of manuscripts but very little is thought of these valuable heritage structures lying unprotected and open. Digitization of cultural heritage will be best preservation for future generations' .The paper suggest for 3 dimensional picturization of these heritages and preserve in digital form. It also suggests for hardware, software and human ware requirements for such work and gives a plan and provides few beautiful slides. A practical LCD projection of digitized heritage site is also prepared.

Keywords : Digital Preservation, Manuscripts, Digitization.

0. Introduction

Information Technology has changed the shape of the face of the Libraries and Library and Information Science. Earlier the activities were limited to management of books but ICT has brought many other activities in its domain. Digitization is one such activity, which has changed the entire gamut of LIS. There are distinct forms of art and culture, which require preservation for the study of mankind and its progress but not in original conditions. Our traditional heritage is quite old and will perish some day if not preserved in time. Thousands of architectural and sculptural beauties have perished. Recently Archeological Survey of India (ASI) has accepted that minarets of 356-year-old Taj monuments are tilting. The international beauty symbol Taj may perish someday. (1) Aggressions, social clashes, natural calamities, and the earthquakes have already demolished hundreds of thousand heritage buildings and architecture and sculptures. Theft and smuggling are two important factors for big loss of such treasure. Foreign aggressions have already caused unreplacable losses. Our heritage has been taken to foreign countries by aggressors and rulers and has either been destroyed or has been kept in their museums. They cannot be replicated by pooling national resources. At present maintenance of monuments is also not so easy.The only way with the available technology is to capture them in digitized form for future interest of the researchers in the form of 3 dimensional models. Now 3D scanners have made it possible to digitally capture these in all their glory. With the help of suitable softwares for storage and maintenance these can be made available on the Internet. We can create virtual museums and virtual palaces, etc. (2)

There were 3200 Hutongs about 800 years old, out of which only 930 are remaining to be digitized. Hutongs are ancient city alley or lone typical in Beizing (China). Museums and online Archive of California (MOAC) is a project for virtual museum in California. (3). In Malwa there are few museums at Ujjain, Indore, Bhopal, Dhar, Mandu, Mandsour, Neemuch, Shajapur, Vidisha, etc containing cultural heritages. The articles found in excavation and pieces of structures demolished by aggressors or due to natural calamities are stored in these museums. Vikram University has one of the richest museum of Malwa. It

has different types of Brahmanical Gods Icons several types of ancient rare coins and some Inscriptions and excavated woodens. Now we need to digitized our cultural heritage and make virtual museum. Cherish before these perish.

The Malwa

The Madhya Pradesh is known as heart of India and its Malwa region keeps an own geographical and historical importance in Indian History. This region was famous to the name of Avanti in the 6th century B.C. From the occurrence of 5th century B.C. this region was comprehensively called as Malwa. This word is related to Malwajan of Punjab in contemporary period. After the invasion of Alexander (King of Babilone) the Malwajan made their residence in Rajasthan and after some time their some branches came in Avanti region and they were reign here hence this region got other name- Malwa.

1. Early Architecture

Malwa region being kept the peculiar heritage from the Palaeolithic period in its lap. Bhimbethaka is one of them discovered by V.S.Vakankar in 1958 at Raisen district. At Bhimbethaka caves we can see middle Palaeolithic (1,00,00-40,000 B.C.) paintings. Paintings are related to hunting of Animals. We can see the tools used in hunting and types of hunting and the types of animal in contemporary age in these lexicon middle and upper Palaeolithic paintings.

The precious heritage art and architectural, Sculptural work, being started virtually from 6 century B.C. In this period religious movement gave an impetus to the development of art and architecture. But from the Shunga-Satvahan-Shaka period (186 B.C.-318B.C.) large stupas and temples began to be constructed. At Sanchi (Dist-Raisen) there is the first instance of true masonry used for constructional purpose in any ancient building. At Sanchi there is the original brick stupa built by Ashoka. It was later encased during the Shunga period. The new structure is about twice the original Mauryan Stupa covering an area 120 feet in diameter with a total height of 54 feet. (4)

There are three Stupas installed at Sanchi, with semi-circular domes, lofty terraces (Medhi). Pradakshinapaths, flattened summits, small square pavilions (Harmika) and railings. On the ground level each stupa has a second processional path, paved with stone, encircling the stupa. In the 1st century B.C., the four elaborate and richly carved gateways were built in the four directions of the Stupa. They are similar in design and 34 feet in height. The bas relieves of four gateways and railings relate to Buddha's (God Buddha) life, showing his birth (Jati), Enlightenment (Sambodhi), first sermon (Dharmachakrapravartan) and death (Mahaparinirvana). Each is represented by its own peculiar symbol: the lotus, the pipal tree, the wheel and the stupa (5,6,7). (Photo 1, 2)

At Udayagiri (district – Vidisha) and Besanagar (District- Vidisha) and in the neighborhood of Sanchi, at the site such as Sonari, Shatadhara, Bhojapur and Andher, some Buddhist stupas and monasteries of the Shunga period have been found. Like the Sanchi Stupas, they are also made of stones. (8) A few freestanding statues of yaksha and yakshi (Local God) of the Shunga period found at Besanagar (District-Vidisha). The Statue of yaksha measures over 12 feet in height, and is decidedly the biggest yaksha image of an early period so far discovered in India. It is the peculiarly Indian in their dress and ornamentation and also in spirit and outlook. It reflects primitiveness in art and indicates the earliest phase of the indigenous art of India and Malawa. (8,9,10) (Photo 3).

2. Gupta Period Architecture

After declining of Shunga-Satavahani Shaka period the Gupta-Aulikara Empire/ (319 to 700 A.D.) came in light in the Malwa. The original capital of Gupta emperor was Pataliputra (Bihar) but after the victory on Western Kshatrapas of Ujjain. Gupta emperor Chandra-Gupta –II (375-414 A.D.) installed the sub-capital at Ujjayini (Ujjain). He is also known as Vikramaditya of Ujjain. During this period, the Aulikaras emperors also occupy an important place in the history of Malwa.The Gupta period was the golden age in the history of India and also Malwa. We found historical buildings in the form of caves and structural buildings of this period.

The earlier examples of structural temples can be seen at Eran (Dist. Sagar). The temples of Vishnu and Varaha, though in dilapidated condition, are rectanguter in plan. The walls of the temple are quite plain, but there is decorative richness on the pillars and doorframes. (11)(Photo 4).

The cave architecture has been found at Bagh (145km to Indore dist.) and Udayagiri, (Bhilsa, dist. Vidisha). The Bagh caves are sacred to Buddhism and Udayagiri are related to Brahmanical religion and Jainism.

2.1 Bagh Caves

At the Bagh, rather there were many caves but owing to the weakness of the rock they has been seriously damaged. Some caves are preserved so far. Cave no. 3 is a pure vihar, cave Nos. 2 and 4 are combinations of the Chaitya (The monuments of prominent person's bodies remains) and vihara. The most important cave is the Great Vihara (No.IV), known locally as the Rangmahal. It consists of a central hall of about 96 feet. In this hall we can see the highly reached development of paintings during Gupta period. In the Rangmahal painting, there are a group of women, playing dance with well-decorated apparel and ornaments. Other one painting is related to procession, which is going on elephant (12,13) (Photo 5).

These paintings are related to indigenous culture and the subjects covered by the paintings are varied and numerous, such as the representation of the Buddha and Bodhisattvas, decorative scroll works, friezes and other patterns. The jatak stories have been beautifully illustrated in the Bagh Caves (14).

2.2 Udayagiri Caves

The Udayagiri Caves situated in Bhilsa (Dist. Vidisha) are twenty in number. They are partly rock cut and partly stone built. Some of the caves contain inscriptions also. Out of these caves all are Brahmanical except one or two Jaina caves. Caves no. 1,2,4,7,16,17 and 19 show distinct features of architectural value. Cave no.4 has shrine, which is much large and more ornate. The cella of cave no.19 is more spacious. We can see the evolution of temple architecure and decorative motifs in various caves. (15)

Cave no.6 of Udayagiri was laid by Maharaja Sankanik (Feudatory of Chandra – Gupta –II) in the year of 401 A.D. It's entrance gate is highly ornate. Both sides of the gate, the piller are carved in the form of trees and at head of the piller lions are sitted in bell shaped (Photo.5). This cave has also the figure of Mahishasurmardini (i.e. the Goddess Durga killing the Buffalo demon or Mahisasura) having 12 arms holding different objects. With her foot, she is shown trading upon the head of the Buffalo. The Umamaheshvara in amorous mood found at this place, is also noteworthy. The peculiarity of Gupta art is, that the two rivers, Ganga and Yamuna make their appearance for the first time in scheme of the temple architecture. On the doorways of cave no. 6 these deities are found as attendants of the great God. (16) (Photo 7,8)

Cave no.5 of Udayagiri keeps important place in the History of Indian caves art, because there is a large vision of Varah. (Incarnation God Vishnu) inscribed on the rock here. In this Icon we can find all alluring vision from divine world to material world and they all became a part of architectural world. This caves has given the wide prestige to Malwa in the Indian art. (Photo.9).

Cave No. 13 shows the fantastic colossal statue of Sheshshayi Vishnu. God Vishnu (Photo.9) is sleeping on the coils of the primeval snake with head resting on the palm of one of his four hands. He is attended by his vehicle Garuda (the Eagle) (17) (Photo 10).

Cave No. 20 is a Jaina Cave, having Jain Tirthankar Parshvanath icon laid foundation at the time of Gupta ruler kumar Gupta I. 5thCentury AD. (Photo 11)

3. Post Gupta Period Architecture

After the decline of Gupta Empire the Pratiharas ruled in Malwa and after their decline the Paramaras (Ujjain and Dhar) became powerful.

3.1 Shitaleshvara-Mahadev Temple

The earliest dated temple of post Gupta period is the Shitaleshvara-Mahadev at Chandravati, near Jhalarapatan (Now in Rajasthan State) founded in 689 A.D. It has been demolished and crudely rebuilt but it still retains some original parts. The pillars of the temple are minutely carved. These are unique example of such intricate stone work (18) (Photo 12).

3.2 Dhamanar Caves

At Mandsour there are 70 caves known as Dhamanar caves of 8-9th Century A.D. It may be possible that the ancient name of Dhamanar would be Dharmanath but in present time this place is known as Dharmarajeshawar. These caves are heritage of Buddhist religion of Malwa in the rocks. Caves no. 6,11,12,13 are known respectively in the name of Badi Kachahari, Bhimabazar, Hathibaghi and Chhotabazar. (19) (Photo13).

There is also a temple of Dharmanath at Dhamanar), originally dedicated to Vishnu (God Vishnu). It belongs to the 8th Century A.D. This monolithic temple is of the same general style as that of the famous Kailash temple at Ellora (Aurangabad) (Maharashtra.) (20).

3.3 Maldevi-temple

At Gyaraspur (Dist. Vidisha) the Maldevi temple, which is partly rock-cut and partly structural, is a mature instance of pratihara style of 9th Century A.D. The roofs of the porch and roofs of the hall is pyramidical composed of horizontal tiers. The hall doorways shows a figure of Chakreshvari as the tutelary image. (21) (Photo.14)

3.4 Bajra- Matha Temple

It is an example of rare class of temple, belong to 8-9th Century A.D. at Gyaraspur Dist. Vidisha) containing three shrine in an arrow. Three shrines dedicated to the gods of the Hindu trinity Brahma, Vishnu and Shiva. The carving of the doorway is exceptionally fine and vigorous. (22) (Photo.15).

4. Parmaras Period Architecture and Sculpture

4.1 Architecture of Parma's

This section studies architecture and sculpture of 9-13th Century A.D. After the decline of Gurjar Pratihara Empire the Paramaras of Ujjain and Dhar became powerfull. The rulling dynasty were inspired by the earlier rich traditions of art, architecture and sculpture and vied with one another in building temples. The temple built during this period in Malwa are known as the Bhumija style of architecture. This style was not confined to Malwa but spread to Rajasthan, Gujarat, Maharashtra and Deccan (23).

4.1.1 Bhojashala and Lata Massed

The Paramara ruler Bhoja was a great patron of art. He has established a college known as Bhojashala, It consist large open court and prayer hall. The pillars and ceilings of the prayer hall are carved delicately. There are numerous slabs of black slate stone carved with the writings of the Parijatmanjari (of Arjun Verma) and Kurmashatak (of Bhoja). Similarly the Lata Maszid (Mosque) of this place is planned with carved pillars and brackets of older temples. (24)(Photo 16, 16a)

4.1.2 Shiva Temple

A magnificent temple of Shiva was built in the 10th Century A.D. during the reign of Bhoja at Bhojapura (20 miles South of Bhopal). This temple is situated on a low rocky hill to the northeast of the great Bhojapura Lake. It is square in plan. Four massive and monolithic columns surmounted by flowered capitals, support a circular lower most of which is decorated with figures of musicians and demi- gods (25).

4.1.3 Nilakantheshvara or Udayeshvara Temple

The temple of Nilakantheshvara or Udayeshvara is the grandest specimen of Paramara architecture by Udayaditya at Udayapur (Near Vidisha) in 1059 to 1080 A.D. It is stellate in plan and a hall with three porches. Great ingenuity has been employed in designing the Shikhara (Cresent) of sanctum, which is decorated with seven vertical and five horizontal, rows of miniature. (26) (Photo.17, 17a)

4.1.4 Siddheshvara Temple

At Nemawar the temple of Siddheshvara is one of the most important ancient monuments of India in Malwa. The temple stands on the bank of the river and has been built on a massive platform of stone. The interior of the main shrine below the shikhara is adorned with numerous decorative carvings and figure sculptures. (27). (Photo18, 18 a)

4.1.5 Sun Temple

The so-called Sun-temple at Jhalarapatan (Dist. Jhalawar at Rajasthan. During that period this region came into Malwa dynasty.) is orthogonal and Saptaratha in plan (Technial term) with a seven storeyed elevation and has a complicated Shikara (Cresent) design. The temple introduces eleborate toranas at the entrance to the porch. (28)). (Photo.19)

4.1.6 Mahakaleshvar Temple

The Mahakaleshvara temple at Ujjain is famous in 12 Jyotirlinga of India. It is mentioned in skand-Puranas Avanti Segment. Its history is related to Satayuga and Treta age but historically it was rebuilt, after the invasion of Iltutmish (Sultan of Delhi at 13 th Century A.D.) by Parmara and Marathas in contemporary age. It is Shiva temple in Bhumija style of temple architecture. (29) (Photo. 20)

4.2 Sculptures of Paramara

In the 10th Century A.D. under the Paramara ruler Bhoja, paramara art was prolific in sculptural output and at the stage of its highest development. Dhar, Mandu, Ujjain, Udayapur, Gyaraspur and Nemawar were the main centers, where there are excellent specimens of Paramara art. These figures were largely and vigorously conceived and were modelled in ample dimensions.

4.2.1 Shiva-Sculpture (God Shiva)

Different types of Shiva images found in Malwa. The dancing Shiva of Jhalarapatan is superb instance of the sculptor's art of the paramara period. An image of Shiva, the lord of dance, belonging to the 11th Centuries A.D. is discovered from Ujjain and is now in aGwalior fort Archaeological museum. Image is dancing in Aindra posture. On the slope of a hill at Udayapur; there is a gigantic unfinished sculpture of Shiva, carved in a single boulder of rock. (Photo.21 , 22).

4.2.2 Vishnu-Sculpture

The extant varieties of Vishnu image found in Malwa. At Udayagiri (district- Vidisha), Dhar, Mandu (Dhar) Dhamanar (Mandsour) Gyaraspur, Bhilsa (Vidisha) and the other large site of Malwa Vishnu images spread-out in numerous. (Photo 23 , 24).

Including other sculptural work we can see the images of Brahma of Modi with four heads, Surya (Sun) image of Gandhwal, deity Saraswati (Vagdevi) of Dhar, Buddha image of Sanchi- God Buddha seated in meditation, Gyaraspur Buddha-Buddha seated in Padmasan. (Photo .25, 26,27,28,29).

5. Medieval & Modern Architecture

Besides these there are numerous medieval and modern structures increasing the beauty of Malwa in which Mandu (Dist. Dhar) is fabulous in natural heritage and famous in India with its Architecture. Mandu was installed as a capital by Husang Shah in 15th Century A.D. There are so many prominent buildings in Mandu like Kamal Mola Mosque, Laat Mosque and Mausoleum of Malik Mugis. The most prominent Architectural buildings are Fort of Mandu, Jama Mosque, Hindola Palace (Swinging palace), Asharfi Palace, and Seven Storey Palace. Victory Piuer, Mausoleum of Husang Shah, Jahaj Palace (Ship Palace), Bajbhadur and Rani Rupmati Palace.

Raja Bhrirtrahari caves, Umbrellas of Durgadas, Kalidas Palace, Kothi Palace are architecture of Medieval and Modern Arts. There is Astrological Observatory in Ujjain installed by Raja Savai Jaysingh (King of Jaipur (Rajasthan) in 18th Century A.D. As a greenwitch of India, Ujjain was famous for determination of time that's why Jaysingh made observatory here. This is one of the five observatories in India.

Rajvara (King Palace made by Holkar Kings in District of Indore and many Umbrellas are importance instances of Maratha Architecture. Jama Mosque of Bhopal and Muslim Structures are of great important architects of Malwa.

6. Management of Digitization

6.1 Digital Collection

Over the past few years there has been an explosion in the number of online Information resources implemented by museums, libraries, archives, historical societies, and other cultural heritage institutions

as they attempt to more aggressively exploit the potential of the web .The benefits of having a rich diversity of quality and authoritative information available online is clear, but the magnitude of that data is meant for many end-users in location of specific, desired resources within the almost overwhelming aggregation of information available. The community continues to struggle in developing new techniques for managing the glut of information and to transform the traditional methods of curator and librarianship for better organization of the available information and make it easier for end-users and to find the specific online information they want.

A digital collection is a group of information items in digital format, related to each other by subject or origin. Digital collections can contain full texts of a wide variety of documents, photographs and images, recordings, videos, or other multimedia A digital collection requires a logical structure, a cataloging or indexing scheme, an archiving policy, and a mechanism by which curators can assess and measure its collection. Digital collections comprise a digital library." In the digital library Collections are transformed through the integration of new formats, licensed content and third-party information over which the library has little or no direct curatorial control. The "Digital Collection" describes a growing collection of original, multi-format content such as:

- Images,
- Data sets,
- Audio/video,
- Text files.

6.2 Images

An image is an online graphic. It may be photographs, line drawings or anything that could be scanned in or created online. Images may be part of a Web page or they may be attached to a hyperlink on a Web page. CONTENTdm is one of the image database software. It was used by WICR (World Civilization Image Repository) and the photos online project. It started using from 1999 onwards as a part of Digital Images. Using this library staff can design the image databases with a range of searching options including Pre-selected searches, such as single hyper-link for a particular search, a drop-down list of search topics, a simple keyword or Boolean- enabled search box, an advanced search engine and the ability to browse all of the objects in a given collection.Collection may also be combined for cross database searching. All of these features can then be placed on a website, with the result that the database interface can be designed for intended audience. The most important thing we need to considered while Digitization of Image is to follow the standards like Dublin Core Metadata .So that we can maintain the changes brought about by the Technologies. By adopting these standards in the open archives initiative ensures that the metadata created in CONTENTdm collection can be harvested by others. This software also allows for several export options, including ASCII and XML.

According to the DiMeMa website, " CONTENTdm provide support for the open Archives initiatives protocol for metadata Harvesting Version 2.0, an emerging standard for Metadata harvesting. So, CONTENTdm servers can function as OAI repositories". It can also burn CD and DVD Disks with full resolution images. Full-Resolution images have been also been made available to the department of general for use by teaching faculty .The metadata from CONTENTdm and the storage of the full resolution images both online and an CD/DVD, we should be ready to migrate the databases to other systems if necessary.

6.3 Photo's Online

This make to create a web-accessible image database for use in website. The photos online oversees the additions of new images complete with metadata to the database.

The FrontPage of photos online includes three search options

- A keyword search
- A dropdown box of predefined searches.
- An advanced search Engines.

Dataset : Datasets are organized collections of related information.

6.4 Audio/Video

Audio has come to mean a method by which sound is recorded digitally and stored. Audio files can be sequenced along with MIDI to form complex arrangements. Many professional musicians and producers use an audio and MIDI combination for the creation of contemporary music.

Video : A moving picture, accompanied by sound. Digital video is useful in multimedia applications for showing real life - such as people talking or real life illustrations of concepts. Here the digital music also has a place. The digital music is an artistic form of auditory communication incorporating instrumental or vocal tones in a structured and continuous manner. The Maine Music Box (MMB) is one of the interactive multimedia digital music library that enables users to view images of sheet music, stores and cover art, playback audio and video renditions and manipulate the arrangement of selected pieces by changing the key and Instrumentation. The library of digital resources which should integral to online music education channel that provides an instructional. The impetus for the endeavor should be a unique collaborative effort within and among diverse institutions and individuals. The music channel should contains metadata and cataloging, music and music education, library science, collection of printed sheet music , scores, graphic design, database design, interactive web programming and network administrations. Through Digitization, musicians, scholars, educators, students and the general public would be able to search textual data and retrieve images of scores or sheet music and cover art, link to the full text of lyrics, hear selected computer generated sound files, and link to other digital conversion of piece .The archive would also be accessible through a web-based Instructional channel integrated with music database.

Digital Conversion: The collection of music scores, manuscripts and sheet music for digitization based on the following criteria. The condition of the original materials, their historical importance and the need to preserve and broaden access to them through digital conversion should be the primary consideration. The other criterion includes:

- Copyright Status,
- Availability of metadata,
- Feasibility of Image Capture,
- Feasibility for second file conversion,
- Relationship to other digital sheet music.

The outsource for the Digitization of the collection will be in the cost effective way:

- TIFF (300 dpi RGB) and bitmap (300dpi-1bit) file formats.
- JPEG images (72 dpi RGB) and thumbnail (115X150 pixels)
- MARC records
- Text of lyrics
- Administrative metadata
- Preservation CDs.

7. Conclusion

Preserving the art architecture, sculptural in the original form is not only difficult but rather impossible. Thus we should work towards processing the contents in different formats. Thus it has to be clearly understood that it is not simply videography or photography. But the advent of the web and other related digital technologies presents a good opportunity for increase content sharing and collaboration in the development of information systems. Making specialized scholarly digital content that is frequently non-textual often hidden within complex database structures and collection contexts more visible and easily accessible requires higher precision search and discovery systems that can exploit richer and more highly structured metadata. Digital preservation expects more care and high expertise for future generation.

8. Suggestions

1. Various agencies engaged in such work should meet and plan for digitization work of art, architecture and sculpture in various regions.
2. Historians at each level should prepare text and mark each object for digitization.
3. There should be an agency in each state, which should be responsible for entire digitization work. There should an apex body at national level for coordination.
4. NGO's should come forward to cooperate in the projects.
5. Sufficient fund may be contributed by various participating agencies.
6. A little tax may be levied from visitors of each buildings of tourist interest to meet financial requirement.
7. NRI can contribute to such a project by providing funds and equipments and technical know how.
8. International agencies like UNESCO may come forward to help the projects.
9. Start now before they perish.
10. A local library should be recognized under the project for collection of material and work as a node for dissemination by providing high broadband connection and large capacity storage devices.
11. Indian Cultural Heritage Information Network (ICHIN) should be developed with the use of ICT.

9. References

1. Hindustan Times. Nov 6, 2004.
2. Yinlu.A memory of old Begging Alleys. Navigating the collections for new services. In: International Conference on Digital Library, New Delhi, 2004. p 418-24.
3. Brown, Heatha: Preserving Cultural Heritage for future generations, hybrid solutions. Ibid. p. 406-17.
4. Jain, K.C. Malwa through the ages (From the earliest Time to1305A.D.). 1972.Delhi, Motilal Banarasidass, .p 103.
5. Morshall John and Alfred Foucher. The Monuments of Sanchi . 3v .p 36-7.
6. Ibid. p36-38.
7. Jain, K. C. Ibid. p 207.
8. Ibid. p.204.
9. Journal of the Madhya Pradesh Itihasa Parishad, Bhopal, No.2, 1960,p 19.

10. Jain, K.C. Ibid. p 208.
11. Morshall John and Alfred Foucher. The Monuments of Sanchi .Delhi. 3 v. p 36-58.
12. Jain, K. C. Ibid. p. 282
13. Upadhyay, Proff, Dr. Vasudev. Ancient Indian Stupa, Guhas & Mandirs.Patna, Bihar Hindi Granth Academy. 3 . p 160-161.
14. Jain, K.C. , IbId. p 294.
15. Ibid. p 283.
16. Ibid.p 290-91
17. Kaval, Ramlal. Temple Architecture in Ancient Malwa.1984.Delhi,Swati..p 122-126.
18. Jain, K C. Ibid. p 423.
19. Kaval, Ramlal. Ibid. p 113-115.
20. Archeaological Survey of India 1905-06.
21. Annual Report of the Archeaological Department, Gwalior State. Gwalior, 1932-33.
22. Jain, K.C. , Ibid. p.432.
23. Jain, K.C, Ibid. p.435
24. Patil, D.R. .The Cultural Heritage of Madhya Bharat, Gwalior, 1952..p94.
25. Archeaological survey of India, Annual Report. Western circle, Poona, Bombay, 1926-27.p.48.
26. Annual Progress Report of the Archaeological Survey of Western Circle. Poona. Bombay. 1914- p.64.
27. Ibid. 1921. p98.
28. Krishna, Deva. Temples of North India.1969.New Delhi. p 56.
29. Kaval, Ramlal.Ibid.
30. Katre, D. S. Pragmatic and Usable Approach for Digital Library Initiatives in India. In: International Conference on Digital Library. 2004. New Delhi, Teri. p 42.
31. Rath, P. N. Digitization as a Method of Preservation of Cultural Heritage. Some theoretical issues. Ibid. p 398.
32. Harianarayana, N. S. Creation of a photogallaery using greenstones: Issues and experiences. Ibid. p1006.

10. Other Referances

1. Nawboodin, V P: Design of Database for Polygonal Models. In : International Conference on Digital Libraries.2004.New Delhi, Teri. p487-494.
2. Rath, P N et.all. Digitization as a method of preservation of Cultural heritage, some theoretical issues. Ibid. p 392-405.
3. Comment, Hearimarc. Archiving Cultural Heritage and History through Digitization, case studies from Russia and Albania.
4. Unison. Memory of the World Programme.
5. Gilland- Swetland, A J and others. Evaluating EAD as an Appropriate Metadata Structure for Describing and Delivering Museum Contents. In: International Conference on Digital Libraries.2004.New Delhi, Teri. p 504-512.

6. Kuffalikar, Chitra Rekha: Rejuvenating Local Historical or Perish. Ibid. p1016.

7 Anil Kumar. M P and Ashok, S. Development and full text Indexing for PDF. Collection

8. Kumar, S and others. Hiring Private Operators in Digitization of Information. In: International Conference on Digital Libraries. 2004. New Delhi, Teri. p. 1024.

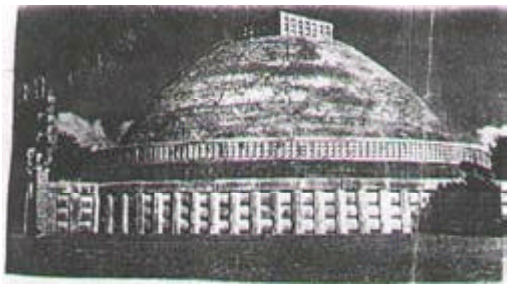
9. Harinarayana, N S and Sunil, M V. Creation of a Photogallery using Green Store: Issues and experiences. Ibid. p1005-06

10. Chopra, H S. Importance of Audio Visual Archives in Preserving Culture Heritage. Ibid. p 994-95.

11. Important Web Sites

1. www.fbi.aiim.wegov2.com
2. www.ansi.org
3. www.nla.gov.ou
4. www.bsi.global.com
5. www.ggbaker.com
6. www.alanhowel.com.au
7. www.ifla.org/VI/4Pao html II 4 (International Preservation News)
8. www.egraph.com
9. http://www.unesco.org/nwhe/
10. www.rit.edu
11. www.iso.ch
12. htttp://64-78.17.7-/India culture/en/aboututs/archive/manuscripts.ntlmus.htm.

Some Heritage Structures Historical)OME HERITAGE STRUCTURES (Historical)







1. Sanchi of Stupa 2.Sanchi North Gate 3.Besnagar Yakshi

About Authors



Prof. S. Kumar is Reader & Head of School of Studies in Library & InformationScience, Vikram University, Ujjain. He holds Master degrees in Arts and Library & information science. Has 33 years of teaching experience to Post Graduates, 1 year to M.Phil and 15 years experience in guiding the research scholars. 6 students have been awarded Ph.D. under his supervision. He has published 90 papers in journals, conferences and seminar volumes. He has received award of Commendations for paper presentation in National Conference. He has two books to his credit. He is life member of various professional bodies. Currently, he is chairman of Board of Studies in Library & information science.



Mr. Mukesh Kumar Shah is a Research Scholar in School of Studies in Ancient Indian History C&A, Vikram Universiry, Ujjain, Madhya Pradesh. He holds B.Sc., M.A. (Ancient History) (Silver Medallists), NET. He has attended 2 National Seminars and also presented papers and also 4 articles have published in reputed journals.



Dr. Leena Shah is Librarian at Govt. College Mehidpur, Madhya Pradesh. She has 45 papers published in various national & International conferences and seminars including the reputed journals. She has presented papers in FID (Jaipur, 1998), and ICDL, (New Delhi, 2004). Besides this, she has attended and presented papers in 5 International & 12 National conferences & 3 seminars. She is Ph.D. in Library Science. She is recipient award of commendation for paper presentation in national conference. She is a life member of 6 national professional bodies. She wrote one bilingual book in 2000. She also attended IFLA Pre Conference and IFLA World Library Conference at Brazil and Argentina in 2004, respectively.

Digital Preservation of Indian Manuscripts - An Over View

Y V Ramana

Abstract

This paper presents a brief over view of Digital Preservation, Digitization of manuscripts and preservation techniques which are currently in use in India. The role of the National Library of India in Digital Preservation of Indian Manuscripts is highlighted. It also deals with the Manuscript Resource Centers and Manuscript Conservation Centers of India. The requirements of Digital Preservation are presented in this paper.

Keywords : Digital Preservation, Manuscripts, Digitization.

0. Introduction

India possesses one of the ancient and richest cultures of the world. India has the largest collection of manuscripts, containing ancient culture and knowledge representing thousands of years of history. The Indian manuscripts, which were written in different languages and scripts are preserved on treated Palm leaves, Birch barks, Silk cloth, Wood, Tamra Patras and hand made paper, inscriptions on stone etc. They are spread all over the country and abroad and are preserved in libraries, museums, temples, Mutts, monasteries etc. These manuscripts contain invaluable knowledge in medicine, science and mathematics, literature, art and architecture, theology, philosophy, music and dance etc. These sources not only provide information on these subjects, but also throw light on the history and culture of the nation.

In the past as a result of natural calamities like floods, wars, fire, and foreign invaders a good collection of old manuscripts were destroyed. Manuscripts and other old documents have been conserved with other artifacts like buildings, sculptures, paintings, monuments etc. Now the concept of preservation has changed. The manuscripts are preserved with the modern digital technology by converting to Analog or Digital copies of the original. At present the preservation techniques are coupled with the word Access, which is to provide information to those who need it in shortest possible time, with the new technologies like Internet, CD-ROM etc.

1. Digital Preservation

Digital preservation refers to a series of managed activities, which are necessary to ensure continued access of digital materials for as long as they are necessary. The term digitization refers to the conversion of material that was originally created in another form to digital form (i.e. which uses a binary numerical code to represent variables). The ultimate goal of preservation is to make the intellectual content to remain in tact as long as possible. The idea of protecting the original documents by reproducing it on a stable media gave rise to digitizing the maps, manuscripts, moving images, music and sounds etc. Digitization of the old and fragile material will not only provide long time preservation but also offers the users to find, retrieve, study and manipulate the information in a colorful environment.

Modern multimedia technology is playing a major role in preservation and promotion of cultural heritage, by digitizing all forms of materials, text, visual, audio/video moving pictures etc. together to represent the holistic form. The World Wide Web is wide reaching medium through which anything and everything could be made available to anyone and everyone around the globe, in fastest way.

2. Digitization of Manuscripts

India has one of the largest and oldest collections of manuscripts in the world. To day the Palm leaf books, paper manuscripts, birch bark texts, drawings, paintings, art and sculpture etc. are either scanned or converted into analog material and are preserved on long lasting digital media for the use of future generations. The most important benefits of digital preservation are:

- ✍ Preservation : Digital reproductions are virtually immortal in the sense, by reproducing multiple digital copies and by putting them for use the originals can be protected. By digitizing the manuscripts, the information can be preserved for a long time on digital media. The paintings and photos etc. of rare manuscripts can be enlarged and reproduction in the colorful environment is possible with digital technology.
- ✍ Dissemination of Information : Most of these manuscripts are stored in museums, libraries, temples and Mutts etc. with a restriction to use them. The digital preservation is not only safeguarding the original documents, but also providing these documents for information dissemination and research purpose via internet and CD-ROM etc.
- ✍ Transcend Originals : Digital imaging promises to generate a product that can be used for purposes that are impossible to achieve with original resources. It uses special lighting to draw out details obscured by aging, use, and environmental damage. Imaging, which makes use of specialized photographic intermediaries or by imaging at high resolution the study of artifactual characteristics has become possible.
- ✍ Collection Management : Digital preservation provides assistance in retrospective cataloguing, researching, assistance with curatorial functions, managing material movement etc.
- ✍ New Revenue Streams : By making available the digital reproductions at lower resolution to scholars as a paid service, sale of high quality posters to art patrons around the world via an e-commerce web site it is possible to generate some revenue.

3. Preservation of Manuscripts in India

Even under the best possible conditions, the physical preservation of manuscripts is a difficult task. The cultural heritage of India, in the form of manuscripts has to be conserved, preserved and documented. With this motivation, from ancient times preservation of manuscripts is done by indigenous methods like wrapping the manuscripts in silk cloth. Some times oil extracts of some natural products, sandal wood powder, black pepper, clove oil etc. are used for preserving palm leave manuscripts. Chemical treatments like, fumigation chambers and Thymol, Chloromate solution are also used to protect the manuscripts.

The Photographic methods, like microfiche, microfilming, photocopying are very important techniques of preservation and access. This method not only may damage the originals but preserve them only for a few decades.

The invention of Scanners has revolutionized the input of data to computer media, which can also damage the manuscripts. Then high-definition film scanner is used to digitize the manuscript as image, which is an expensive method. Before 1998, digital cameras were used, which could copy only a few pages and turns out to be quite expensive.

From 1999, improved still cameras are used to meet the needs of in-house digital copying. National Institute of Advanced Studies (NIAS) used this method by digitizing Bhagavad Gita into two CD-ROMs. The availability of Bhagavad Gita in digital form and its inclusion in computer database has rendered its access through Internet.

NIAS started a new method of preservation called NiDAC, to share the rare manuscripts via the Internet or CDs, for educational and research purposes. Instaed of using a scanner to digitize each page as a computer graphic, the NiDAC procedure begins with the DV video format. The DV video format simply records everything in binary code onto a mini DV tape. The Camcorder connects to high end computers via an IEEE1394 cable and card. The digital image can be manipulated as a graphic or converted into alphanumerical list. Images will be compressed into JPEG image formats and the computerization is completed with various forms of storage like rewritable media, CDs etc. The NiDAC also used a mega-pixel digital still camera with extra large memory cards and this method is one of the cost effective methods and it is ten times faster than downloading via a parallel or serial cable. This method is superior to the DV digitization and also works for extended field trips to archives, if a laptop or a computer with adequate storage is available. The NiDAC procedure allows in house copying of acid paper books such as yellowed and crumbled books. In the NiDAC procedure, DV (Digital Video) video format simply records everything as binary code on to a mini DV tape. DV digitizing method can be utilized for work in remote archives for extended times with no computer access and uncertain power supply etc.

4. Role of National Library of India

The National Library of India, located in Kolkata is collecting, disseminating and preserving the national heritage of the country. Digitization of manuscripts is one of the initiatives the library has taken up with its own holdings. The National Library of India holds the following manuscripts:

✍	Paper manuscripts	3,000 volumes
✍	Correspondence and diaries	250 volumes
✍	Palm leaf manuscripts	334 volumes
✍	Persian	955 volumes
✍	Arabic	681 volumes
✍	Bengali	168 volumes
✍	English	255 volumes
✍	Hindi	5 volumes
✍	Tamil	370 volumes
✍	Sanskrit	790 volumes

100 volumes of Xylographs, comprising more than 800 items (which are presented by honorable Dalai Lama) are block prints made from the bark of rare Nepali trees. The Arabic and Persian manuscripts bear beautiful illustrations, fine calligraphy and elegant binds.

As a sample project the Persian manuscript *Tutinamah* was chosen by the National Library for digitization. This manuscript consists of well known 52 tales of a parrot written in Indian Taliq within gold and color ruled borders and contains colored illustrations made through vegetable and organic dyes, on a hand made paper.

5. The Process of Digitization

The National Library of India carried out the digitization project in two operational areas. They are

Image Capture Station : The Image capture station consists of a digital camera (Nikon D100 with Bayonet mount 28-70mm f/2.8 ED-IF AF-S Zoom-Nikkor lens) mounted vertically on the photographic copystand with side illumination through 40 watts incandescent lamp. The digital camera had special colorimetric filters that enabled the camera to capture a broader spectrum of colors than the scanners.

Image Processing Station : It is a HP Brio Pentium – IV processor, having image processing software like, Kodak, Imaging, Adobe Photo shop –6 etc. The image transfer device, which is connected to the USB port, gathers images from the memory card of the digital camera.

Process : While taking care of the condition of the document, page number orders etc. the lighting is adjusted using light meter.

Image capture : The images were taken for all right handed side pages first, and then left handed pages, in color as uncompressed 8 – bit per channel (24 bit RGB) TIFF files at 300 dpi.

The Image processing

The images were first transferred to the image processing station, where they were renamed as per the page sequence and checked for their quality with the originals. Later they were edited and converted into three basic formats like PDF (Portable Document Format), which can describe documents containing any combination of text, graphics and images in a device independent and resolution independent format, TIFF (Tagged Image File Format), which is a file format used for still – image bitmaps, stored in tagged files) JPEG (Joint Photographic Exports Group). The JPG file is small and compressed by 90% of the original size. Finally, the manuscript is put in E-Book Format, in which the PDF image files were tagged and a composite PDF file is prepared as per the original document pagination and sequence. The composite PDF containing the individual pages were in E-Book form, with the object of access. The images were stored in CD-ROM and were made resident in hard disk of the central server.

6. Manuscript Resources of India

✍ Total number of Manuscripts in India	5,000,000
✍ Indian Manuscripts available in European countries	60,000
✍ Indian Manuscripts in Asian countries	150,000
✍ Percentage of manuscripts in Sanskrit	67%
✍ Percentage of manuscripts in other Indian languages	25%
✍ Percentage of manuscripts in Arabic /Persian/Tibetan	8%
✍ No. of manuscripts recorded in catalogue	1,000,000

Out of these Indira Gandhi National Center for Arts (IGNCA) has 2,50,000 manuscripts, Indian National Trust for Art and Cultural Heritage (INTACH) has surveyed more than 300 sites in 3 districts and prepared an inventory of 47,000 palm leaf and paper manuscripts.

7. Mission for Digital Preservation of Manuscripts

The National Mission for manuscripts was launched by the Department of Culture, Ministry of Tourism, Government of India, with the Indira Gandhi National Center for Arts as the national nodal center to save the India's most valuable heritage. It selected four agencies for digitizing manuscripts in 5 states of India. They are :

1. NIC, New Delhi, Illustrated Manuscripts of Orissa.
2. MSP, Bangalore, Siddha Manuscripts of Tamil nadu.
3. CDIT, Delhi, Kuddiyattam Manuscripts of Kerala
4. CIL, New Delhi, Kashmir Manuscripts of Ikbal Manuscript Library.
5. NIC, New Delhi, Vaishanava Manuscripts of Majuli Island, Assam.

8. Manuscript Resource Centers

The Mission has selected 24 Manuscript Resource Centers (MRC) in the country for coordinating its activities pertaining to survey and documentation of manuscripts. These centers are registered libraries, museums, oriental institutions, universities etc. with considerable manuscript holdings and necessary infrastructure to provide support for survey, documentation and digitization of manuscripts. Strategies for cataloguing, preservation and storage of manuscripts is drawn in consultation with the experts and evolved a standard format for the preparation of a comprehensive national electronic registration of manuscripts.

9. Manuscript Conservation Centes

The mission has identified 15 Manuscript Conservation Centers (MCC) all over the country as nodal centers for preservation and conservation of manuscripts. MCC are to provide training in preservation and conservation, taking up the task of conservation of manuscripts in different institutions, work for introducing the new technologies for conservation of manuscripts. Some of the MCCs are:

- ✍ Indira Gandhi National Center for Arts, Delhi
- ✍ Orissa Art Conservation Center, Bhubaneswar
- ✍ Rampur Raja Library, Rampur, U.P
- ✍ Saraswati Mahal library, Thanjavur
- ✍ Salarjung Musuem, Hydarabad
- ✍ Khuda Baksh Library, Patna
- ✍ Rajasthan Oriental Research Institute, Jodhpur
- ✍ Oriental Institute, M.S. University, Baroda
- ✍ INTACH, Chitrakala Parishat, Bangalore
- ✍ Bhandarkar Oriental Research Institute , Pune. Etc.

All these centers are well equipped with conservation laboratories and have expertise in providing conservation and preservation services for manuscripts.

National Electronic Register: The mission started for evolving a National Electronic Data base of manuscripts, which integrates information regarding different aspects of manuscripts like, type of material, script, language, subject, place of availability, illustrations, number of pages etc. The database integrates information regarding the organizations with their manuscript holdings, catalogues in the country and abroad, along with the bibliographical details of the manuscripts. National Informatics Center prepared the software and data for 20,000 manuscripts are entered in the database. A copy of the software is given to MRCs and the work of entering the Meta data is started in these centers.

10. Institute of Asian Studies Project

A study by Institute of Asian Studies, Madras indicates that there are about one hundred thousand palm leaf manuscripts, in Tamil language in south Indian Repositories and are lying in a destroyed shape. These manuscripts are related to subjects like Siddha, Ayurveda, Yunani, Human anatomy, Art & Architecture, Temple art, Ship building, Carpentry, Metal working, Astrology & Astronomy, Yoga, Martial arts, Physiognomy etc. The institute deputed a team of highly qualified specialists for the task of preservation of these manuscripts. As a first step the team identified, collected and conserved the manuscripts and then microfilmed and preserved the manuscripts. In the second step, these manuscripts are translated, edited and catalogued.

Digitization : An international team of scholars from Germany, University of Cologne, University of Berkeley, U.S.A, are working in collaboration with Online Tamil Lexicon project, to digitize the manuscripts and after digitization, they will be disseminated as online databases, CD-ROMs and conventional publications as books.

11. Digital Preservation Requirements

Digital preservation encompasses a broad range of activities designed to extend the usable life of machine readable computer files and protect them from media failure, physical loss and obsolescence. Digital preservation will add little values to the research process if it serves only as an alternative form of storage. Preserving digital materials in formats that are reliable and usable will require long term maintenance of structural characteristics, descriptive Meta data, display and computational and analytical capabilities, which demand mass storage and software for retrieval and interpretation. The digital preservation is a process that requires the use of the best available technology, careful thought, administrative policy and procedures.

12. Conclusins

Preservation of manuscripts is not new in India. Along with traditional methods of preservation, modern techniques of digital preservation are also adopted. The Government of India is trying to preserve its cultural heritage by proposing strategies and policies at national level. By giving the responsibility of the conservation and preservation to National Library, National Informatic Center, National Archives and many individual libraries and information centers in the country. But the invaluable manuscripts of India are scattered among libraries, museums, temples, individuals etc. of the country and aboard. Therefore, it is the responsibility of each of these institutions to preserve them with modern digital technologies. The present technology and available expertise is enough to digitize the existing manuscripts, but one of the important limiting factors is motivation and monetary support, which should come from private business houses, religious bodies and individuals and a system of sharing the benefits should be worked out among owners of manuscripts, sponsors and universities.

13. References

1. www.library.cornell.edu/iris/tutorial/terminology/preservation.htm
2. National mission of manuscripts, namami.nic.in
3. http://gistnic.ap.nic.in/san/
4. www.ndl.go.jp./en/publicaton/cdnlao/047/473.html
5. http://xlweb.com/heritage/asian/index.htm
6. http://www.tifac.org.in/abt/ab/.htm

About Author

Mrs. Y V Ramana is currently working as an Assistant Librarian at Vellore Institute of Technology, Vellore in Tamilnadu.She has about 12 years of experience as Assistant Librarian in various Institutes of Engineering & Technology. She holds MA (Soc.), MA (AIH&C), MLISc. She has contributed number of papers in seminars and conferances.

E-mail : yvramana58@yahoo.co.in

A Novel Approach for Document Image Mosaicing Using Wavelet Decomposition

P Shivakumara G Hemantha Kumar D S Guru P Nagabhushan

Abstract

There are some situations where it is not possible to capture or scan a large document with given imaging media such as scanner or Xerox machine as a single image in a single exposure because of their inherent limitations. This results in capturing or scanning a large document image into number of split components of a document. Hence, there is a need to mosaic the several split components into a single large document image.

In this work, we present a novel and simple approach for mosaicing the two split document images using wavelet decomposition to generate single and large document image. The proposed method uses the wavelet decomposition to speed up the mosaicing process by means of Multi Resolution Analysis (MRA). The pixel based and Column –Block matching procedures are used here to identify the overlapping region in the split images. The overlapping region is a common region which helps in obtaining mosaiced image from its split images. The proposed methods work based on assumption that the overlapping region is present at the right end of split image1 and at the left end of split image2

Keywords : Wavelet decomposition, Pixel value matching, Column-Block matching, Overlapping region, Document image mosaicing.

0. Introduction

The concept of image mosaicing is a phenomenon that occurs in the vision system of human beings because the human brain mosaics the split images of a large object that are automatically captured through eyes. Each eye functions as a camera lens. But, it is impossible to cover very large area with the help of an eye than a pair of eyes. Keeping this in mind, one can infer that two eyes capture the two split images of a large object but essentially with certain amount of overlap between the split images, which are later mosaiced into a single complete large image by deriving the knowledge from the Overlapping Region (OLR). Similarly, even in the real world the concept of mosaicing is essential because it may not be possible to capture a large document with a given camera or a Xerox machine in a single exposure. It has got to be captured as two or more split images due to the inherent limitations of the capturing medium. In such man made multiple camera exposures to cover a large image, the split images should necessarily contain OLR between the images, so that the stitching of two or more such split images into a single image becomes easier. Therefore, the proposed technique demands small amount of OLR in the split images such that the OLR is present at the right end of the first image and the left end of the second split image respectively.

There is a great demand for developing an algorithm for mosaicing the split images obtained by scanning of the large document part by part in order to restore original and large document image.

The structure of paper is as follows. The proposed methodologies are discussed in section 2 with suitable algorithms and mathematical models. In section 3, the comparative study is given. The experimental results are reported in section 4. Finally, the conclusion is given in section 5.

1. Document Image Mosaicing

Mosaicing is defined as the process of assembling the multiple components that are obtained either by scanning or capturing a large document part by part, in order to restore an original image without any duplication of portions. For example, a Xerox machine handles the documents of sizes A_4 (210mm X 297mm) and A_3 (297mm X 420mm). But the document of sizes A_2 (420mm X 594mm) such as a full newspaper cannot be scanned in a single stroke because of its inherent limitation. Hence, bigger sized documents such as newspapers have got to be split into number of smaller documents of A_4 or A_3 dimensions with little overlap between the split images.

Several researchers addressed the methods for obtaining the large image from its split images. (Schutte and Vossepoe, 1995) described the usage of flat bed scanner to capture large utility map. The method selects the control points in different utility maps to find the displacement required for shifting from one map to the next. These control points are found from pair of edges common to both the maps. However, the process requires human intervention to mask out the region not common to both the split images in image mosaicing.

The researchers (Zappala et al., 1997; Peleg, 1997) have worked on Document Image Mosaicing (DIM). A feature-based approach through estimation of the motion from point correspondence is proposed. They have exploited domain knowledge, instead of using generic corner features, to extract a more organized set of features. The exhaustive search adopted was computationally expensive because of the rotation of an image employed during matching. However, the approaches are limited to only text documents and are prone to failure in case of general documents containing pictures. But in practice, a typical document contains both text and pictures.

An automatic mosaicing process for split document images containing both texts and pictures, based on correlation technique is proposed by (Whichello and Yan, 1997). Here correlation technique was used to find the position of the best match in the split images. However, accuracy is lost at the edges of the images. Moreover, the correlation of two images of practical size is computationally very expensive. In order to find a solution, additional constraints like *a priori* knowledge were introduced. Here, the sequence in which the images were captured and their placement (generally, referred as image sequencing) is known. Template matching procedure was used to search OLRs, present in the split document images. Usually, template-matching procedure is a time consuming method. In addition, this approach assumes that the printed text lies on straight and horizontal baselines, which is not always possible in many of the pragmatic applications.

2. Proposed Methodology

The authors of this paper have proposed (Shivakumara et al, 2001) a new technique to tackle the above-mentioned problems. The proposed technique works for any type of document without considering the nature of the content present in the document to produce a complete large document image without having *a priori* knowledge about the order of image sequence. The proposed technique demands at least one pixel wide (1-2%) OLR in the split images. The OLR is present at the right and the left ends of the first and the second split images respectively. The technique is based on PMA (Pattern Matching Approach). A PMA is employed to determine the OLR in the split images of a large document image. The order of image sequence is obtained by considering the split images in all sixteen possible ways of matching sequence exist between them, as each image is associated with four faces. Sixteen possible ways of matching sequence exist if we consider the perfect square shape images. Out of sixteen possible matches, some of them are accepted as possible mosaiced images based on the overlapping region existing in the split images. Subsequently, the original complete image of the document is obtained by

mosaicing the split images without any duplication of portions in the mosaiced image. The presented technique requires 16 O(n2) + 16n comparisons search time for finding the right sequence for two split images under the worst case, where n is the length of String of Column Sums (SCS). However, this method is time consuming method.

We propose two methods one is simple and pixel based and Column-Block matching procedure to identify the overlapping region in the split images to produce a single large document image. The two methods work based on wavelet decomposition. The proposed method is based on assumption that the overlapping region is present at the right end of split image1 and at the left end of split image2. The following sub sections explain the wavelet decomposition in detail, simple algorithm for mosaicing and column-block matching procedure for mosaicing.

2.1 Wavelet Decomposition

The wavelet transform of a 2D image f (x, y) is defined as the correlation between the image and a family of wavelet function { $ø_{s,t}$ (x, y)}: W_f (s , t; x, y) = f (x, y) * $ø_{s,t}$ (x, y) .Wavelets are generated from a mother wavelet function as follows: $Ø_{s,t}(x, y) = 1/s\ ø\ (x – t_x/s, y – t_y/s)$ Where s is the scale parameter, and (t_x, t_y) the translation parameters in the x axis and y axis. in most practical applications, one never explicitly calculates the mother wavelet. The pyramid-structured wavelet decomposition operation (mallt, 1989) produces four subimages f_{LL} (x, y), f_{LH} (x, y), f_{HL} (x, y) and f_{HH} (x, y) for one level of decomposition. f_{LL} (x, y) is a smooth sub image, which represents the coarse approximation of the image. f_{LH} (x, y), f_{HL} (x, y) and f_{HH} (x, y) are detail subimages, which represent the horizontal, vertical, and diagonal directions of the image, respectively. The 2D pyramid algorithm can iterate on the smooth subimage f_{LL} (x, y) to obtain four coefficient matrices in the decomposition level. Fig.1 depicts one stage in multi-resolution pyramid decomposition of an image (Tsai and Chiang, 2002).

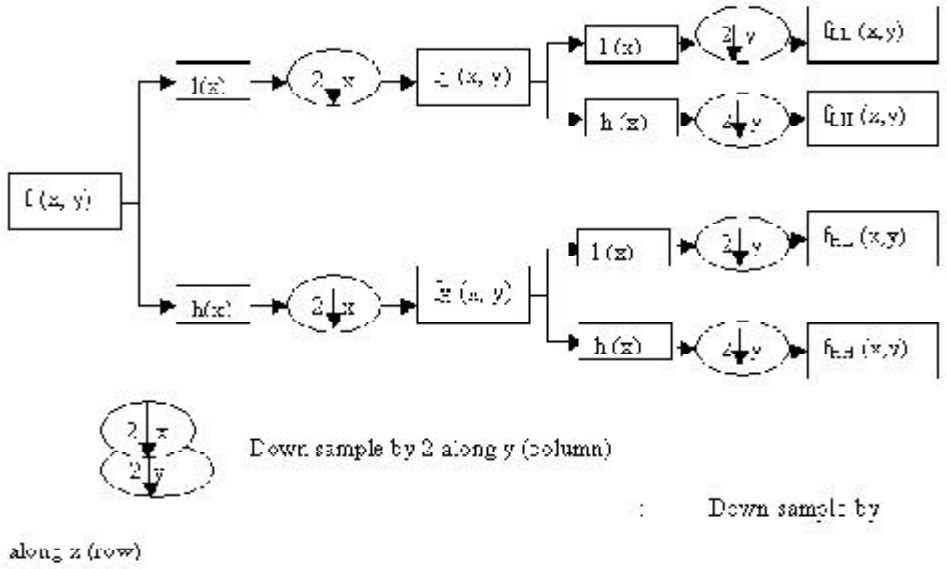


Fig.1 One stage in multi-resolution image decomposition

The reduction factor of an image size is given by 4j, where j is the number of decomposition levels. For level-2 decomposition, the size of an original image can be reduced by a factor of 16. This results in great computational saving in the matching process. In practice, the effective size of the smallest subimages in the decomposition should be used as a stopping criterion for determining the maximum number of decomposition levels. If the decomposed subimage has an over-down sampling size, the locations and wavelet coefficient values of object features may change dramatically from one sample to sample, and generate a false match accordingly. The experimental results on a variety of test images showed that the smallest size of decomposed template subimage should be larger than 20X20 pixels. The matching process can be performed either on the decomposed smooth subimage or on the decomposed detail subimages at a lower multi-resolution level. In this study, we consider decomposed smooth subimage for mosaicing purpose.

2.2 Simple Pixel based Algorithm for Mosaicing

This section presents a simple approach to generate mosaiced image from its split images containing wavelet coefficients at one level. The wavelet coefficients are obtained by the above section (section 2.1). Let S_1 and S_2 be the split image1 and split image2. The method compares the values of pixels of First column (F_c) of S1 with First column of S_2. If match occurs then the pointer (i) pointing column of S_1 moves to next pixel value in the same column. Similarly, in S_2 also the pointer (j) moves to next pixel value of same column. If it doesn't match then the pointer moves to next column mean while the pointer of S_2 comes back to first column. This procedure repeated till all the columns match continuously since once overlapping starts it ends at the end of image in S_1. If whole column match (CM) in S_1 and S_2 then both the pointers go to next column. The number of matching columns decides the overlapping region in the split image. Finally, the algorithm terminates when the pointer of S_1 reaches n where n is the number of column in S_1 without overlapping region. The algorithm also terminates if pointer of S_1 reaches n with overlapping region.

Algorithm: Simple

Input: Split image1 (S_1) and Split image2 (S_2)
Output: Mosaiced image
Method:

Step1: For each Column (C) of S_1 and S_2
For each pixel value of C of S_1 and S_2
If ($P_i = P_j$) in C_i of S_1 and S_2 (Where P_i is the pixel values of column of S_1 and P_j is the pixel values of column of S_2)
i = i + 1 and j = j + 1 (i is pointing to Column values of S_1 and j is pointing to column values of S_2)
if ($P_i = P_j$) and (i = E_c of S_1) (here E_c is represents the end of column in)
CM = 1 (if the whole column matches in S_1 and S_2)
Else exit from the for loop
else exit from for loop
Step2: If (CM = 1)
$C_{s1} = C_{s1} + 1$ and $C_{s2} = C_{s2} + 1$
Else $C_{s1} = C_{s1} + 1$ in S_1 and Cs2 in S_2 comes back to F_c
(C_{s1} is the pointer pointing to Column of S_1 and C_{s2} is the pointer pointing to column of S_2)
For end
Step3: If (CM = 1) and (C_{s1} = n) then OLR = 1
Else OLR = 0
If (OLR = 1) Mosaic the split images
Else algorithm terminates with out overlapping region
Method ends
Algorithm simple ends

2.3 Algorithm Column-Block Matching Procedure for Mosaicing

This section presents an algorithm Column-Block matching procedure for mosaicing of two split images containing the wavelet coefficients to produce single large document image. Let S_1 and S_2 be the given two split images containing local FT coefficients obtained by the above algorithm (section 2.1) The algorithm begins by matching the pixel values of F_c (First column) of S_1 with F_c of S_2. IF match is found then it goes to next pixel values of corresponding columns of S_1 and S_2. After finding whole column match (CM) the algorithm considers rest of the portion in the split images as a block from next to CM to end of S_1. Similarly in S_2 also. Next the method computes total sum of the values of pixels in both the blocks of S_1 and S_2. If sums are match then that portion is considered as actual overlapping region in the split images. If the pixel values in the column or sums do not match then the pointer C_p pointing to S_1 moves to next column mean while the pointer C_p pointing to S_2 comes back to F_c. This is because of assumption that the overlapping region is present at the ends of the split images. That means the overlapping region in S_1 begins at middle column and in S_2 the overlapping region begins from first column of S_2. The algorithm terminates when C_p of S_1 reaches n where n is the end of column of S_1 without overlapping region. The algorithm also terminates if the overlapping region is found in the split images.

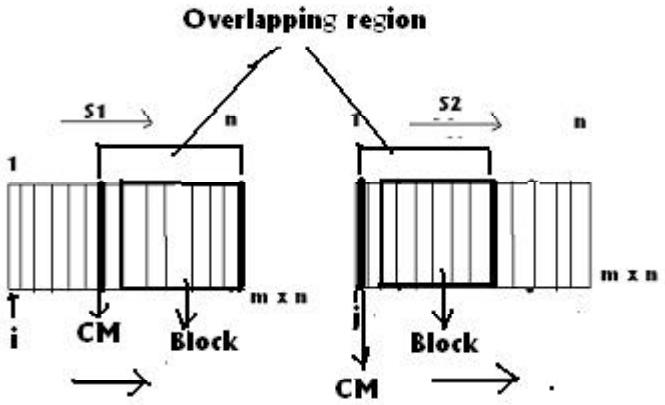


Fig. 1. The method to finds overlapping region in the split images

In Fig. 1, the Column Match (CM) denotes the matching column in the split images and Blocks denotes rest of the overlapping region in the split images. i and j are the pointers pointing to split image 1(S_1) and split image 2 (S_2). The actual overlapping region is represented by both CM and Blocks of split images.

Algorithm for CB
Input: S_1 and S_2 containing local FT coefficients
Output: Mosaiced image
Method:

Step1: For each Column (C) of S_1 and S_2
For each pixel value of C of S_1 and S_2
If ($P_i = P_j$) in C_i of S_1 and S_2 (Where P_i is the pixel values of column of S_1 and P_j is the pixel values of column of S_2)
i = i + 1 and j = j + 1 (i is pointing to Column values of S_1 and j is pointing to column values of S_2)
if ($P_i = P_j$) and (i = E_c of S_1) (here E_c is represents the end of column in)
CM = 1(if the whole column matches in S_1 and S_2)

Else exit from the for loop
else exit from for loop
for end
Step2: If (CM = 1) then B_1 = N – CM + 1th = W in S_1
B_2 = CM + 1th to W in S_2 (B_1 represents the Block of S_1,
N is number of column in the S_1, W is the width of Block of S_1 and B_2 represents the Block of S_2)
Else C_p = C_p + 1 in S_1 and C_p in S_2 comes back to F_c (C_p is the pointer pointing to Column of S_1 and S_2)
For end
Step3: For B_1 of S_1

$Sum_1 = \sum_{p=1}^{N} \sum_{q=1}^{M} B_{pq}$ (Where p and q are the pointers of Block and N is the number of rows in Block and M is the number of column in the Block)

For B_2 of S_2

$Sum_2 = \sum_{p=1}^{N} \sum_{q=1}^{M} B_{pq}$

Step4: If (Sum_1 = Sum_2) then OLR = 1 (overlapping region is found)
Else OLR = 0
If (OLR = 1) Mosaic the split images
Else if (i = n) algorithm terminates with overlapping region is not found
Method ends
Algorithm ends

3. Comparative Study

In this section, we present comparative study of two methods by considering time for obtaining wavelet coefficients, time for matching area is to be found and number of comparisons with respect to levels for particular data set.

Table 1. Comparative study of Simple and Column-Block method with respect number of comparisons and its time in second

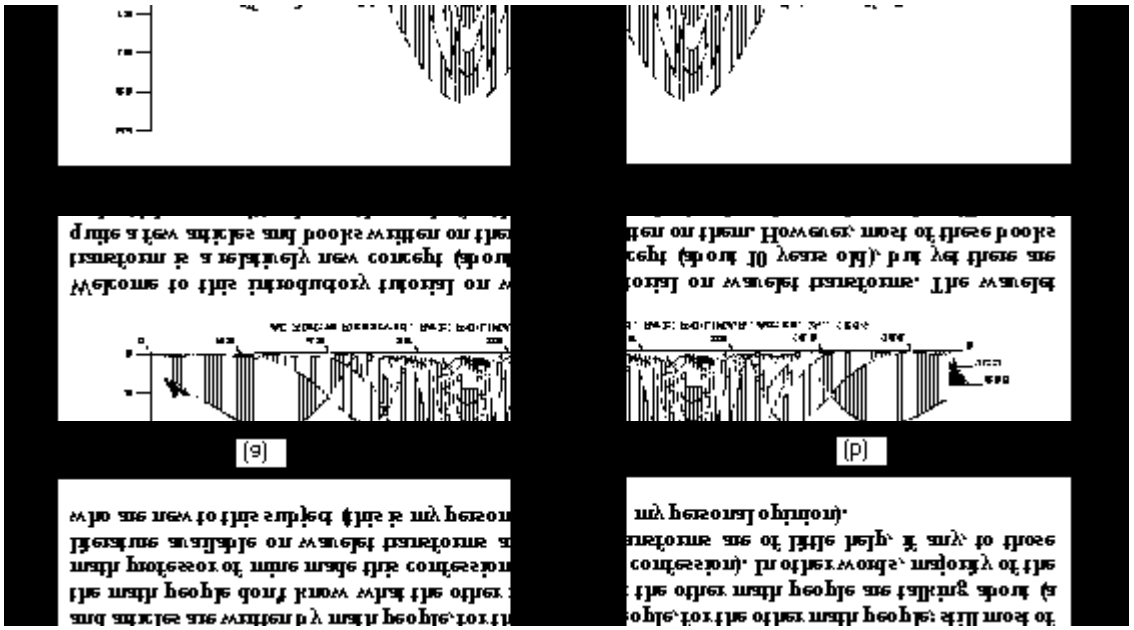
Levels	Column-Block method			Simple (pixel based)		
	TFW	TC	No.C	TFW	TC	No.C
128X128(I)	0.49 sec	0.05 sec	27	0.55 sec	0.05 sec	87
64X64(II)	0.60 sec	0.05 sec	13	0.87 sec	0.05 sec	43

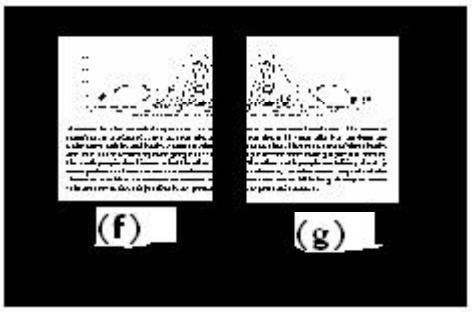
From the above table, it is observed that the column-block matching procedure takes very less number of comparisons compared to simple method in both the levels. In the above table TFW denotes time for obtaining wavelet coefficients at first level, TC means time for comparisons i.e time required for getting a matching area in the split images and No.C means number of comparisons required to identify the actual overlapping region in the split images. With this we conclude that the column-block matching procedure is better than simple method in all the way. This is because in column-block method after whole column matches it requires one match to decide overlapping region whereas in simple method every values in the columns should be compare to decide overlapping region.

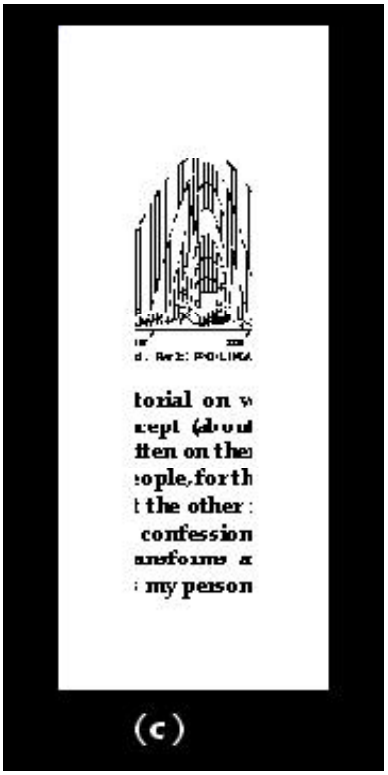
4. Experimental Results

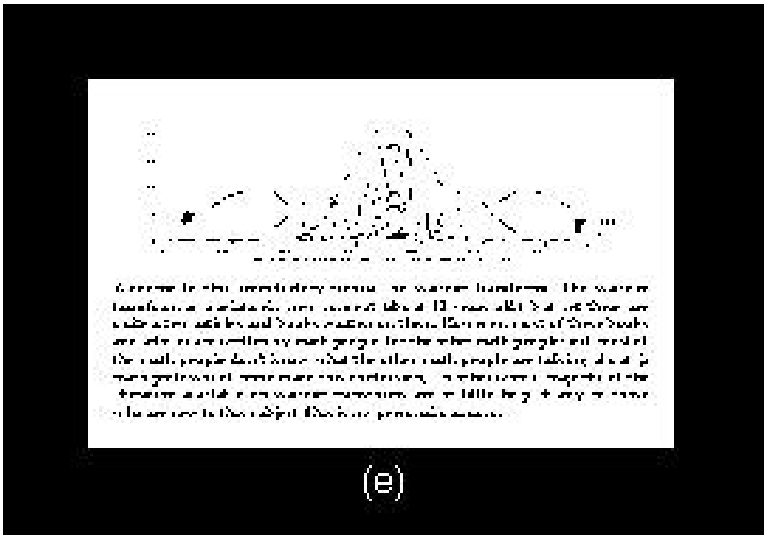
In this section, we present experimental results based on proposed methodology. We have conducted several experimental results out of them a few of are presented here. For both the methods we get same results. Therefore, we have given only one data set for both the algorithms. The experimental results showed that the proposed method work for any type of documents. In the following Examples Figs. (a) and (b) are the input images, Figs. (f) and (g) are the results of wavelet decomposition at one level, Fig. (c) is the overlapped images and Fig. (e) is the mosaiced image in wavelet domain and Fig. (d) is the reconstructed and original mosaiced image. For sample we have given wavelet decomposition effect for only one data set. In all examples, the proposed methods use the reduce sized split images. After finding control points of overlapping region the methods reconstruct the original image (output image) using inverse wavelet transform.

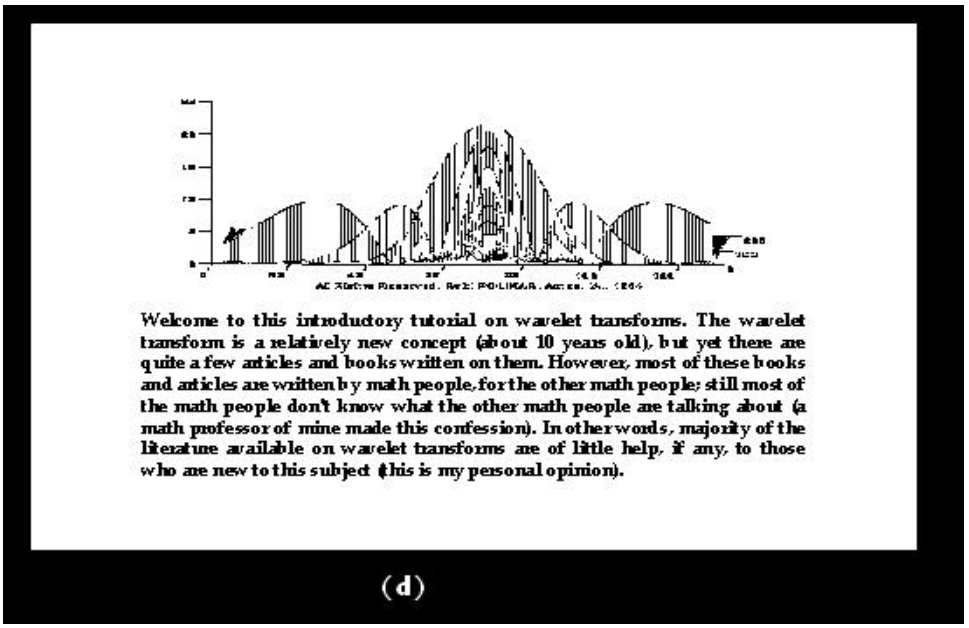
Example1: Here, the split images contain text with graph



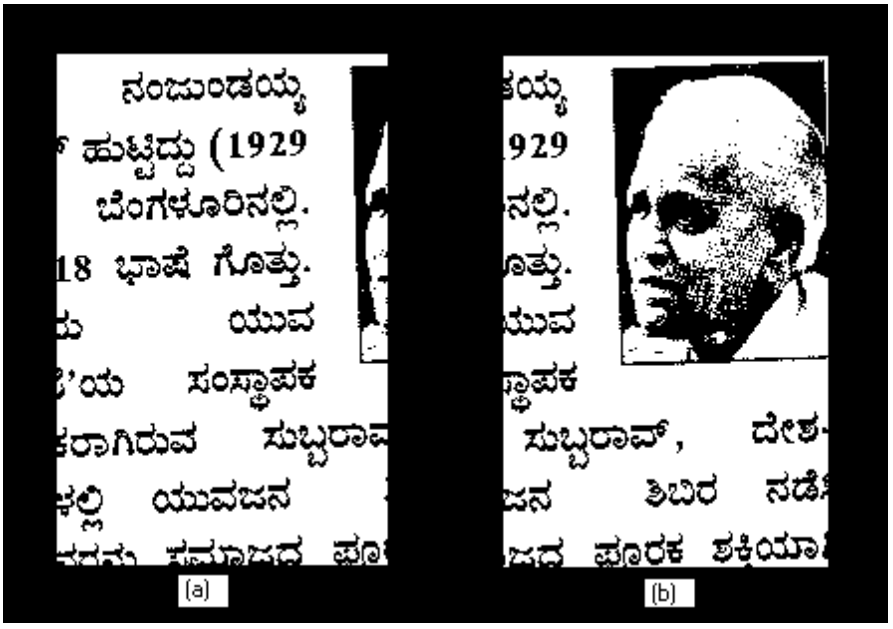


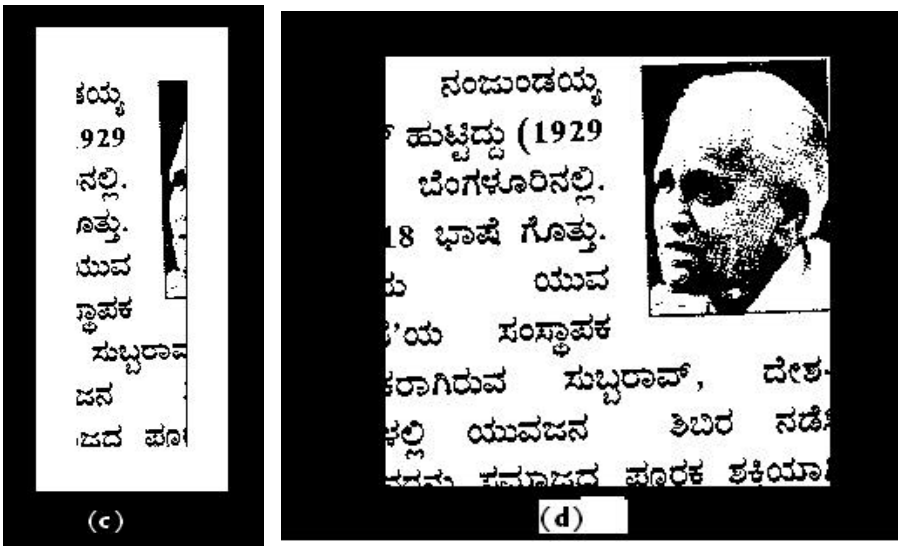

(c)


(e)

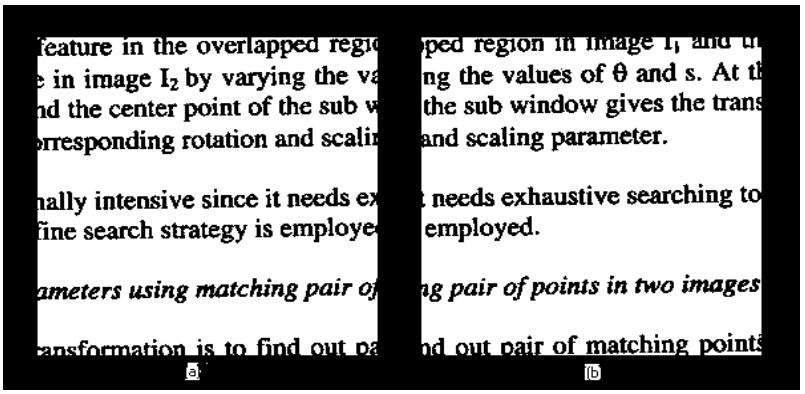


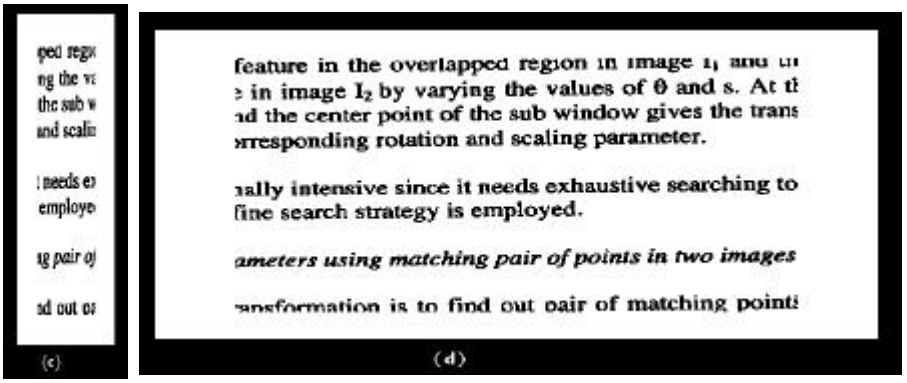
Example 2: Here, the split images contain Kannada with picture





Example 3: Here, the split images contain only English text

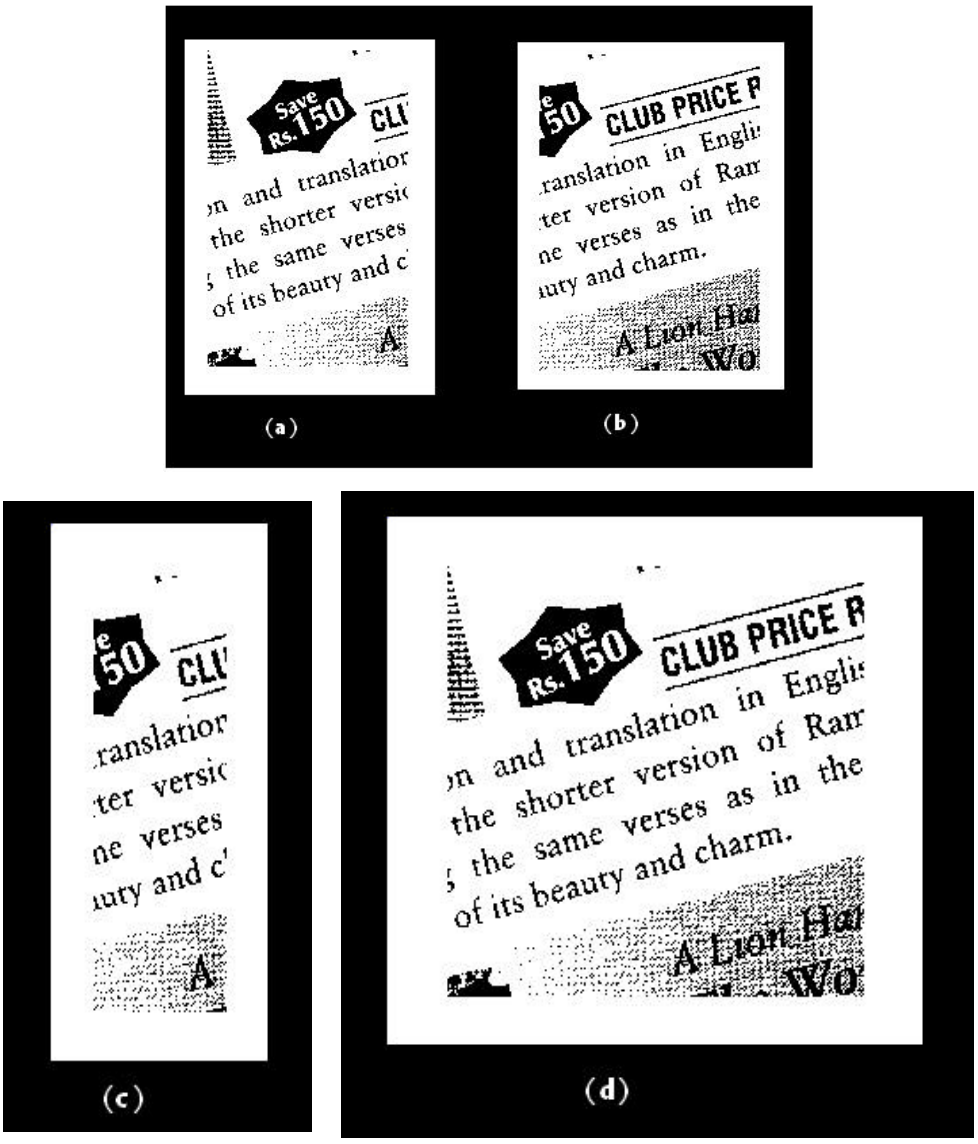




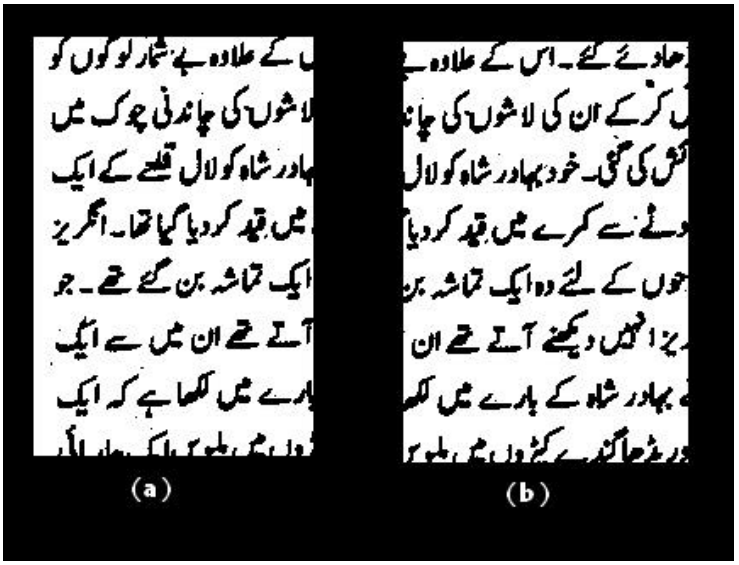
Example 4: Here, the split images contain Malayalam language

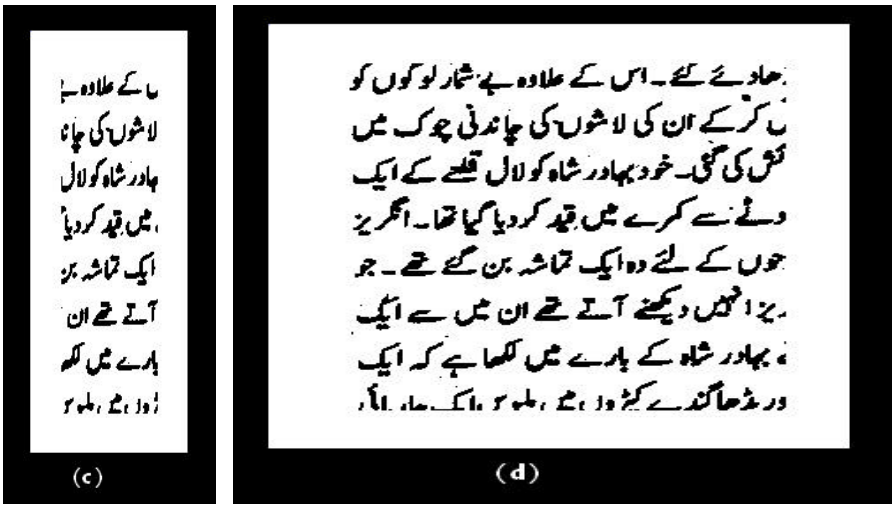


Example 5: Here, the split images contain English text with pictures



Example 6: Here, the split images contain the Urdu language





6. Conclusion

The wavelet decomposition based methods are presented in this paper. The comparative study of two methods is also given. The proposed methods take very less computational burden since the methods involve the multi resolution analysis property of wavelet. The experimental results and comparative study showed that the column-block matching procedure is better method in solving real world applications compared to simple method. The proposed methods assume that the overlapping region is present at the right end of the split image 1 and at the left end of the split image 2. The proposed methods fail when the split images are rotated, scaled and skewed differently. In addition, the methods fail to mosaic unequal sized split images.

7. Acknowledgement

The authors acknowledge the support extended by Dr. D.S Ramakrishna, Principal, APS College of Engineering, Somanahalli, Bangalore – 82 and Mr. Bhavani Shankar Hiremath

8. References

1. Shivakumara P, Guru D. S, Hemantha Kumar G and Nagabhushan P, *Document Image Mosaicing: A* Novel Technique Based on Pattern Matching Approach. Proceedings of the National Conference on Recent Trends in Advanced Computing (NCRTAC-2001), Tamil Nadu, Feb 9-10, 2001, pp 01-08.
2. Shivakumara Guru D. S, Hemantha Kumar G and Nagabhushan P, Pattern Matching Approach based Image Sequencing useful for Document Image Mosaicing. Proceedings of the National Conference on Document Analysis and Recognition (NCDAR-2001), Mandya, Karnataka, July 13-14, 2001.
3. Shivakumara Guru D. S, Hemantha Kumar G and Nagabhushan P, Mosaicing of Color Documents: A Technique based on Pattern Matching Approach. Proceedings of National Conference on NCCIT, Kilakarai, Tamilnadu, 24th and 25th September, 2001, pp 69-74.
4. Shivakumara Guru D. S, Hemantha Kumar G and Nagabhushan P, Mosaicing of Scrolled Split Images Based on Pattern Matching Approach. Proceedings of Third National Conference on Recent Trends in Advanced Computing (NCRTAC – 2002), Tamil Nadu, Feb 13-15, 200
5. A drian Philip Whichello and Hong Yan Document Image Mosaicing, Imaging Science and Engineering laboratory, Department of Electrical Engineering, University of Sydney, NSW 2006,1997.
6. Shmnel Peleg, Andrew Gee, Haifa Research Laboratory, Virtual Cameras usingImage mosaicing, Herbrew University, October 1997
7. Zappala A.R, Gee A.H and Taylor M.J Document Mosaicing. In proceedings of the British Machine Vision Conference, volume2, pages 600-609, Colchester, 1997
8. Tsai and Chiang, Rotation-invariant pattern matching using wavelet decomposition, Pattern Recognition Letters, 23, pp 191-201, 2002.

About Authors



Shivakumara P, obtained B.Sc, M.Sc. and M.Sc. Technology by research in Computer Science from the University of Mysore in the year 1996, 1999 and 2001 and his B.Ed. degree from Bangalore University in the year 1997. He has submitted his Ph.D dissertation to the University of Mysore for the award of Doctoral Degree. He has participated in 13 National/ International Conferences/Workshops. He has published around 60 papers in various journal/conferences on Document image Mosaicing, Skew Detection, Character Recognition and Automatic Face Recognition. His research focuses on Pattern Recognition and Image Processing in general, Document Analysis and Document Image Mosaicing in particular. Currently he is working as Assistant Professor in APS College of Engineering, Bangalore.



G. Hemantha, obtained M.Sc and P.h. D in computer Science from the University of Mysore in the year 1988 and 1998. He has published around 80 research papers in various journals/conference proceedings. Currently he is guiding 6 P.h.D candidates and 4 MSc Technology by research candidates. He has delivered talk at many Universities in India and abroad. His research focuses on Pattern Recognition and Image Processing, Document Analysis, Speech Processing, Computer Network and Simulation. He is currently the Reader and Chairman, Department of Studies in Computer Science, University of Mysore.



D. S. Guru, obtained B.Sc. degree, and M.Sc degree in Computer Science from the University of Mysore in 1991 and 1993 respectively. The same University awarded him the Ph.D degree for his work in the field of object recognition in the year 2000. Currently he is the senior faculty member in the Department of Computer Science in University of Mysore. He holds rank positions both at B.Sc and M.Sc levels. He has been identified as reviewer for journals and conference proceedings. He is currently guiding 6 Ph.D and 4 M.Sc Technology (by research) scholars. He has authored about 60 research papers in international/national journals and conference proceedings. His areas of research include Image Processing, Pattern Recognition, Advanced Software Engineering, Data mining, Image retrieval, Object Recognition, Digital Image Processing.



P. Nagabhushan, F. I. E obtained B.E from University of Mysore, M.Tech from Birla Institute of technology and then Ph.D from University of Mysore in the years 1980, 1983 and 1988. He is currently the professor, Department of Studies in Computer Science, University of Mysore. He was the dean of Faculty of Science and Technology, University of Mysore and is working as an advisor to many other Universities in India and abroad. He has authored about 230 research papers. Currently, he is guiding 7 Ph.D and 4 M.Sc Technology (by research) candidates. He has been a visiting Professor at a few abroad Universities. His areas of research include Pattern Recognition, Image Processing, Dimensionality Reduction, Data Mining, Document Analysis, Simulation.

Enhanced Information Retrieval

R Bhaskaran

Abstract

This paper presents a Novel Scheme for improved Information Retrieval. The Proposed concept of this paper "FRIEND AGENT" is to enhance the need for Effective Information Retrieval. It helps in gathering the required information Independent of the location. Friend Agent offers Network Bandwidth Consumptions, Optimal routing, Security, Effective Data Transfer.

Keywords : Information Retrieval, Friend Agents, Mobile Agents, Friend Network

0. Introduction

The advances in Computing Technology have produced Miniaturization. The rapid deployment of wireless Communication Technologies such as Cellular Network, Wireless Ad-hoc Networks support the universal connectivity.

Overcoming the challenges facing the network can help access the Global network easily and efficiently. The most important problem concerning mobile networking is Quick Access to Information independent of its Location.

The other concern to be considered is the network security as constant location changes of mobile station occurs. Authentication is needed between mobile and base station. It also requires immediate update of the changes in the location.

In this Paper, We Propose a Novel and Noble scheme for Retrieval of Information and Routing Information Update based on mobile traffic pattern. We begin Section 2 with Mobile Agent Concept. In Section 3, we present the concept of Friend Network. In Section 4, we introduce the concept of "Friend Agent". In Section 5, Performance analysis and other related issues are discussed. In Section 6, conclusion is presented.

1. Mobile Agents and their Challenges

Mobile agents are computational software processes capable of roaming wide area networks (WANs) such as the WWW, interacting with foreign hosts, gathering information on behalf of its owner and coming 'back home' having performed the duties set by its user. These duties may range from a flight reservation to managing a telecommunications network. However, mobility is neither a necessary nor sufficient condition for agent-hood.

Challenges :

- Transportation
- Authentication
- Low Band-width or Busy Network
- Dynamically Re-plan Migration path to adapt circumstance

2. Friend Network

The Static Network is inadequate to support Mobile Networking. We require the combined function of Mobile Host (MH), Base Station (BS), and Mobility Support Router (MSR) to be a part of Mobile Networking. Friend Network forms a

"Ad-hoc Network" whenever required. It does not require any fixed infrastructure to work. It provides greatest possible Flexibility. It forms Instant Infrastructure for Disaster Relief and Remote area Communication.

A network is considered as Friends Network of mobile host if the following conditions occur :

- ✍ Substantial network traffic exists between the network and the mobile host
- ✍ The network is trusted network of the mobile host and equipped with Mobility support router.
- ✍ Prohibiting the routing information into unfriendly networks can also Provide the network security.

3. Friend Agent

The introduced concept "FRIEND AGENT" has the advantages of friend network and overcomes the challenges facing the mobile agents.

It is the combination of Friend network concept and the Mobile agent Technology.

Overcoming the barriers of Mobile agents : Friend Agent provides the solutions for overcoming the challenges faced by the mobile agents. The challenges mentioned above can be solved by providing an optimal solution through the Friend Agent.

Network-Bandwidth Problem : Friend agents solve the network bandwidth problem. Network bandwidth in a distributed application is a valuable (and sometimes scarce) resource. A transaction or query between a client and the server may require many round trips over the wire to complete. Each trip creates network traffic and consumes bandwidth. In a system with many clients and/or many transactions, the total bandwidth requirements may exceed available bandwidth, resulting in poor performance for the application as a whole. By creating an agent to handle the query or transaction, and sending the agent from the client to the server, network bandwidth consumption is reduced. So instead of intermediate results and information passing over the wire, only the agent need be sent.

The Network Bandwidth used for routing information update can be greatly reduced by achieving Optimal Traffic Routing. Information should be propagated

only to those routers or hosts who need to communicate with the mobile host, instead of Broadcasting. This assumption is quite normal in mobile traffic scenarios.

Dynamic Re-plan Migration :

Case (i) : When the encapsulated packet destined for mobile host arrives to mobility support router, it is de-capsulated and finds whether the mobile host is available within the Home Location Register (HLR). If yes, it is delivered to mobile host via its own mechanism, through Base station.

Case (ii) : If mobile host is out of HLR, and MSR updates the information of the location of the MH. In the mean time, the MH on reaching the other location registers in the corresponding Visiting location Register (VLR), which is updated in the HLR, the packet is re-encapsulated and routed through the MSR and VLR for serving the MH. In this way, the Migration is re-planned dynamically.

4. Performance Analysis of Friend Agent

Many factors needed to be considered in Network routing, and the information retrieval. InAd-hoc networks each node must be able to forward data for other nodes. The routing table must somehow reflect these changes in topology have to be adapted. It takes very minimum time for the routing table to be updated. The topology changes during distribution, LIR- (Least Interference Routing) can be used to avoid the wastage of Bandwidth, and while Topology changes occur during distribution.

Comparison between Traditional Information Retrieval and Friend Agent Information Retrieval

Traditional Information Retrieval :

Transfer Information : 100KB*1S/KB= 100S

Transfer Back Results : 1KB*1S=1S

Total Time spent in communicating: 100S+1S=101S

Friend Agent Information Retrieval :

Transfer Agent : 1KB* 1S =1S

Carry Information Locally:100KB*0.1S/KB=10S

Transfer Back result : 1S

Total Time spent in communicating: 1S+10S+1S=12S

As a result of the Performance Analysis, Time for retrieving information is reduced approximately EIGHT times by using Friend Agents in comparison to Traditional Retrieval Systems.

5. Conclusion

We have proposed a novel concept called “Friend Agent” which overcomes the challenges in mobile agents using the advantages of Friend Network, and it provides Secured and Faster Information Retrieval .It offers the advantage reduced Network resource consumption by providing Optimal routing and authentication only to Friend Networks thus paves a way for the Network Security. In addition to this, it also offers greatly reduced time for Data transfer when compared to traditional Information Retrieval Systems. This Concept can be deployed to a wide range of Information Retrieval Applications.

6. References

1. An Adaptive Routing Scheme for wireless Mobile Computing – Ruixi Yuan
Wireless and Mobile Communications by Allied Publishers in Association with Kluwer Academic Publishers- Jack M.Holtznan David J.Godman Rugers University.
2. Mobile Communication – Jochen Schiller
3. Computer Networks - Andrew S.Tanenbaum

4. http://www.comsoc.org/livepubs/surveys/public/4q98issue/bies.html
5. http://www.computer.org/concurrency/pd1999/pdf/p3080.pdf
6. http://www.javaworld.com/javaworld/jw-06-1998/jw-06-howto.html
7. http://more.btexact.com/projects/agents/publish/papers/review3.html
8. www.csc.ncsu.edu/faculty/ mpsingh/papers/positions/maas-97.pdf
9. www.objs.com/agent/00-12-05.ppt
10. www-eksi.cs.umass.edu

About Author



R. Bhaskaran, is studying in M.E., (CSE) Final Year, Raja College of Engineering and Technology, Veerapanjan, Madurai, Tamil Nadu.
E-mail : bhaskaran_mdu@yahoo.com, bhaskaran_1981@rediffmail.com

Preservation and Maintenance of the Digital Library : A New Challenge

K R Mulla Shivakumara A S M Chandrashekara

Abstract

Libraries, archives, and museums play a critical role in organizing, preserving, and providing access to the cultural and historical resources of society. Digital technologies are used increasingly for information production, distribution, and storage. The institutions that have traditionally assumed responsibility for preserving information face technical, organizational, resource, and legal challenges in taking on the preservation of digital holdings. Maintenance will be critical to digital libraries; especially those who promote broad access to diverse, informal materials. If ignored, maintenance issues within the digital library, especially those relating to its materials will threaten its usefulness and even its long-term viability. We perceive the maintenance problem to be both technical and institutional, and this paper considers the preservation and maintenance of the digital library. The paper examines collection maintenance from several vantage points, including software architecture and the type of collection. The paper ends with an examination of potential technical solutions.

Keywords : Digital Library, Preservation.

0. Introduction

As with any new technology-based idea, there has been considerable controversy over the definition and possibilities of the term “digital library” to the computer science community, the new technical possibilities. However, as traditionalists in the library community might point out, important issues are being ignored. This paper promotes a view of collections and the long-term consequences of their operation, based on the consideration of digital libraries as social institutions.

This runs contrary to the substantial body of digital library research that focuses on creating the initial Preservation, collections and providing access mechanisms. We believe that the problems must be recast to include long-term issues. By centralizing those issues surrounding the maintenance of institutions and their artifacts, especially the library collection, important considerations for the long-term success of digital libraries emerge. To distinguish our concerns from traditional collection management, we call these materially increasing accessibility and content issues, over the long run, collection maintenance. We use “maintenance” to deliberately invoke “software maintenance” and its often ignored importance for software systems. As discussed below, collection maintenance is likely to be a significant problem in the digital library.

This paper begins by discussing the differing notions of the digital library, anchoring the issues in an analysis of institutional needs and practices. We then examine the various types of collections, including those that include dynamic and informal materials. This consideration of collection types and their control lends itself to analyzing the institutional arrangements and resulting maintenance issues for digital libraries. Maintaining collections, that are extensions of traditional collections (with delineated boundaries), not surprisingly require only extensions of traditional methods. Existing institutional arrangements and resources can be modified to handle these requirements, and maintaining collections that include dynamic and informal information will be possible only with new technical solutions.

1. Digital Library Collections and Preservation

Libraries are about many things. But, collections have always been at the heart of libraries, whether digital, traditional brick and mortar, or hybrid. Collections will retain that role in the future. However, the concept of what constitutes a collection in the networked environment of digital libraries is undergoing a transformation from the age-old concept of library collection signified by ownership. New concept of a digital collection is evolving incorporating adaptations of many old features and standards, and creation of many brands. This conceptual and pragmatic evolution is far from over. What are the digital library collections ? The question looms as a large problem for practice and for research and development.

The concepts and processes of collection development and collection management are undergoing a transformation. New processes and tools for collection development have emerged which are used for development and management of both, traditional and digital collections. The process of collection management became more closely connected than ever before with means, ways, and policies for access, adding an additional dimension. But digital collections also present distinct and serious challenges related to preservation and archiving. Many libraries and other institutions worldwide are concerned with these issues. Libraries started including digital preservation as a vital part of collection management. A number of national and international bodies are developing standards, tools, and practices related to preservation.

a) Collection Types and Control

Libraries have always managed their collections, selecting and removing items from their shelves. This has been viewed as a critical function of library management. According to the collection management literature, the practices of collection management are dependent on the type of library collection. We will argue here that new types of collections in the digital library will lend themselves to new types of maintenance issues. Table 1 delineates four types of digital collections.

This is not the only method of distinguishing among “digital libraries”; for example, we could have included access methods or network topologies. Additionally, an actual digital library could have elements of any or all of these types. In a traditional, or paper-based, library, there is considerable control over the collection. Library staff can decide what is and what is not in the collection. Maintenance of the collection is within the preview of the institutional members.

b) Collection Maintenance

The variation in collection control determines the type of institutional and technical maintenance possible. Collections that are closer to those in traditional libraries can use more traditional control and maintenance mechanisms. Digital libraries that incorporate more individualistic, dynamic, and informal information may need to find new maintenance mechanisms.

Traditional libraries, have developed methods for maintaining their core set of institutional ideals, their community of practice, and their collections of materials. They were based on a constrained collection; i.e., a selection from the bibliographic universe. A traditional library could never cope with much ephemera; it would require too many resources. Going digital, however, changes the cost structure, and collection and maintenance costs need to be revisited. As Figure 2 shows, however, the dynamism and volatility of organizational memories tend to be close to that of traditional libraries. It will be interesting to see how existing maintenance procedures within organizations will adapt.

c) Preservation of Documents

Distributed bin packing problems and the file allocation problems are known to be NP-hard. This is one reason we have not sought to find an optimal placement for data collections. Moreover, these problems are even harder when the number of sites and sizes of collections are not known in advance. In our model, reliability is more important than performance, configuration like mirrored disks are preferred if possible. However, due to the dynamic nature of the system, with new sites and new collections appearing at any time, it is not possible to statically assign copies to mirrored disk pairs, and a more dynamic data allocation scheme must be devised. For this reason, much work has been done to ensure correctness and consistency for distributed transactions.

Data is replicated so that it can be read despite temporary site failures or network partitions. This is a different goal than long-term reliability, which seeks to preserve data despite permanent site failures or data corruption. Digital library researchers have begun to examine the archiving problem. Some projects have focused on maintaining collection metadata, or on dealing with the data formats. Each of these issues is important, and complements the basic bit-level reliability we seek to provide here. Our trading algorithm could be used in systems like these to place data in the most reliable manner.

2. Key Components

Digital Library Repository (DLR) is formed by a collection of independent but collaborating sites. Each site manages a collection of digital objects and provides services (to be defined) to other sites. Each site uses one or more computers, and can run different software, as long as it follows certain simple conventions that we describe in this paper. Our architecture is based on following key components.

i) No Deletions

The deletion is dangerous when sites are managed independently; in particular, it makes it hard to distinguish between a deleted object and one that was corrupted (Morphed into another) and needs to be restored. Ruling out deletions is natural in a digital library, where it is important to keep a historical record. Thus, books are not burnt but removed from circulation.

ii) Data preservation overview

Archiving sites are autonomous units, managed by different organizations, and thus are not under any centralized control. Each site has a quantity of archival storage. New archival storage may be added to a site at any time. The basic unit of archived data is the data collection. A collection represents a set of related data, for example a group of files, a database table, a set of documents etc. Collections may contain any number of data items, and different collections may be of different sizes. Clients can also read archived data at their own site or at a site storing a copy of the collection. We also need a method for determining reliability. The goal of an archiving system is to reliably protect data, despite the potential for site failures.

Thus, we use the following concepts:

a) Site reliability : The probability that a site will not fail.

b) Local data reliability : The probability that the collections owned by a particular site will not be lost.

c) Global data reliability : The probability that no collection owned by any site will be lost.

iii) Digital Preservation Responsibilities

Digital preservation is not an isolated problem affecting only large libraries and archives. Some digital materials exist in holdings of other libraries for which the institution assumes preservation responsibility. In fact, the research libraries, which tend to be larger than the archives, museums, and special libraries, are not quite as likely to hold digital materials as the other types of institutions.

iv) Digital Preservation Policies and Practices

Digital preservation policies and practices are not well developed in academic institutions. One common reason that institutions appear not to develop digital preservation policies is that they have not yet assumed responsibility for preserving materials in digital form. However, taking responsibility for digital preservation does not necessarily mean that institutions use policies to govern their digital preservation activities. Only half of the institutions with digital materials in their holdings have written digital preservation policies.

v) Preservation Practices

Effective digital preservation requires life-cycle management of digital information from the point of creation through storage, migration, and providing access on a continuing basis. Few institutions have established methods in place for digital preservation. Institutions limit the acceptable formats to flat files while others accept several different formats such as PDF, TIFF, SGML, and word processing formats.

vi) Problems and Threats to Digital Preservation

The technological obsolescence is the greatest threat to loss of digital materials, followed closely by insufficient resources and an inefficient policy or plan for digital preservation. The lack of resources for digital preservation is the greatest threat in institutions that have policies in place for digital preservation.

vii) The problem of digital changes and user expectations

Digital technology and high speed networks are leading to sweeping changes throughout society, and moving image production and distribution are in no way immune to either the technological changes or to the social expectations that these changes have induced. In the past completed digital effects were transferred back onto film and inter cut with the rest of a production, but as general moving image production itself becomes increasingly digital, this intermediary transfer to film will become far common. Small-budget independent productions are increasingly being shot and edited in digital form. According to director Mike Figgis, “there is clearly a technical revolution taking place you can edit a film on a laptop, and there is the Internet, the streaming and downloading capabilities. These are the technical elements of the revolution” (Silverman 2000).

viii) Problems with preserving anything Digital

Information encoded and stored in digital form is fragile, but quite different from film stock. Digital storage shares some characteristics with video storage but it is different from analog storage formats i.e. film and video. But print archivists and special collection librarians, who aggressively pursue print-based collection development in their particular specialty areas, claim that it should be the responsibility of computing staff of their organisation to pursue collection development of material originating in digital form. The costs for handling digital materials diminish and as strategies for long-term maintenance of

digital files become better known, reasons for handling digital material separately will start to fade, and administrators will begin to realize that digital files of moving images have much more in common with film and video than with word-processing files and databases.

ix) General Approaches to Digital Preservation

This is a brief history of the approaches to preservation of simple types of digital materials. In the mid-1990s, the library community began to worry about the fragility of works stored in digital form. The Commission on Preservation and Access and the Research Libraries Group formed a task force to explore how significant this problem was really. The Task Force report sounded an alarm "Rapid changes in the means of recording information, in the formats for storage, and in the technologies for use threaten to render the life of information in the digital age as, to borrow a phrase from (Task Force 1996)".

The wide body of digital works in Task Force 1996 involves periodically moving a file from one physical storage medium to another to avoid the physical decay or the obsolescence of that medium. Two key approaches have been proposed to deal with the problem of changing file formats (Task Force 1996): migration and emulation.

3. Conclusion

This paper examined some of the mechanisms required to maintain the digital library. Collection and maintenance is a significant issue in the digital library more than in traditional libraries or current organizational memory repositories. The broadly construed and narrowly construed digital libraries were examined in this paper. The narrowly construed library is analogous to the traditional library where the collection has known boundaries. Because of the possibility of control over the collection in the narrowly construed library, many of the institutional mechanisms for maintaining collections can be assimilated from the traditional library.

The inclusion of dynamic and informal materials in the collection, leads to serious control and long-run maintenance issues. Because of the lack of control over the collections, technical mechanisms will be needed for collection maintenance. While we need several technical possibilities for collection maintenance, we perceive this problem to be both technical and institutional. However, a strictly technical emphasis will not lead to an adequate understanding of the long-run issues in digital library use. The digital library is more than a set of technologies; it is also a social institution with long-term needs and maintenance requirements.

By combining their vast set of skills in handling of analog objects as well as moving to new paradigms provoked by the digital age, moving image archivists can continue to play a critical role in preserving our cultural heritage and ensuring that today's works will last well beyond the life of the team that produces them.

4. References

1. Benjamin, Walter (1978). "The Work of Art in the Age of Mechanical Reproduction." Illuminations. New York: Schocken.

2 Besser, Howard (forthcoming). Longevity of Electronic Art, Proceedings of the International Conference on Hypermedia and Informatics in Museums (September 2001).

3. Besser, Howard (2000a) Digital Longevity, in Maxine Sitts (ed.) Handbook for Digital Projects: A Management Tool for Preservation and Access, Andover, MA: Northeast Document Conservation Center, pages 155-166.

4. Besser, Howard (2000b). Bibliography of Moving Image Indexing (website) (http://www.gseis.ucla.edu/~howard/Classes/287-mov-index-bib.html)

5. Besser, Howard (1997). The Changing Role of Photographic Collections with the Advent of Digitization, in Katherine Jones-Garmil (ed.), The Wired Museum, Washington: American Association of Museums, pages 115-127.

6. Besser, Howard (1994). Fast Forward: The Future of Moving Image Collections, in Gary Handman (ed.), Video Collection Management and Development: A Multi-type Library Perspective, and Westport, CT: Greenwood, pages 411-426

7. Besser, Howard (1987). Digital Images for Museums, Museum Studies Journal 3 (1), Fall/Winter, pages 74-81.

8. Council on Library and Information Resources (2000). Authenticity in a Digital Environment, Washington, D.C.: Council on Library and Information Resources May.

9. Davis, Ben (2000). Digital Storytelling, Razorfish Reports #24, June 16 (http://reports.razorfish.com/frame.html?rr024_film)

10. Digital Light Processing (2001). Where is DLP Cinema? (Website) (http://www.dlp.com/dlp/cinema/where.asp)

11. Goodrum, Abby (1998). Representing moving images: Implications for developers of digital video collections, Proceedings of the 1998 Meeting of the American Society for Information Science.

12. Hummelen, Ijsbrand and Dionne Sille eds. (1999). Modern art: who cares? : An interdisciplinary research project and international symposium on the conservation of modern and contemporary art, Amsterdam: Foundation for the Conservation of Modern Art and Netherlands Institute for Cultural Heritage.

13. Laguna Research Partners (2000). Notes from INFOCOMM 2000 (Industry Brief), June 22 (http://www.lrponline.net/infocomm2000_1.PDF)

14. Laurenson, Pip (2000) Between Cinema and a Hard Place: The Conservation and Documentation of a Video Installation by Gary Hill, Talk to Preservation of Electronic Media session, 28th Annual Meeting of American Institute for Conservation of Historic & Artistic Works, Philadelphia, June 9, 2000.

15. Lyman, Peter and Howard Besser (1998). Defining the Problem of Our Vanishing Memory: Background, Current Status, Models for Resolution in Margaret MacLean and Ben H. Davis (eds.), Time and Bits: Managing Digital Continuity, Los Angeles: Getty Information Institute and Getty Conservation Institute, pages 11-20.

16. MacLean, Margaret and Ben H. Davis (eds.) (1998), Time and Bits: Managing Digital Continuity, Los Angeles: Getty Information Institute and Getty Conservation Institute.

17. Mallinson, John C. (1986). Preserving machine-readable archival records for the millenia, Archivaria 22 (summer), pages 147-152.

18. National Institute for Standards and Technology (2001). Digital Cinema 2001 Conference, Gaithersburg, MD, January 11-12 (http://digitalcinema.nist.gov/)

19. Rothenberg, Jeff (2000). An Experiment in Using Emulation to preserve Digital Publications, Den Haag: Koninklijke Bibliotheek, 2000. - (NEDLIB Report series; 1) (http://www.kb.nl/coop/nedlib/results/emulationpreservationreport.pdf).

20. Rothenberg, Jeff (1999). Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation, Washington, D.C.: Council on Library and Information Resources, January (http://www.clir.org/pubs/abstract/pub77.html)

21. Rothenberg, Jeff (1995). Ensuring the Longevity of Digital Documents, Scientific American, 272(1): pages 42-7.

22. Sanders, Terry (1997). Into the Future: Preservation of Information in the Electronic Age, Santa Monica: American Film Foundation (16 mm film, 60 minutes)

23. Silverman, Jason (2000). Digital Cinema Plays with Form, Wired News, April 12 (http://www.wired.com/news/culture/0,1284,35098,00.html)

24. Task Force on Archiving of Digital Information (1996). Preserving Digital Information, Commission on Preservation and Access and Research Libraries Group (http://www.rlg.org/ArchTF/tfadi.index.htm)

25. Turner, James M. (1999). Metadata for moving image documents. Proceedings of the 20th National Online Meeting, New York, May 18-20, 1999, edited by Martha E. Williams. Medford, NJ: Information Today, 477-486.

Annexure - I

1).

		DIGITAL LIBRARY		
	Traditional library	Organizational repository (organizational memory)	World wide Web	Usenet or other computer-mediated communication system
Types of collection	Monographs, Serials, Special collections	Documents and files	Text and multimedia notes	Messages
Authoring agencies	Authors and their publishers	Organizational members or sanctioned individuals	Anyone	Anyone
Collection control	Selection by organizational member (e.g. bibliographer)	Selection by organizational member. May have requirements of other approval	None (or individual)	None (or moderator)

Table 1: Characteristics of digital libraries (by type)

2).

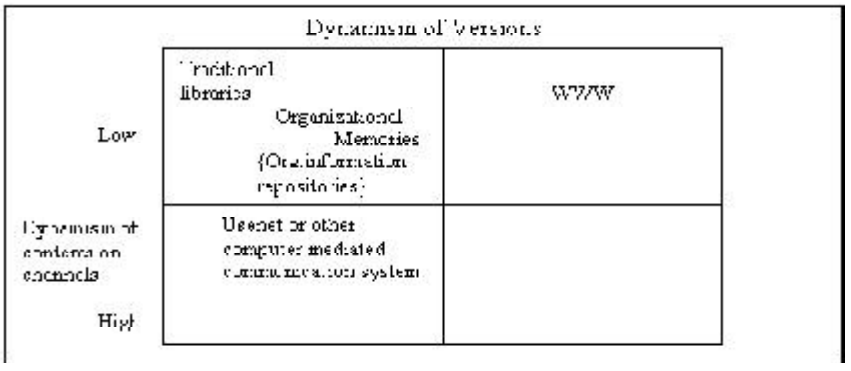


Figure 2. Dynamism (volatility) in types of digital library collections

About Authors

K R Mulla is working in HKBK College of Engineering, Nagawara, Bangalore.
E-mail : mulla_kamalasab@yahoo.com

Shivakumara A S is working inHKBK College of Engineering, Nagawara, Bangalore.
E-mail : shivaksi@yahoo.com

M Chandrashakara is working as Lecturer in Department of Library Information Science., University of Mysore, Mysore.
E-mail : chandram5@yahoo.com

Meta Search in Distributed Electronic Resources : A Study

M Krishnamurthy

Abstract

The explosion of information on the Internet and information technology in general has created challenges for libraries to focus on developing more effective ways to meet the information needs of users. One practical approach is through customized portals performing simultaneous database searching. This paper presents a new library portal service used at the Indian Statistical Institute Library at Bangalore (ISIB). The key feature of a library portal is to allow searching across multiple databases without having repeat search. This feature is generally referred to as meta-search, parallel search, broadcast search or federated search. Discussion also includes strategies of local customization and the impact on library management in an electronic environment.

Keywords : Electronic Resources, Portal.

0. Introduction

The explosion of information on the Internet and information technology in general has created challenges and more opportunities to information professionals for redefining their roles for the present and future. Web based information services are receiving much attention from the library community. With the advent of Internet and more specifically the World Wide Web, libraries have undergone a revolution in the way that they build information source collection, operate and provide information services to users. The information professionals are uniquely positioned to play a major role in the emerging information age driven by convergent technologies.

1. Library Portals

While information is readily available, overwhelming at times, on the World Wide Web, libraries are focusing more into developing more effective alternative ways for users to efficiently navigate among the library resources distributed on library's Websites. One of the approaches to accomplish this goal is through customized portals

performing simultaneous database searching. As Strauss has noted that a portal as a special kind of gateway to Web resources" a hub from which users can locate all the web content they commonly need" (Strauss 2003). The design and implementation of local portal will be the responsibility of individual institutions. Local portals will allow tailored access to a selection of data sets of importance to a particular institution, plus integration with other locally licensed data sets and local products, abstracts and indexing and citation databases, which user has access collection of electronic journals including OPAC (Webb, 1998),

The content, structure, and relationship of information in the information space are most commonly represented as alphabetically based or hierarchical categories. Because of information overload on the Web, hypertext-browsing abilities in locating information may be significantly augmented by content based searching (Chen, *et al*, 1998)

The key feature of a library portal is to allow searching across multiple databases without having repeat search. This feature is generally referred to as Meta search, parallel search, broadcast search or federated search.

2. Model of Access at the Bangalore Center Library

2.1 ISI Bangalore Library

The Indian Statistical Institute (ISI) has three centers at Calcutta, Bangalore and New Delhi. The Library at the Bangalore Centre (ISIB) established in 1978, has a good collection on Statistics, Mathematics. Quality Engineering/ Management and Library and Information Science it) India. Presently it subscribes to about 300 international journals for its 300 library users including faculty, students and research scholars. The overall collection of books and reference materials, reports, reprints, directories, encyclopedias, CDs etc., is approximately 50,000.

The ISIB Library provides from its homepage Web listings to its journal resources, separately linked according to procurement agreement with content providers (Fig. I). They are, namely, ISIB Online Subscriptions, ISIB Periodical Holdings; *Science Direct* consortium, Union Catalogue of Publications in Bangalore Libraries.

2.2 Web Access to ISIB E-Journal Subscriptions

The first listing is the -E-journals available for ISI Bangalore." The ISIB Library has a total of 278 online journal subscriptions since 2000. This listing, written in html format, provides access to ISIB licensed electronic resources either to the full-text articles or to the abstract of the articles. Linking data to these electronic resources are embedded in this hypertext file. They include such data as the details of the URL path and sometimes the session information needed to establish authorization at the vendor/publisher server. Through this dynamic linking, access to the Library services is enhanced with immediacy, accuracy and currency of information, which are the most important factors in fulfilling user information needs.

2.3 Web Access to ISIB Periodicals Holdings

The second listing is the "Periodicals holdings of ISIBC Library-2001." Tile ISIB periodical collection has a total of 412 records of holdings since 2001. This listing, also written in html format, provides access to the ISIB journals alphabetically by the first letter of the title word. It includes publisher information and the beginning date of the periodical holding, and bibliographic notes such as the former and/or later titles. This Web list is continually updated in the source code by library staff to incorporate new titles when they are added to the library.

2.4 Web Access to Science Direct Consortium

The third listing is the "Science Direct Consortium: Online Access to Elsevier Science and Academic Press Journals." In 2003, a contract to use Elsevier Science Direct electronic publications was signed by the Indian Statistical Institute to establish a consortium networking the three IST Centre libraries. The consortium was formed to contribute to the development of research through the acquisition of electronic publication to the participating libraries of Indian Statistical Institute using the LAN, in addition to their individual current journal collection. In this consortia gateway, the ISIB Library maintains a total of 133 online journal subscriptions. They consist of consortium subscription to e-Journals from various Elsevier publishing groups such as Pergamon, Saunders and Academic Press.

2.5 Web Access to Union Catalog of Publications in Bangalore Libraries

The fourth listing is the "Union Catalog of publications in Bangalore Libraries." In 1997, a project to promote access to journals available in Bangalore was launched to create a union catalog of current journals subscribed by the major libraries in the region. The result is a formation of a union catalog of current journals compiled by the National Centre for Science Information (NCSI). The catalog consists of five separate annual listings from 1997 through 2001. It provides bibliographic information as well as the holding library information. It also provides a subject search to the journal titles. However, each annual listing is searched independently and the title word searching is limited by adjacent search only.

3. Analysis of Need

The cooperative environment of library resources has created an environment that the ISIB Library can serve with a host of journal resources available in the region. The development of the Library's homepage reflects the active participation of each component of the Web listings. However, users often need to navigate among and between the various Web pages to locate the journal holding or full-text articles in question. Since the search is mainly conducted as a browsing operation, knowledge of title words is required as key word or truncated searches are not possible. There are also issues of high maintenance for the upkeep of these hypertext-markup lists,

Perhaps the biggest difficulty is the lack of normalized data across the various Web listings. For example, forms of abbreviation were not standardized for name of the journal title, publisher, place of publication, or library code of ownership. As a result, the search, regardless of how thoroughly conducted, would be futile and meaningless. As electronic and print resources are increasing to overlap in various journal collections, the best practice for providing an efficient and accurate access is inevitably through a library portal that is customized to navigate simultaneously across tile various lists.

4. Technologies and Strategies for Portal Customization

4.1 Normalization of Metadata in a SQL database

The Web pages at the ISIB homepage were created to display individual listings of journal resources. In order to achieve normalization of data, a MS Access database was created as the main resource registry. Four tables were then constructed sharing many identical metadata field names and information, such as Title, ISSN, Place of Publication, URL, Subject, etc.

Data from the Web lists at the original ISIB site were downloaded as sources to populate the corresponding registry. For the union catalogs of current serials from 1997 through 2001, the five years data were merged into one registry. A Visual Basic application was composed to eliminate duplicates and consolidate the holdings for multiple years.

Spelling and use of abbreviations were standardized through the registry. This is a crucial feature for producing accurate and meaningful results for the end users.

4.2 Programming for Simultaneous Searching

As searches are performed in multi-user situations, the best practice is to install the metadata registry over a MS SQL server, which allows better transaction throughput over the Internet (Mischo and Schlembach, 1999).

Unlike the Library's original Web access to the library resources, the new Web applications are typically composed of several Active Servers Pages (ASP). The embedded VB scripting commands process and parse the user search strings into appropriate SQL commands to retrieve data across the four tables defined in the metadata registry.

The new information portal provides a gateway to the entire spectrum of library journal resources. However, it also allows end-users to choose among them as desired for more efficient search strategy. Fig. 2 shows the newly designed ISIB Library search aid as a gateway to its journal resources.

The entered search term, computes and retrieves a number of matches simultaneously from a pre-selected list of information sources, i.e., in this case all are selected: Journals at ISIB Library, *Science Direct* Journals at ISIB Library, Periodicals Holdings at ISIB Library, and Union Catalog of Bangalore Libraries. Users may proceed to choose from here with one single click to locate more details about the journals of interest.

5. Evaluation of the New Library Portal Service

With the advent of digital libraries and of wide area network, enormous amount of textual information is made available all over the world. Searching and browsing are the two resource discovery paradigms mostly used to access this information. To improve access to the information stored in our library portal we incorporated a library search aid.

5.1 Efficiency of Searches

The advantage of simultaneous searching across multiple sources of library information is evident. For users, the new library portal design presents a gateway to navigate fruitfully from a simple step rather than making a series of search attempts indiscriminately across the previous Web lists. The biggest improvement of this approach is exemplified particularly in searching the 5-year union catalogs into one database, as it eliminates the need to make multiple searches one at a time as required previously.

With the new approach, collection scope is well defined for users to make intelligent decision for making their selection. Resource sharing is better served as interlibrary loan staff can easily determine from where their request is best to be filled.

5.2 Maintaining and Updating Issues

One of the major concerns in providing library resources is the maintenance issue. Information is only as good as it is up-to-date. Consequently, the foul Web listings on the ISIB Library homepage present several challenges in maintaining the information current. For example, new title information must be updated in hypertext language. It is also quite cumbersome to make changes, such as library holdings, to the existing files.

In contrast, the new library portal performs a uniform SOL searching across a database with uniformly defined metadata tables. Each table has identical field structure so that information can be kept in standardized format. Normalization of title words, library codes and abbreviations improves the accuracy of search matches as well.

6. Conclusion

The basic concept of the library portal is to provide databases, and localization. It is therefore necessary for the library to organize the information in order to reflect institutional or consortia information resources licensed by and made available to the users. The efforts to enhance access involve both public as well technical services in order to maintain a successful performance. We feel that good planning should be low maintenance but highly efficient one. Our solution to present a library portal service achieves this goal with ASP technologies, a registry accessible over a SOL server, and a uniform metadata structure of tables to produce effective simultaneous searching of matched results.

7. References

1. Schatz, B.R., et al. Internet browsing and searching. User evaluations of category map and concept space techniques, Journal of the American Society for Information Science. Vol.49; 1998; p582-603,
2. Mischo, W.H. Library portals, simultaneous search, and full-text linking technologies. Science and Technology Libraries. Vol.20; 2001; p133-147.
3. Mischo, W.H.; Schlembach, M.C. Web-based access to locally developed databases. Library Computing. Vol. 18: 1999; p51-58.
4. Strauss, H. Web portals: the Future of Information Access and Distribution, The Serials Librarian. Vol.44: 2003; p27-35.
5. Webb, J.P. Managing licensed networked Electronic resources in a university library. Information Technology Libraries. Vol.17; 1998; p198-206.

About Author

Dr. M Krishnamurthy is working as Librarian in Indian Statistical Institute, Bangalore.
E-mail : krish@isibang.ac.in, mk13murthy@hotmail.com

Web Based Library Services : A Study on Research Libraries in Karnataka

Vijayakumar M B U Kannappanavar Madhu K N

Abstract

The paper presents how research libraries are effectively using web technology to provide their services. It is observed from the study that all the research libraries under the study are following Hybrid culture (manual & electronic) in providing the library and Information services to their clients. The results also conclude that, majority of the aerospace research libraries effectively use web technology to provide library services when compared to biological and health science libraries. But all selected academic libraries provide web based library services when compared to aerospace and Biological & Health Science libraries.

Keywords : Library Services, Research Library.

0. Introduction

Web technology has become pervasive in all walks of our life, be it research, business, entertainment or others. In other words, web technology is talk of the day and need of the hour in order to remain competitive. It is changing the way people think, work or run a business. On one hand web technology is changing the way in which the activities of work place are being carried out and on the other hand there are challenges like poor infrastructure, lack of knowledge and information on effective use of web Technology. It becomes evident in such a situation to go for reassessment and reevaluation of the entire activities of work place where web technology is used to provide the library services.

1. Review of Literature

Literature search is an essential link in the process in research. It helps to know what the other researchers in specified subject have done. So an attempt has been made here to identify the related literature published in the area of study. The information sources consulted for the review includes, books, journals, articles, library and information science abstract [LISA], International dissertation abstracts, conference/ seminar papers and other resources.

2. Need for the Study

This is a crucial moment to bring up successful case studies in Web Technology application in libraries as well as raising awareness about the use of Web Technology in Research libraries specifically information disseminating. Also it is apt time to carry out case studies to examine existing application of Web technology in research libraries in a region, particularly in developing country.

3. Objectives of the Study

This study is carried out to know, how research libraries are effectively using the web technology for providing their library services and

1. Current scenario of web technology in research libraries.
2. To know the different library services provided through web technology

4. Research Methodology

The study is designed, developed and carried out to determine and analyze the existing position and use of Web technology for providing library services. The principal tool for data collection covering the research libraries spread across the Karnataka state is "Questionnaire". We grouped National Aerospace Laboratory (NAL) library, Hindustan Aeronautical Limited (HAL) library, Aeronautical Development Agency (ADA) library and Centre for Air Borne System (CABS) library in to Aerospace libraries. Central Food Technological Research Institute (CFTRI) library, National Institute of Mental Health and Neuro Sciences (NIMHANS) library, Kidwai Memorial Institute of Oncology (KMIO) library and National Center for Biological Sciences (NCBS) library into Biological and Health Science Libraries and Indian Institute of Science (IISc) library, Indian Institute of Management Bangalore (IIMB) library, Indian Institute of Astrophysics (IIA) library and Raman Research Institute (RRI) library into Academic Libraries for our convenience.

5. Analysis of the Data

The data collected was analyzed in terms of various aspects related to the study using Statistical Package for Social Sciences [SPSS] software. The data was interpreted in terms of objectives defined.

5.1 Library Services

A library service is "acid test" for effective library functioning. Library services are means of sharing knowledge to wide spectrum of user community. Librarianship is a media of service in a library for users to meet their needs like research, educational, intellectual and others. Effective library services make functional library qualitative. Library, even with a small collection, can provide effective library services provided it effectively uses available information technology applications.

5.2 Level of Library Services Provided

To identify the level of library and information services provided at institutional level, local/regional level, national and international level, a questionnaire was distributed to the respondents in this regard. The collective responses are depicted in the Table 1 for necessary statistical interpretation.

Table 1: Level of Library Services Provided

Libraries	Parent Institution		Local level		Region level		National level		International level	
	No.	%	No.	%	No.	%	No	%	No.	%
Aerospace Libraries	4	100	2	50	3	75	2	50	1	25
Biological and Health science Libraries	4	100	2	50	3	75	3	75	2	50
Academic Libraries	4	100	3	75	3	75	3	75	1	25
Total	12	100	7	58.33	9	75	8	66.67	4	33.33

Table - 1

Table 1 also points out that all research libraries provide their service to users of its parent institution. While 9 (75%) research libraries are providing their service at regional level, 8 (66.67%) at national level, 7 (53.33%) at Local level, only 4 (33.33%) research libraries provide their services at international level.

From the table 1 it is clear that, libraries are the heart of the Parent Institution and work towards the achievements of the parent organizational goal, as all libraries provide their services to parent Institution.

5.3 Way of Providing Library Services

To identify weather research libraries provide library services through manual mode, automated mode, LAN of library or through its homepage, respondents were asked to describe, how were they providing library services. The responses elicited from them are depicted in Table 2 for necessary statistical analysis.

Table 2: Media of Library service

Libraries	Manual		Automated Library		LAN of Library		Web Page of Library	
	No.	%	No.	%	No.	%	No.	%
Aerospace Libraries	4	100	4	100	4	100	3	75
Biological & Health science Libraries	4	100	3	75	3	75	2	50
Academic Libraries	4	100	4	100	2	50	4	100
Total	12	100	11	91.67	9	75	9	75

Table – 2

Table 2 clearly indicates that all research libraries provide library services manually. Out of 12 selected research libraries, 11(91.67%) research libraries provide the library services through automated library followed by 9 (75%) libraries who provide services through LAN and Home page of the library.

Further it is very interesting to note here that, all selected Academic libraries are providing web based library services and aerospace libraries provide automated and LAN based library services

5.4 Library Services through Web Technology

A good and well designed library web page is not only excellent media for publicizing the library functions, activities, programmes, resources and services but also helps to bring to the notice of users all significant information which they must know in developing and using their library [3]. To assess the degree of web technology in the library services, the survey included a questionnaire about the web-based library services offered by each library. The collective responses are depicted in the Table 3 for further statistical interpretation.

Table 3 : Web based Library services

Services	Aerospace Libraries		Biologica & Health Science Libraries		Academic Libraries		Total	
	No	%	No.	%	No.	%	Nc	%
Literature Search	2	50	2	50	2	50	6	50
Indexing	2	50	1	25	1	25	4	33.33
Reference	2	50	1	25	1	25	4	33.33
Referral	1	25	1	25	2	50	4	33.33
News Clipping	2	50	0	0	1	25	3	25
Bibliographic	2	50	1	25	1	25	4	33.33
Translation	0	0	0	0	2	50	2	16.67
Abstract	2	50	0	0	2	50	4	33.33
SDI	1	25	1	25	1	25	3	25
ILL	1	25	1	25	2	50	4	33.33
Technical Enquiry	2	50	1	25	0	0	3	25
Publication	2	50	1	25	0	0	3	25
New arrivals	2	50	1	25	1	25	4	33.33
Retrospective Search	2	50	1	25	1	25	4	33.33

The computed data in the Table 3 demonstrate that, out of 12 libraries surveyed, nearly half of the libraries provide web based literature services, 33.33% of the libraries provide web based Indexing, Reference, Referral, Bibliographic, Abstract, Inter library loan, New arrival and retrospective search services. Remaining 25 % of the libraries provide web based News clipping, SDI, Technical Enquiry and Publication services.

When we made observation between the responded questionnaire and respective library website, we came to know that, all most all library websites provide their link to other sources or libraries, but in the questionnaire only a few members have responded that, they are providing web based referral services. As per our knowledge linking is also directs the library users from one source to other for particular information, which is nothing but referral service. Another aspects is that all most all libraries in the study have maintained the Online Public Access Catalogue (OPAC), but a few of the respondents answered that they are providing retrospective services through manual, automated, LAN and web based modes. We hope retrospective search can be done through OPAC.

5.5 Summing up

It is observed from the study that all the research libraries under the study are following Hybrid culture (manual & electronic) in providing the library and Information services to their clients. It clearly indicates that even though we have sophisticated technologies to provide faster services to its clients, we still retain the manual mode, because a small group of users still needs the manual mode of information services or lack of the infrastructure facilities or budget. But a hybrid type of services is essential in any kind of libraries.

6. References

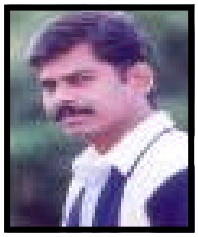
1. Buckland, M. (1992). Redesigning library services: A manifesto. Chicago: American Library Association.

2. Kannappanavar, B U. and Vijayakumar, M. (2001). Use of IT in University of Agricultural Science Libraries of Karnataka. A comparative study. DESIDOC Bulletin of Information Technology, Vol.21 (1), p21-26.

3. Malhan I V and Vijay Gupta. Publicizing the University Library Resources and services through the University Library Home page. In Academic Libraries in Internet Era. Ed by PSG Kumar, CP Vashishth. Ahmedabad: INFLIBNET, 1999. P - 378

4. Sharma, Sumati. (2000). Information Technology in special library environment. DESIDOC Bulletin of Information Technology, Vol. 19 (6), p 19.

About Authors



Mr. Vijayakumar M presently he is working as HOD – Library, MVJ College of Engineering, Bangalore. He has over eight-year professional experience in Dubai and India as a practicing and teaching library professional. Associated with more than 5 library professional association and published more than 15 papers in journals and presented more than 20 papers in national and International conferences. Attended many conferences, workshops and training programmes. He received a certificate copy of ISO -9001 for preparing manual of the systems and procedures of MVJCE Library for ISO 9001 Certification and implemented ISO applications in MVJCE Libraries and passed through NET-UGC exam in 1997. Now he submitted his PhD thesis at Dept. of LIS, Kuvempu University, Karnataka under the guidance of Dr. B U Kannappanavar.



Dr. B U Kannappanavar is presently serving in Kuvempu University library and guiding 4 research students which results in Doctoral degree. He has 18-year of professional experience as a practicing library professional and teaching professional. He acted as BoS Member and BoE Chairman for Under Graduate Courses in Library and Information Science. He is life member for many library professional associations. Published more than 6 papers in journals and books and presented the papers in 28 National Seminars/Conferences. He is also authored for 2 books.



Mr. K N Madhu presently is working in MVJ College of Engineering Library, Bangalore. He holds MLISc from Kuvempu University and pursuing PG Diploma in Computer Applications in Annamalai University, Tamil Nadu. He also attended ASSIST 2004 National seminar.

Digital Library of Theses and Dessertations

G Rathinasabapathy

Abstract

Building of Digital Library largely depends on the nature of content and quality of digital resources. The digital library resources include electronic journals, electronic books, full-text CD-ROM databases, etc. In research and academic institutions, theses, dissertations and research reports play a vital role as primary sources of information since they contain information of research value. Therefore, most of the research scholars refer to these documents regularly. To ensure easy and wider access to these documents, a digital library of theses and dissertations has been established in Tamilnadu Veterinary and Animal Sciences University and connected with the Intranet of the university. The digital library provides access to the constituent units of the university including remote colleges. This paper discusses in brief about the establishment of the digital library of theses and dissertations.

Keywords : Networked Digital Information Service, Digital Library, Digitization, Theses, Dissertations

0. Introduction

The concept of digital library is outcome of the popular use of information technology. It is a library without walls which provides digital information environment in which all the information resources are available and performed through the use of digital technology. It is not merely equivalent to a digitized collection with information management tools but an environment to bring together collection, services and the people in support of the full life cycle of creation, dissemination, use and preservation of data and information. The physical feature of digital libraries do not reside in a specific building. Since the scope is widespread and unlimited, the user can have access to any part of the collection and the information can be made available on user's desktop. The digital library system consists of resources of both text-based and non-text information such as photographs, drawings, illustrations, works of art, numeric data, digitized sound, and moving visual images. A digital library if networked with other library networks such as Local Area Network, Metropolitan Area Network or Wide Area Network and World Wide Web can be accessed world over by any one or anywhere.

1. Defination

There are a number of definitions available for digital library. A 'digital library" may be "remote access from any point in the work at library content and services, combining bibliographic news, electronic texts, image banks and all kinds of information by means of a computer network. The digital library can be called as a 'wall free' electronic workstation to access universal knowledge irrespective of distant location of information.

2. Characteristics of a Digital Library

The following are some of the major characteristics of a digital library.

- Accessibility to network
- Compatibility to handle large amount of information
- Speedy
- Searching and retrieval
- User-friendly interface
- Compatibility to multimedia

2.1 Prerequisites

The following are some of the important prerequisites for establishing a digital library:

- High speed computer
- Multimedia kit
- Fast Scanner
- Server segments
- Printer
- LAN connectivity
- Internet connectivity
- Trained library staff

3. Establishment of Digital Library of Theses and Dessertations

3.1 Building Resources

The basic requirement in creating a digital library will be the building the digital library resources. The digital library resources include various resources such as electronic journal, books, full-text CD-ROM databases, theses and dissertations, etc. Theses and dissertations are playing a vital role as reference sources in academic and research institutes.

Since the Tamilnadu Veterinary and Animal Sciences University is engaged in education, research and extension activities in the fields of veterinary and animal sciences, the theses and dissertations are playing an important role in reference service. It is also observed that a large number of external users also consult these documents frequently. Considering the improved use of these documents by a large number of users, it has been decided to digitize the theses and dissertations.

In the theses and dissertations, the abstract portion is very important since it profiles the brief picture of the whole research work. So, the research scholars always show interest to refer to the abstract part of the theses/dissertation. So, in view of providing the needed materials in digital format and to complete the digitization task with the limited available financial resources, it has been decided to digitize the abstract part of all the theses and dissertations. The theses/dissertations which have been awarded for quality research work have been selected for full-text digitization.

3.2 Digitization

The primary method of building a digital collection is digitization. It is the process of converting analogue information to a digital format. All the 114 awarded Ph.D. theses have been fully scanned using scanners. Abstract part of the remaining theses and dissertations have also been scanned. All the scanned documents have been subjected for proof reading by spell check facility available in MS-Office package and final proof reading manually.

3.3 Designing of the Database

Theses and dissertations will have a number of access points viz., title, author, subject, chairman, etc. It is necessary to provide a searchable database of theses with the required access points to ensure fast and accurate retrieval of required information. Therefore, a database with the following ten access points has been designed using MS-Access.

- Title
- Author
- College
- Degree
- Thesis No.
- Department
- Chairman
- Year
- Keyword

3.4 Designing of the Interface

To ensure easy retrieval of the information, it is necessary to provide a user-friendly interface. In this case, the theses and dissertations have been digitized and a database with all required access points has also been developed. The library user will need an interface to use the digital contents. Therefore, a search screen has been designed using VB-Script so as to search the database using all the available fields.

A PIV multimedia computer with 40 GB HDD is used as the library server to deliver digital contents to the users. The digital library of theses and dissertations has been hosted in the server. The online public access terminals of the library have been connected to the digital library server so as to access the digital library through the OPAC terminals.

3.5 Digital library services through Intranet

The academic departments and research units of the university are functioning in various parts of Tamilnadu viz., Chennai, Madhavaram, Namakkal and Tuticorin. The Namakkal campus and Tuticorin campus are around 400 kilometers and 600 kilometers away from the main library, respectively. Since the research scholars and academicians from those campuses also need these information, it has been decided to extend the digital library facilities for them.

Therefore, the library server has been connected to the Intranet of the university which connects all the academic departments, constituent colleges and research units. With this facility, the research scholars, academicians and students of the university who are in remote areas such as Namakkal and Tuticorin also access the theses and dissertations from their desktops without visiting the library.

3.6 Advantages of the Digital Library

The digital library of theses and dissertations offer several advantages for the library users and few of them are furnished below:

- Fast, accurate and timely access
- More material can be included (in terms of quantity and type)
- Full-text searching facility
- Economical
- Wider access
- Paper-less information retrieval
- User-friendly
- Resource sharing

3.7 Hardware and Software

The following hardware and software were used for the digitization of the documents.

Hardware :

- PIV Computer
- HP Scanjet Scanner
- CD Writer
- Re-writable CDs
- Printer

Software :

- Digit
- MS-Office
- Internet Explorer 5.0

3.8 Cost-effective way to establish digital library of Theses and Dissertations

Nowadays, most of the research scholars are preparing their theses/ dissertations by using computers. It is needless to say that they will have a soft copy of the document with them. So, for future use, the academic institutions may insist the research scholars to submit a soft copy of the theses in the form of a CD along with the print theses. This will help the libraries to a great extent to avoid expenses on digitization of the documents.

4. Conclusion

The digital library of theses and dissertations established in our library has been well received by the library users. The number of persons using the digital library is increasing tremendously. Now, we have insisted the submission of a soft copy of the theses/dissertations to the library while submitting the thesis to the university. This helps us to avoid cost of digitization. Now, we are planning to digitize the annual reports and other publications of the university. It is important to note that the emergence of digital library technologies offers new opportunities to librarians. The digital library accelerates the information capabilities, accessibility and utilization. Libraries have benefited from the increased access to resources, the opportunities for communication and the facilitation of new services that were not possible in pre-digital era. While problems and challenges still exist, it is the need of the hour for any academic library to establish digital library at least with their own resources to ensure better access.

5. References

1. Commings, K. (1997). Libraries of the future. Computers in Libraries, September-1: 136-139.

2. Hulser, R.P. (1997). Digital Library: Content preservation in a digital world. DESIDOC Bulletin of Informtion Technology 17 (7): (1997). 7-14.

3. Shetty, S.C. and Ramaraj, R. (2000). Role of Digital Libraries in Medical Education. University News 38 (37): 36–39.

Screen Shots

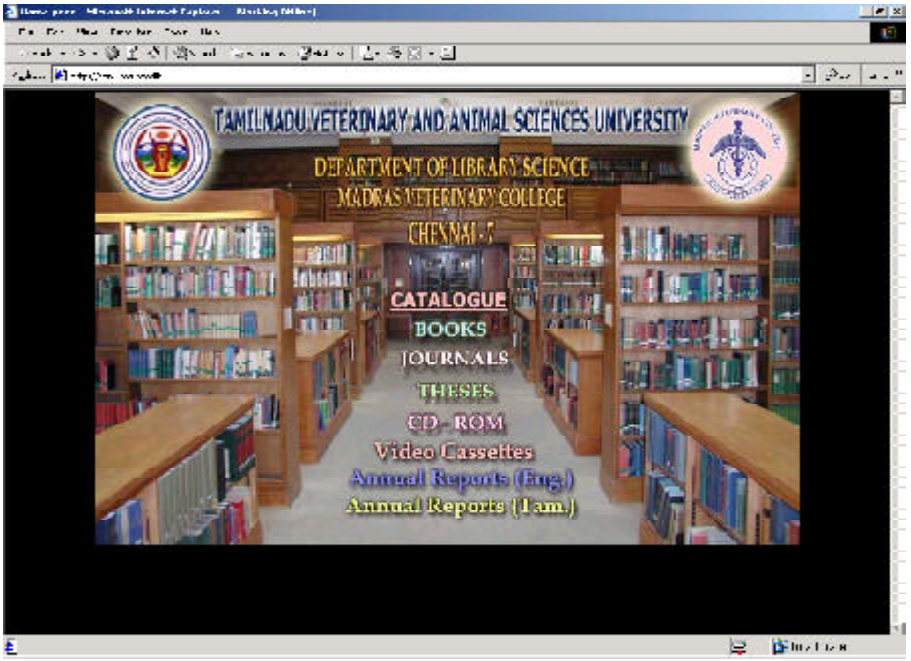


Fig. 1. First page of the Library OPAC



Accession Number - Microsoft Internet Explorer

Abstract Details

Thesis Number	DT 1957
Thesis Title	**EXPERESSION OF F AND HN GENNES OF NEWCASTLE DISEASE VIRUS IN EUKARYOTIC CELLS**
Year	2003
Author Name	**Senthilvelan A**
Chairman	Purushotharnan V
Degree	**Ph.D**
Department	Veterinary Microbiology
College	**MVC**
Abstract	Newcastle disease is regarded throughout the world, as one of the most important diseases of poultry and other birds. It causes severe mortality among chickens and the recurrence of outbreak is not uncommon even after planned vaccination programmes. The causative strains often responsible for outbreaks might be derived from vaccine viruses which had been in use since a long period of time. Hence a local isolate of NDV was subjected to molecular and expression studies. The objectives covered cloning, sequencing and expression of two most important genes, F and HN of a ND viral isolate. The local isolate of NDV was found to be lentogenic in ICPI test in day old chicks. The PGR amplified P and HN genes were cloned in to pUC57 vector and the recombinant plasmids were designated as p57F and p57HN. The clones of p57P and p57HN were sequenced with the help of M13 forward and reverse primers, F and HN internal sense primers to determine the orientation and to read the entire coding region of 1746 bases and 1909 bases of F and HN respectively. Nucleotide sequence analysis of F and HN genes typed the NDV isolate as lentogen. The amino acid motif in the FPCS region of the F gene was "G-R-Q-G-R-L"7. Biocomputational study of both F and HN genes were found to code 553 and 577 amino acids respectively. The F gene had three major hydrophobic regions, six N-linked glycosylation sites and 13 cysteine residues. The HN gene had one major hydrophobic region near the N-terminus, five N-linked glycosylation sites and 12 cysteine residues. Phylogenetic analysis of F and HN genes were carried; out using CLUSTAL W programme. The phylogeny based on F gene sequences grouped the NDV isolate along with other lentogenas such as BI, LaSota, ConnecticatarideFulter. The inserts released by double digestion from recombinant plasmids p57P and p57 HN were purified and ligated to precleaved eukaryotic expression vector VR1012. After transformation and-selection, four clones from F and three i HN were subjected to restriction enzyme digestion to check the presence iserts. The released inserts from the P and HN recombinant plasmids were ndized with DIG labelled F and HN gene probes. The recombinant plasmids e designated as VRF and VRHN respectively. The CEF cells were transfected with the recombinant plasmids VRF and VRHN with the help of CELLFECTIN reagent individually and together to study the coexpression of both proteins in CEF cells. In a similar way both plasmids were also transfected into Vero cells separately. The transfected cells were subjected to IFAT and the presence of apple green fluorescence in the cytoplasm revealed the expression of respective proteins by the cells. The co-transfected CEF cells produced syncytium with fluorescence which revealed the involvement of F and HN proteins in the fusion process and CPE formation. The CEF and Vero cells transfected with VRHN plasmids were also gtreated with chicken and guinea pig erythrocytes 72 h post-transfection and they were found to adsorb chicken and guinea pig erythrocytes respectively. The haemadsorpdon test revealed not only the expression of HN proteins by the cells but also the conservation of structural integrity of the molecules after expression as any change in the conserved structure would affect the important biological function of the protein. Successful expression of both proteins proved the

Fig. 2. Abstract of the Selected Theses

About Author



Mr. G. Rathinasabapathy is working as Assistant Librarian in Tamilnadu Veterinary and Animal Sciences University, Chennai. He holds M. Com., MLISc, M.Phil. He has over 10 years of professional experience in the field of Library and Information Science. Published four books and more than 100 popular articles in the area of LIS, Career counseling and Higher education. Areas of Specialization include : Library management, Digitization, Virtual libraries. Member in ILA, TLA, IASLIC, SALIS and MALA. Attended a number of professional events and presented papers
E-mail : grspathy@yahoo.com, grsaba@vsnl.net

Preservation of Digital Cultural Heritage Materials

P Lalitha T A V Murthy

Abstract

Cultural Heritage materials are being immensely converting into digital forms. The digital materials are at risk of being lost and their preservation for the benefit of present and future generations is an urgent issue to address. The present paper gives a view on the issues related to digital preservation, approaches for digital preservation. It also emphasizes the importance of metadata in digital preservation.

Keywords : Digital Preservation, Media obsolescence, Encapsulation, XML, Universal Virtual Computer (UVC), Migration, Emulation, OAIS, Metadata.

0. Introduction

Our cultural, scientific and information heritage is increasingly converting to the digital forms. Libraries, archives, museums and research institutes are potentially responsible for preserving digital heritage. Modern digital technologies have made it a reality to exhibit large collections of works from multiple cultures. Most of the cultural and heritage materials are being converted into the digitized forms knowing that permanent access to this heritage will offer broadened opportunities for creation, communication and sharing of knowledge among all peoples, as well as protection of rights and entitlements and support of accountability. Since an enormous amount of historical and cultural materials have been created, both storage and distributions raise many challenges [14]. Digitized cultural artifacts should be useable in future for many possible applications. The data should be therefore being preserved for the long term retaining as much of the information as possible [7].

1. What is Digital Preservation ?

According to Sue Mckemmish, digital preservation means “Enable reliable, authentic, meaningful and accessible records to be carried forward through time within and beyond organizational boundaries for as long as they are needed for the multiple purposes they serve.”[16].

According to Cornell University Library, Digital Preservation encompasses a broad range of activities designed to extend the usable life of machine-readable computer files and protect them from media failure, physical loss, and obsolescence.[3]

Digital preservation consists of the processed aimed at ensuring the continued accessibility of digital materials. [12]

2. Need for Preservation

The long term preservation of the intellectual and cultural record of society has occupied librarians, archivists and museum curators for centuries.[9] The long term future of digital resources must be assured, in order to protect investments in digital collections and to ensure that the scholarly and cultural records are maintained in both its historical continuity and media diversity.[10] It would seem that preservation would be much easily achieved with digitized resources, but there are a number of issues that complicate the maintenance of digital objects over a long period. To address few of them:

a. Media Decay which may be caused due to physical, chemical and magnetic fields, etc.,

b. Media Obsolescence

   i. Physical and logical format incompatibilities.

   ii. Unavailability of suitable “drives” or “controllers”.

c. Changes in applications and operating systems which may cause unavailability of operating systems, input output devices, etc., for required software and unavailability of hardware required to run required software.

A digital artifact is completely software dependent as its structure and content could only be understood by the programme that has created it. That particular programme only runs that software. Through this is evident that all digital documents or artifacts are software dependent. So in the present context of media obsolescence, the threat is very real and insidious and will eat away the future of our cultural heritage, knowledge economies, and information society. [1]

In 1964, the first electronic mail message sent from either the Massachusetts Institute of Technology, or the Carnegie Institute of Technology of Cambridge University. The message does not survive however, and so there is no documentary record to determine which group sent the path breaking message. Satellite observations of Brazil in 1970s, critical for establishing a time-line of changes in the Amazon basin, are also lost on the now obsolete tapes t which they were written. [15]

3. Approaches for Digital Preservation

The below given are some important preservation strategies through which digital preservation can be achieved.

3.1 Technology Preservation

To preserve the technology required to access original records for as long as those records are required. But support for the software and hardware eventually ceases and the parts required to maintain the hardware become more and more scarce as manufacturers discontinue obsolete components. The number of machines available that are capable of recording old files continues to decrease, for computers do not last for ever. The skills required to operate the hardware and software also become rare and eventually disappear.

3.2 Printing to paper

Although this approach is still in practice, printing all records to paper is not a viable preservative method. Printing to paper loses functional or behavioral traits that the records had in them digital form. Certain information may also be lost.

3.3 Encapsulation

The encapsulation approach retains the record in its original form, but encapsulates it with a set of instructions on how the original should be interpreted. This would be needed to be detailed format descriptions of the file format and what the information means. The process can be well understood through Fig.1.

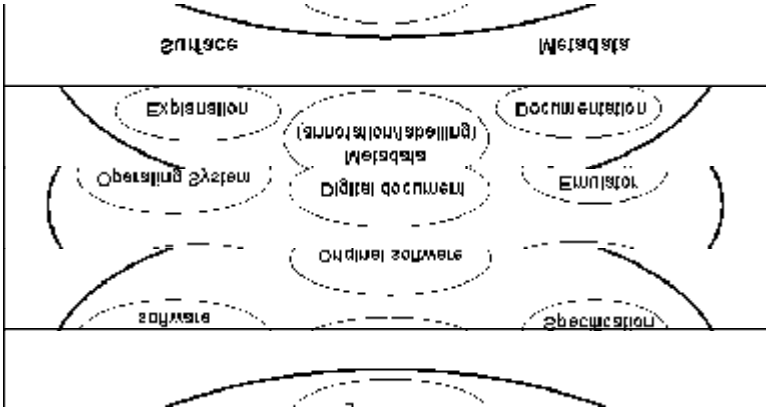


Fig. 1.Encapsulation

3.4 Virtual Machine Software

Raymond Lorie of IBM Almaden has proposed this approach. This addresses the problem of interpreting data files in the future by programming a set if instructions to carry out these interpretations in the machine language of a “Universal Virtual Computer” (UVC). This programme would be written at the time the record was archived and would be preserved together with the record. In order to interpret the record on a future computer, a UVC interpreter would be required and this could be produced fro the specifications of the UVC. With this process the data can be stored in any format and the knowledge required to decode it is encapsulated in the UVC programme.

3.5 XML

Extensible Markup language is a text-based markup language for describing the structure and meaning of data. As it is text based, it is human readable, but desired primarily to be easy to process using the computers. It is a open standard defined by World Wide Web Consortium. Conversions of records to XML format can be seen as a particular type of migration. It is often regarded as a very promising present day data format for archiving and interoperability and so deserves to be considered as an approach in its own right. XML is of the greatest importance for digital preservation, not just because of this widespread uptake, but also because it protects the Achilles’ heel of digital documents: the dependence on obsolete operating systems and application software. It does this by being platform- and software-independent. [5]

3.6 Migration

Migration can be defined as the transfer of files from one hardware configuration or software application to another configuration or application. Problem associated with migration is that the results are often unpredictable, mostly because of a lack of or because the process has not been fully tested. The results of migration are difficult to predict, unless a substantial amount of work is first done regarding the specifications of the source and target formats. Migration can influence the authenticity of a document. Each document that is preserved must be preserved ‘authentically’, otherwise the meaning and validity of the archival record cannot be guaranteed. This has both legal and archivist implications. [6]

3.7 Emulation

The theory behind Emulation is that the only way to ensure the authenticity and integrity or the record over the long term is to continue to provide access to it in its original environment, i.e., its original operating system and software application. This can be only done by preserving not only the record, but also emulator specifications, which contains enough details about the original environment for that environment to be recreated on a future computer whenever necessary. Emulating strategies would involve encapsulating a data object together with the application software and to create or interpret it and a description of the required hardware environment. From the Figure. 2., it is understandable that there are three technical options in emulation.

a. Emulate Applications

b. Emulate Operating Systems

c. Emulate Hardware platforms [8]

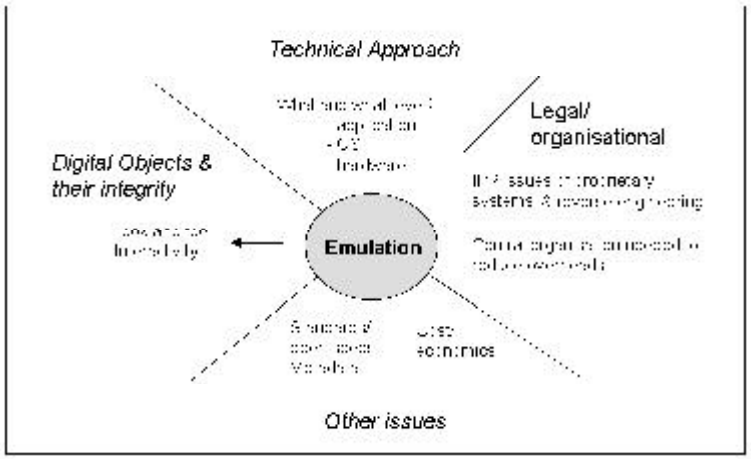


Fig.2. Emulation

It is suggested that these emulator specifications formalisms will require human readable annotations and explanations (metadata). [4] Preservation strategy may be emulation based or migration based, it is that both will have same role that is the long-term preservation of digital information that involves the creation and maintenance of metadata. Within an archive, metadata accompanies and makes reference t each digital object and provides associated descriptive, structural, administrative, rights management, and other kinds of information. This metadata will also be maintained and will be migrated from format to format and standard to standard, independently of the base object it describes. A Digital object enters a repository as a set of sequence of bits; it is accompanied by a variety of metadata related to that object. With proper storage management, replication and refreshing, this set of sequences of bits can be maintained indefinitely. [11] For example, The Pittsburgh Project, The UBC Project, The SPIRT Record keeping Metadata Project are some research projects and practically-based initiatives have been concerned with the development of record keeping metadata schemes and standards.

4. Open Archival Information System (OAIS)

This model has been developed by CCSDS: Consultative Committee for Space Data Systems (NASA) – ISO: 2002. This Reference Model:

- Provides a frame work for the understanding and increased awareness of archival concepts needed for long term preservation and access;
- Provides the concepts needed by non-archival organizations to be effective participant in the preservation process;
- Provides a framework for describing and comparing different long term preservation strategies and techniques;
- Expands consumers on the elements and processes for long-term digital information, preservation and access, and promotes a larger market which vendors can support.

The reference model defined common terminologies like :

- AIP (Archival Information Package)
- SIP (Submission Information Package)
- DIP (Dissemination Information Package)
- PDI (Preservation Description Information)

It has also discussed many important issues like

- Ingest formats and processing
- Use of standards
- Metadata
- Existing Records (bibliographic record in world Catalog).

As the present paper emphasizes on the digital preservation, Preservation Description part of the OAIS information model has been briefly described.

The OAIS information model divides Preservation Description Information into four categories:

i. Reference Information : It describes identification systems, and the mechanisms for providing assigned identifiers, used to unambiguously identify the Content Information both internally and externally to the archive in which it resides.

ii. Context Information : It documents relationships of the Content Information with its environment, including the reasons for its creation and relationships to other Content Information objects.

iii. Provenance Information : It documents the history of the Content Information, including its origin, changes to the object or its content over time, and its chain of custody.

iv. Fixity Information : It provides the Data Integrity checks or Validation/Verification keys used to ensure that the particular Content Information object has not been altered in an undocumented manner.

In a nutshell, Preservation Description Information records the identity, relationships, history and integrity of the archived Content Data Object. With this it is understandable that Effective Metadata is a necessary condition for effective digital preservation. The elucidation and maintenance of Preservation Description Information, however, is the keystone to building an information infrastructure to support the processes associated with digital preservation. [13]

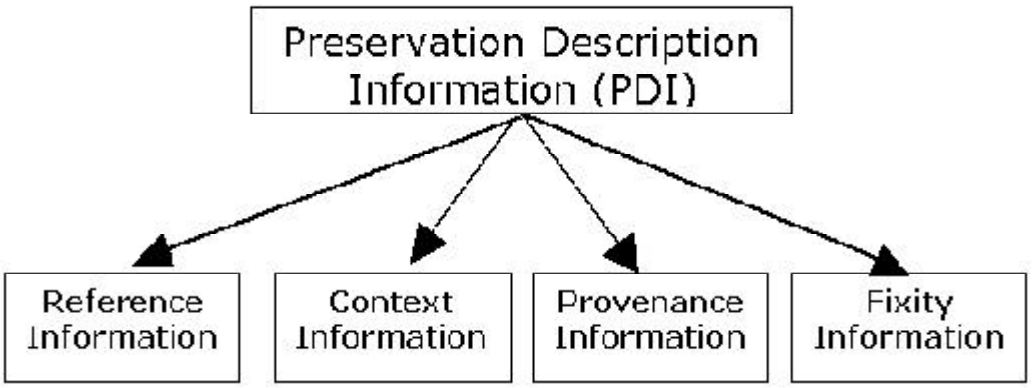


Fig. 3. PDI

In OAIS model the Digital Migration is defined to be the transfer of digital information, while intending to preserve it, within the OAIS. It is distinguished from transfers in general by three attributes:

- a focus on the preservation of the full information content,
- a perspective that the new archival implementations of the information is a replacement of the old; and
- full control and responsibility over all aspects of the transfer resides the OAIS.

But recognized that: the Digital Migrations are time consuming, costly, and expose the OAIS to greatly increased probabilities of information loss. Therefore, an OAIS has a strong incentive to consider Digital Migration issues and approaches. [2]

5. Conclusion

Even though it is difficult to decide which digital materials are to be preserved, there should be a strict plan for preservation according to the organizational agendas. In possible environments, it is always better to use the standard formats for preservation. Although the Preservation Community is at its disposal for providing solutions for digital preservation, till date there are no scalable solutions for the general problem of digital preservation.

6. References

1. Beagrie, Neil., 2004. “The Continuing Access and Digital Preservation Strategy for the UK Joint Information Systems Committee (JISC)”, *D-Lib Magazine*, 10(7/8). http://www.dlib.org/dlib/july04/beagrie/07beagrie.html Accessed on 14.09.04.

2. CCSDS: Consultative Committee for Space Data Systems (NASA), 2002. http://www.ccsds.org/documents.650xob1.pdf Accessed on 20.9.04.

3. Cornell University Library, 2003.Digital Preservation Management: Implementing Short-term strategies for Long-term Problems. http://www.library.cornell.edu/iris/tutorial/dpm/terminology/preservation.html Accessed on 06.03.04

4. Day, Michael., 1999. "Metadata for Digital Preservation: an Update", *Ariadne*. 22. http://www.ariadne.ac.uk/issue22/metadata/intro.html Accessed on 06.03.04.

5. Digital Bewarning, 2002. Digital Preservation Test bed White Paper, XML and Digital Preservation http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_xml_en.pdf Accessed on 21.10.04.

6. Digital Bewarning, 2003. Digital Preservation Test bed White Paper Emulation: Context and Current Status. http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.pdf Accessed on 21.11.04.

7. European Commission on Preservation and Access, Amsterdam, 1997. Digitization as a Means of Preservation?: Recommendations for the digitization of microfilm, http://www.clir.prg/pubs/reports/digpres/digpres3.html Accessed on 12.10.04

8. Granger, Stewart; 2000. "Emulation as a Digital Preservation Strategy", *D-Lib Magazine,* 6(10). http://www.dlib.org.october00/granger.html Accessed on 16/8/04

9. Hedstrom, Margaret., Digital Preservation: a time bomb for Digital Libraries.

   http://www.uky.edu/~kiernan/DL/hedstrom.html Accessed on 06.03.04.

10. Lavoie, Brian; Dempsey, Lorcan., 2004. "Thirteen Ways of Looking at ...Digital Preservation", *D-Lib Magazine*, 10(7/8). http://www.dlib.org/dlib/july04/lavoie/07lavoie.html Accessed on 14.09.04.

11. Lynch,Clifford., 1999. "Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information". *D-Lib Magazine*, 5(9). http://www.dlib.org/dlib/september99/09lynch.html Accessed on 06.03.04.

12. National Library of Australia, 2003. Guidelines for the Preservation of Digital Heritage. http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicities/unesco_guidelines.pdf. Accessed on 04.09.04.

13. OCLC/RLG Working Group on Preservation Metadata, 2002. Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects. http://www.oclc.org/research/pmwg/ Accessed on 03.04.04.

14. Report of the DELOS-NSF Working Group on Digital Imagery for Significant Cultural and Historical Materials. http://www.delosnsf-imaging.unifi.it/assets/wg-final_report.doc Accessed on 16.09.04

15. Research Libraries Group, Inc., May 1, 1996. Preserving Digital Information: Report of the Task Force on Archiving of Digital Information, Commission on Preservation and Access http://lyra2.rlg.org/ArchTF/tfadi.index.htm Accessed on 01.04.04

16. Rothenberg, Jeff., 2002. Digital Preservation: The State of the Art. http://www.kb.nl/kb/hrd/dd/dd-links_en_publicaties/workshop2002/ rothenberg.pdf Accessed on 10.04.04

About Authors



Ms. P Lalitha is the Librarian at Kesar SAL Medical College & Research Institute. She holds MLISc and pursuing Ph.D. in LIS from University of Bundelkhand , Jhansi. She is also teaching LIS students of IGNOU at Ahmedabad Regional Centre. She has published three papers, and her area of interest is Digitization of Artifacts related to Indian Heritage and Culture.
E-mail : lalithapoluru@yahoo.com.



Dr. T A V Murthy is Director of INFLIBNET Centre, Ahmedabad and holds B Sc, M L I Sc, M S L S (USA) and Ph.D. He is President and Fellow of SIS-India, Hon. Director of E.M.R.C., Gujarat University and Member Secretary of ADINET, Ahmedabad. He carries with him a rich experience and expertise of having worked in managerial level at a number of libraries in many prestigious institutions in India including National Library, IGNCA, IARI, Univ of Hyderabad, ASC, CIEFL etc and Catholic Univ and Case western Reserve Univ in USA. He has been associated with number of universities and has guided number of Ph.Ds and actively associated with the national and international professional associations, expert committees and has published good number of research papers. He visited several countries and organized several national and international conferences and programmes.
E-mail : tav@inflibnet.ac.in

Preservation and Digitisation of Rare Collection of Dr. Panjabrao Deshmukh Smruti Sangrahalaya, Amravati

Vaishali G Choukhande Jitendra Dange

Abstract

The paper emphasises on the preservation, digitisation and dissemination of information in digitised or electronic form to the end user. It is a practical approach to digitise the rare collection of Dr. Panjabrao Deshmukh Smruti Sangrahalay, Amravati. The author also highlights the issues of copyrights faced during the work.

Keywords : Preservation, Digitization.

0. Introduction

Computerisation is changing forever the way information is being created, managed and accessed and is revolutionising our ability to communicate, analyse and reuse that information. At the same time electronic data needs active management from its creation, if it is to survive and be kept accessible in a technological environment where there is rapid change and evolution of hardware and software.

The increasing use of the Internet and World Wide Web has developed awareness and concerns about access and retrieval of information across networks. Digital library is a new concept. The concept has brought phenomenal change in the information collection, preservation and dissemination scene of the world.

Digitisation is also a high-speed data transmission technique. It is the conversion of any fixed or analog media (such as books, journals, articles, photos, painting, maps, microforms etc) into electronic forms through scanning, sampling or rekeying by using various technologies.

Digitisation refers to the conversion of an item in printed text, manuscript, image or sound, film and video recording from one format (usually print or analogue) into digital. The process basically involves taking a physical object is captured using a scanner or digital camera and converted to digital format that can be stored electronically and accessed via a computer.

Digitisation is quite simply the creation of computerised representation of printed analogue. It may also refer to all the steps in the process of making available collection of historical material available in digital form.

The International Encyclopaedia of Information science defines digitisation as “The process of converting analogue information to digital format. In communication this is the process of converting analogue signals to digital signals”. In Information systems, digitization often refers to the process of converting an image (such as photograph) using some type of scanning device (or digitizer), into digital representation so that it can be displayed on a screen and manipulated”.

According to Canadian Heritage Information Network “Digital technology have helped Institution’s goal either highlighting particular aspects of local history or reaching national and international audience. Collections that were once too remote to be viewed are now accessible; object that were once too fragile to be handled or exhibited can now be seen by broad audiences. By making it possible to bring together diverse materials or collections from scattered locations for comparison and research, digital technology can be a powerful teaching aid, especially when institutions work together to create a critical mass of complementary material.”

1. Statement of the Problem

There are several reasons for choosing to preserve and digitise Dr. Panjabrao Desmukh's special collections available in Dr. Punjabrao Deshmukh Smruti Sangrahalaya, Amaravti. Typically, special collection items are unique, to provide service to scholars they need to be put in digitised form. Special collection items are often fragile, oversize or it is in need of particular care, and as unique items they are irreplaceable. So that replacing their use with the use of some digital surrogate helps with preservation. It can also become much faster to look at them digitally than to browse materials, which must be handled slowly and carefully to avoid deterioration.

Today's library and information centres have passed through long evolutionary sequence. Now the concept and the very name "Library" has changed to " Library walls", Network library", Desktop library, "Logical library", Virtual library, " Information Nerve centre", Information Management centre" and lastly "Digital library".

Today we are living in exciting times of digital libraries, whose history spans a mere dozen years. They will surely figure amongst the most important and influential institutions of the new century. The information revolution supplies the technological horsepower that drives digital libraries and accessing information. If information is the currency of the knowledge economy, digital libraries will be the banks where it is invested.

One reason for using digital library technology is to manage large amounts of digital content such as millions of lines of textual material, thousands of images or hundreds of audio clips. Advances in storage technologies have enabled large amounts of contents to make available locally at increasingly affordable costs.

Dr. Punjabrao Deshmukh Smruti Sangrahalaya, Amravati was established in the year 1926. It has a very rich collection of 2000 books, 500 bound volumes, 1200 photographs, 56 manuscripts, 10 diaries 500 circular letters, 10 audiovisual material including films, magnetic tape, videocassette, camera and gramophone records. The collection includes Agriculture, Education and Law subject. At present 30 research scholars are working on the vast literature available in Dr. Punjabrao Deshmukh's collection in Smruti Sangrahalaya, Amravati. The collection is valuable and hence need to be digitised. Indeed this is a challenging and promising task but one has to undertake such kind of activity which will not only help librarians, but the entire humanity as a whole.

2. Significance of the study

The study emphasises on the preservation and Digitization of rare collection of Dr. Punjabrao Deshmukh Smruti sanghralaya, Amarvati.

The study helps the researchers to get instant access to the rare manuscripts, photographs etc. from Dr. Punjabrao Deshmukh Smruti sanghralaya, Amarvati.

3. Aims and Objectives

- ✍ To create, manage and preserve the collection in digitised form.
- ✍ To make the digitised form available to the users.
- ✍ To preserve the information content as back-up on a long-lasting medium. Here long-lasting would mostly mean more than 100 years.

- To preserve the manuscripts for easy dissemination.
- To enhance intellectual control through creation of new finding aids, links to bibliographic records and development of indices and other tools.
- To increase and enrich use through the ability to search widely, manipulating images & text and to study disparate images in new contexts.
- To encourage research scholar's use through the provision of enhanced resources in the form of widespread dissemination of local or unique collections.
- To enhance use through improved quality of image, for example, improved legibility of faded or stained documents; and
- To create a "virtual collection" through the flexible integration and synthesis of a variety of formats or of related material scattered among many locations.

4. Methodology

The experimental method was used to conduct study. Dr. Punjabrao Deshmuk's collection available in Smruti Sangrhalaya, Amravati was categorised and classified into Personal collection, Agricultural, Co-operative, Parliamentary, Social work, Educational and Law.

There are four steps involved in the process of digitisation : Scanning, indexing, storage and retrieval.

Scanning : The scanning process involves acquisition of an electronic image through it original that may be a photograph, text, manuscript etc. into the computer using an electronic image scanner.

Indexing : Indexing of a document converted into an image or text file is the second step in the process of document imaging. The process of indexing scanned image involves linking of database of scanned images to a text database.

Storage : The most tenacious problem of a document image relates to its file size and therefore, to its storage. Every part of an electronic page image is saved regardless of presence or absence of link. The file size varies directly with scanning resolution, the size of the area being digitised, compression ratio, content and the style of graphic file format used to save the image.

Retrieval : Once scanned images and OCR text documents have been saved as a file, a database is needed for selective retrieval of data contained in one or more fields within each record in the database.

5. Problems & Prospects

Library and Information Science profession has remained witness to the last 200 years history of Intellectual Property rights movements. While digitising the rare collection, the researcher faced many problems due to copyright law and related issues i.e. copying, accessing, archiving and preservation. As the rare collection belongs to the Shivaji Education society, Amravati, and Dr. Panjabrao Deshmukh was the founder of the society, they permitted to preserve the rare collection in electronic form to have an access to the end user.

6. Conclusion

The manuscripts and photographs were digitised in the image form, while the book form collection was converted into text format to save the space and bytes. Digitisation is a challenging and promising task but one has to undertake such kind of activity, which will not only help librarians, library professionals but the entire humanity as a whole.

Technologies pose challenges as well as opportunities before us. It is for us to use these technologies for survival/success and to cope up with time. There is great hope that digital technology can help to preserve and make more accessible many rare and fragile items, because the quality of digital image is high and the use of electronic surrogates is easier than that of any form. Preservation is the art of managing risk to the intellectual and physical heritage of a community and all the members of that community have a stake in it. Dr. Panjabrao Deshmukh Smruti Sangrahalay, Amravati is an example of making a small beginning in the digitisation process. Digitisation will help the end user to make use of the digitised resources and create a new wave in the library use.

7. References

1. Choukhande, Vaishali G. Digitisation of manuscript collection of Shardashram Ethihas Sanshodhan Sanstha Wachnalaya, Yavatmal. Edited by A. Vaishnav and S. Sonawane. *In* Impact of digitisation on development of Information professionals organised by Dept. of Library & Information Science, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, 2003.
2. Nagnath, Ramdasi Digitisation and emerging roles for LIS professionals relating to intellectual property rights in India. Edited by S. P. Satarkar. *In* Press and registration act, delivery of books act and copyright act. Nanded: Swami Ramanand Teerth Marathwada University, 2004.
3. Vatnal, R. M. and Ramesha. Digital archiving of manuscripts: A case study of Karnatak university library. Edited by M. Bavakutty and others.New Delhi: Ess Ess pub. , 2003.

About Authors



Dr. Vaishali Choukhande is librarian at Shri Shivaji Science College, Amravati, India. She holds B.Sc. , MLISc. , D.C.P., D.C.P., MS-CIT, Ph.D. She worked as a Principal in Nagar Wachanalaya Mahavidhyalaya, Deptt. Of Lib. & Inf. Science, Yavatmal for 7 years. She has to her credit 17 articles published in fetscrift proceeding and presented papers in International, National, and State level conferences. Organised one day seminar on “Medium of Instruction in LIS education and innovation of teaching methods in LIS education”. And also organised one day workshop on “Research methods for teachers in LIS”. 16 students were under supervision for MLIS dissertation.



Mr. Jitendra Dange holds B.A.(English literature) and is studying MLIS in Vidyabharati Mahavidhyalaya, Amravati.

Web Services and Interoperability : Security Challenges

S K Sharma G K Sharma P N Srivastava

Abstract

The Web services framework intends to provide a standards-based realization of the service-oriented architecture (SOA) over Internet, which has emerged in response to a fundamental shift from program-to-consumer (B2C) to program-to-program (B2B) interactions. Fully Interconnected enterprises are being replaced by business networks in which each participant provides the others with specialized services. This new service architecture defines a set of requirements that distinguish SOA from other services architecture. Security and Web services are consistently reported among the top technologies of interest to business. Concerns about the security technology are major deterrent to companies considering use of the technology. This paper attempts to explain the new Web Services security and mentions the main initiatives and their respective specifications.

Keywords : Web Services, XML, XML-Signature, XML-Encryption, WS-Security.

0. Introduction

Everyone knows roughly what a “Web Service” is, but there is no universally accepted definition. The definition of web service has always been under hot debate within the W3C Web Services Architecture Working Group[1]. The precise definition of Web Services is still evolving as witness by the various definitions in the literature. One such the definition is that a Web Services is convergence between SOA and web with at least the following additional constraints:

- Interfaces must be based on Internet protocols such as HTTP, FTP, and SMTP.
- Except for binary data attachment, messages must be in XML format.

It is also defined as the web applications that are self-contained, self-describing, modular that can be published, located, and invoked across the web. They perform functions, which can anything from simple request to complicated business process [2]. W3C (World Wide Web consortium) define a Web service as a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards [3].

At a minimum, web services can be any piece software that makes itself available over the Internet using standardized web services messaging system and interface [4]. This paradigm has the potential to deliver many benefits for business. Some of these are:

- Developer will be able to respond quicker with the demanding business needs to link up partners or to provide access to existing business assets within the firewall.
- Functionality from the heterogeneous development platform (.NET, CORBA, J2EE) can be quickly integrated into business applications.

- ✍ The web services paradigm, being platform agnostic, solves the basic EAI (Enterprise application integration) problem of having uniform interface to business applications. In the future, packaged software application will expose service-oriented interface based on web service standard.
- ✍ Investments made in your internal infrastructure can be leveraged to provide new channel for providing service to your customers.
- ✍ Current applications that have developed in any language can be exposed as web services quickly using web services tools, web services tools will be available on all development platforms.
- ✍ The services-oriented paradigm that web services builds on promoters the use of clean interfaces that allow a business asset being wrapped by an interface to be replaced, changed, or outsourced as business needs requires without affecting customers or business process that depend on the assets.
- ✍ Developers will be able to use number of web services provided be third parties to deliver more powerful and integrated business solutions.

Due to these immediate benefits, most of IT department are implementing this technology with the higher-priority objective of making them operable leaving aside, at least until later stages, the problem related to security. XML, eXtensible Markup Language, lies at the core of web services interoperability, and enables it to provide a language-neutral and platform-independent way of linking applications [5].

However, these characteristics also present new security threats and challenges. Security is being considered as the biggest roadblock facing widespread implementation of web services technologies.

The security solutions that are commonly implemented in today's web-based applications, such as SSL (Secure Socket Layer), do not provide a sufficient security infrastructure for web services [6]. In contrast to web site applications that require a security solution between the client browser and the web server, the web services applications can involve calling one or more intermediary services, thus requiring more comprehensive security solution. If a transaction passes through intermediary systems the integrity of the data and security of information that flows with it may be lost [7]. Also, the user credentials cannot be easily passed through each stop in the transaction chain. To solve the security issues in the above scenarios, web service security architecture requires mechanism that provides security for the entire transaction.

The Web Services reference Model

Interactions among Web Services involve three types of participants: **Service Providers, Service registry, and service users** (Fig1.0).

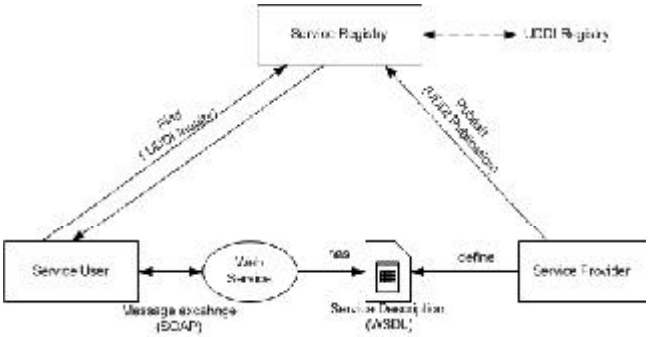


Fig 1.0

Service Providers are the parties that offer services. They define descriptions of their services and publish them in the service registry, a searchable repository of service descriptions. Each description contains details about the corresponding service such as its data types, operations, and network location. Service users use a find operation to locate services of interest. The registry returns the description of each relevant service. The user uses this description to invoke the corresponding web service.

Three major standardization initiatives have been submitted to the W3C consortium to support interactions among Web Services.

- WSDL (Web Service Description Language) : WSDL [8] is an XML-based language for describing operational features of Web Services. WSDL descriptions are composed of interface and implementation definitions. The interface is an abstract and reusable service definition that can be referenced by multiple implementations. The implementation describes how the interface is implemented by a given service provider.
- UDDI (Universal Description, Discovery and Integration) : UDDI [9] defines a programmatic interface for publishing and discovering Web Services. The core component of UDDI is the business registry, an XML repository where businesses advertise services so that other businesses can be find them. Conceptually, the information provided in a UDDI business consists of white pages (contract information), yellow pages (industrial categorization), and green pages (technical information about services).
- SOAP (Simple Object Access Protocol) : SOAP [10] is a lightweight messaging framework for exchanging XML formatted data among Web Service. SOAP can be used with a variety of transport protocols such as HTTP, SMTP, and FTP. A SOAP message has a very simple structure: an XML element, the header includes features such as security and transactions. The second element, the Body includes the actual exchanged data.

2. Interactions in Web Services

Web Services allow interactions at the communication layer by using SOAP as a messaging protocol. The adoption of an XML-based messaging over well-established protocols (e.g. HTTP, SMTP, and FTP) enables communication among heterogeneous systems.

At the content layer, Web Services use WSDL language. WSDL recommends the use XML Schema as a canonical type system (to associate data types to message parameter). However, the current version of WSDL does not model semantic features of Web Services. For example, no constructs are defined to describe document types (e.g. whether an operation is a request for quotation or a purchase order).

Web Services are still at a maturing stage. Hence, they still lack the support for interactions at the business process layer. To date, enabling interaction among Web Services has largely been an ad hoc process involving repetitive low-level programming. Standardization efforts such as BPEL4WS (Business Process Execution Language for Web Services) are underway for enabling the definition of business process through Web Services composition.

WSDL does not currently include operations for monitoring Web Services such as checking the availability of an operation or the status of a submitted request. Additionally, neither UDDI nor WSDL currently define quality of service parameters such as cost and time. In terms of adaptability, changes may occur in operation signatures (e.g. name), messages (e.g. number of parameters, data type), service access (e.g., port address), and service and operation availability. The process with changes is currently ad hoc and manually performed.

3. Main Web Services security Issues

Security in Web Services needs to be addressed at different levels including communication, description, and firewall. Some of the major security issues that web services technologies must address:

- Authentication: Any Web Services that participate in an interaction may be required to provide authentication credentials by the other party (e.g. a pair username/password or an X.509 certificate).
- Authorization: Web Service should include mechanisms that allow them to control access to the services being offered. They should be able to determine who can do what and how on their resources.
- Confidentiality: Keeping the information exchanged among Web Services nodes secret is another of the main properties that should be guaranteed in order to consider the channel secure.
- Integrity: This property guarantees that the information received by a Web Service remains the same as the information that was sent from the client.
- Non-repudiation: In the Web services world, it is necessary to be able to prove that a client utilized a services and that service processed the client request. This security issues is covered by Digital Signatures.
- Availability: The need to take care of the availability aspects for preventing denial-of-service attacks or to arrange redundancy systems is a crucial point in Web Services technology.
- End-to-End Security: Network topologies require end-to-end security to be maintained all across the intermediaries in the message's path. " When data is received and forwarded on by an intermediary beyond the transport layer, both the integrity of the data and any security information that flows with it may be lost" [7]. This forces any upstream message processors to rely on the security evaluations made by previous intermediaries and to completely trust their handling of the content of messages.

In addition, the above mentioned issues, which are inherited from the distributed computing classical scheme, Web Services should also address the issues arises from the new threats created by its own nature such as:

- Availability of higher number of standard specifications;
- Most of specifications are in draft state;
- XML standard format needed to structure the security data;
- Application-level, end-to-end and just one-context-security communications;
- Interoperability of the requirement and online security elements;
- Audit, automatic and intelligent contingency processes aimed at being machine-to-machine interactions not controlled by humans;
- Online availability management in critical business process;

4. Core Web Services Security Specifications

The core Web Services specifications are XML, SOAP, WSDL, UDDI. These specifications have been broadly adopted by the industry, and constitute the basic building blocks on which Web Services are

being designed and implemented. The bad news is that these four operative services specification allows creation of Web Services but they do not say anything about how to secure them.

XML and SOAP both specifications do not say anything about how to obtain integrity, confidentiality, and authenticity of the information that they respectively represent and transport.

Numbers of questions are associated with the UDDI and WSDL specifications such as: “Is the UDDI registry located in a trustworthy location? How can we be sure that the published data has not been maliciously manipulated? Was the data published by the business it is supposed to have been published by? Can we rely on the business that published the services? Are the services available at any moment? Can we trust the transactions that are produced from the execution of the services?” As we can see from all these questions, an in-depth analysis of the security problems that UDDI and WSDL architecture implies is needed [9].

Two new security initiatives designed to both account for and take advantage of the special nature of XML data are XML Signature and XML Encryption. Both are currently progressing through the standardization process. XML Signature is a joint effort between the World Wide Web Consortium (W3C) and Internet Engineering Task Force (IETF), and XML Encryption is solely W3C effort. In addition, XML-key Management System standard is associated with these two standards.



Fig 2.0

4.1 XML-Digital Signature

It defines how to digitally sign XML content and hot to represent the resulting information to an XML schema. A digital signature grants information integrity and non-repudiation [11]. Thus, for example entity cannot deny the authorship of digitally signed documents.

According to the XML Digital Signature specification, digital signature can be applied any kind of digital content, including XML. It has ability to sign only specific portion of the XML tree rather than the complete document. This is important when a single XML document may need to be signed by multiple times by a single or multiple parties. This flexibility can ensure the integrity of certain portion of an XML document, while leaving open the possibility for other portions of the document to change. The signature validation mandates that the data object that was signed be accessible to the party that interested in the transaction. The XML signature will generally indicate the location of the original signed object.

4.2 XML Encryption

It provides a model for encryption, decryption, and representation of full XML documents, single XML elements in an XML document, contents of an XML element in an XML document, and arbitrary binary content outside and XML document [11].

XML encryption solves the problem of confidentiality of SOAP messages exchanged in Web Services. It describes the structure and syntax of the XML elements that represent encrypted information and it provides rules for encrypting/decrypting an XML document (or part of it).

The specification states that encrypted fragments of a document should be replaced by XML elements specifically defined in the recommendation. In order to recovery the original information, a decryption process is also specified.

4.3 XML Key Management Specification (XKMS)

It is an XML-based way of managing the public key infrastructure (PKI), a system that uses public-key cryptography for encrypting, signing, authorizing and verifying the authenticity of information in the Internet [12]. It specifies protocols for distributing and registering public keys, suitable for use in conjunction with the proposed standard for XML Signature and XML Encryption.

XKMS allow implementers to outsource the task of key registration and validation to a “trust” utility. This simplifies implementation since third party does the actual work of managing public and private key pairs and other PKI details.

5. WS-Security Family Specification

IBM and Microsoft with other major companies have defined Web Services security models that guarantee end-to-end communication security. The center of these specifications is composed of WS-*Policy*, Ws-*Trust*, WS-*Privacy*, WS-*SecureConversation*, WS-*Federation*, WS-*Authorization*, and WS-*Security*.

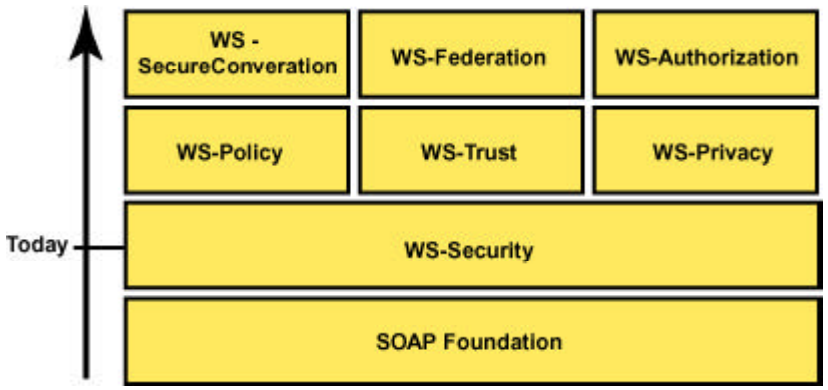


Fig 3.0[4]

5.1 WS-Security

The most important work in this area is *WS-Security* specifications from IBM, and Microsoft, and VeriSign [13]. The three companies jointly developed the new specification, known as *WS-Security*, and have submitted it to two major standardization organizations: W3C, World Wide Web Consortium, and the OASIS, Organization for the Advancement of Structured Information Standards.

WS-Security describes enhancements to SOAP messaging to provide *quality of protection* through message integrity, message confidentiality, and single message authentication. These mechanisms can be used to accommodate a wide variety of security models and encryption technologies.

WS-Security is placed at the base of the security specification pile. Other specifications that directly relate to security issues are being developed based on WS-Security. In the protocol stack and left on top of the WS-Security, we find WS-policy, WS-trust, and WS-privacy. WS-Policy will describe how senders and receivers can specify their requirements and capabilities. WS-Trust defines XML Schema as well as protocols that allow security tokens to be accessed, validated, and exchanged. WS-Privacy will describe how organizations state the privacy policy so that incoming request make claim about the sender's adherence to these policies.

The top layer consist three protocols, which are follow-on specifications. WS-SecureConversation will describe how a Web service can authenticate requester messages, how requesters can authenticate services, and how to establish mutually authenticated security contexts. It is designed to operate at the SOAP message layer so that the messages may traverse a variety of transports and intermediaries. This does not preclude its use within other messaging frameworks. In order to further increase the security of the systems, transport level security may be used in conjunction with both WS-Security and WS-SecureConversation across selected links. WS-Federation specifications define how to construct federated trust scenarios using the WS-Security, WS-Policy, WS-Trust, and WS-SecureConversation. It also defines the mechanism for managing trust relationships. WS-Authorization specifications describe how access policies for a Web service are specified and managed. In particular it will describe how claims may be specified within security tokens and how these claims will be interpreted at the endpoint.

5.2 Security Assertion Markup Languages (SAML)

SAML is an Extensible Markup Language standard (XML) [2] that supports Single Sign On. SAML allows a user to log on once to a web site and conduct business with affiliated but separate web sites. SAML can be used in B2B and B2C transactions.

There are three basic SAML components: Assertions, protocol, and binding. Assertion can be one of three types: authentication, attribute, and authorization. Authentication assertion validates the identity of the user. The attribute assertion contains specific information about the user. While, the authorization assertions identifies what the user is authorized to do.

The protocol defines how SAML request and receives assertions. There are several available binding for SAML. There are bindings that define how SAML message exchanges are mapped to SOAP, HTTP, SMTP and FTP among others. OASIS is the body developing SAML.

5.3 XACML: Communicating Policy Information

XACML is an Extensible Markup Language standard (XML) based technology, developed by OASIS for writing access control policies for disparate devices and application. It includes an access control language and request/response language that let developers write policies that determine what users can access on a network or over the Web. XACML can used to connect disparate access control policy engines.

5.4 Liberty Alliance Project

The Liberty Alliance Project is led by Sun Microsystems, and it s purpose is to define a standard federation framework that allows services such as Single Sign-On [14].

Thus, the intention is to define an authentication distributed system that allows intuitive and seamless business interactions. This purpose is the same as those of the WS-Federation specifications and Passport's .NET technology. Once again, this is another example of the previously so-called overlap problem in Web Services Security solutions.

6. Summary of the Current Web Services Standards

Authentication	WS-Security, WS-Trust (Draft), XKMS, SAML, Liberty Alliance Project, WS-Federation (Draft)
Authorization	XACML, WS-Authorization, (Draft)
Confidentiality	XML-Encryption, WS-Security
Integrity	XML-Digital Signature
Non-repudiation	XML-Digital Signature, WS-Security
Security Policy	WS-Policy, WS-SecurityPolicy (Draft), XACML
Trust authority	WS-Trust (Draft); XKMS
Security Context	WS-SecureConversation (Draft)
Delegation/Proxy	WS-Trust (Draft), Delegation has not yet been fully addressed
Privacy	WS-Privacy

7. Conclusion

In spite of the amount of specifications, there are many unresolved security issues that will have to be addressed in the future. The explosion of specifications and concepts, and lacking of a global standardization initiative is causing overlapping solutions to similar problem. In, addition, the problems relating to security vulnerabilities, which would be introduced in complex WS implementation using different security tokens, have not been sufficiently addressed. This fact will require an extra effort in the future not only for the specifications to unify and make themselves interoperable but also for industry to adopt and easy implement them.

8. References

1. WSAS Web Services Architecture Draft 8 August 2003 (2003). Seehttp://www.w3.org/TR/2003/WD-ws-arch-20030808/

2. SAML Specifications V1.1 –http://www.oasis-open.org/committees/download.php/791/sstc-acml-1.1-cs-02.zip

3. Web Services architecture W3C working group note 11th Feb, 2004 See http://www.w3.org/TR/ws-arch/.

4. WS-Security Specifications V1.0 – Chris Kaler (Editor). *WS-Security, Version 1.0*. An IBM, Microsoft and VeriSign joint specification. April 5, 2002.http://www-106.ibm.com/developerworks/webservices/library/ws-secure/

5. Gottschalk K., Graham S., Kreger H., and Snell J., Introduction to Web Services architecture IBM System journal, Volume 41, Number 2, 2002 http://www.research.ibm.com/journal/sj/41/gottschalk.html.

6. Mark Curphey (OWASP), et. al. A Guide to Building Secure Web Applications and Web Services. Version 1.1. Sep. 22, 2002. http://www.owasp.org/guide/

7. Security in a Web Services World: A Proposed Architecture and Roadmap, IBM/Microsoft White paper – http://www-106.ibm.com/developerworks/security/library/ws-secmap/?dwzonw=security

8. WSDL Web Service Description Language (WSDL) 1.1 – W3C Note Mach 2001. See http://www.w3.org/TR/wsdl

9. UDDI version 3.0.1 – UDDI spec Technical committee Specification 14 October 2003. See. http://uddi.org/pubs/uddi-v3.0.1-20031014.htm.

10. Don Box (DevelopMentor), et. al. SOAP: Simple Object Access Protocol 1.1 W3C Note. May 8, 2000.http://www.w3.org/TR/2000/NOTE-SOAP-20000508/.

11. Ed Simon, Paul Madsen and Carlisle Adams. An Introduction to XML Digital Signatures. August 8, 2001. http://www.xml.com/pub/a/2001/08/08/xmldsig.html

12. XML Key Management Specification 2.0 (XKMS) W3C Working Draft 18 March 2002, http://www.w3.org/TR/xkms2/

13. Chris Kaler (Editor). WS-Security, Version 1.0. An IBM, Microsoft and VeriSign joint specification. April 5, 2002. http://www-106.ibm.com/developerworks/webservices/library/ws-secure/

14. Liberty Alliance Project Specification Archive V.1.1 –http://www.projectliberty.org/specs/archive/v1_1/index.html

About Authors



Mr. S K Sharma is Scientist-B, leader of Networking, testing and Quality control group in Information and Library Network Centre, UGC-INFLIBNET Ahmedabad. He has a M.Sc. in Physics and Master of Computer Application (MCA). He has nearly 7 years rich experience in IT and published papers in national conference/journals. Currently he is doing research in Web Services Security. His major research interests are distributed computing, Network security and management.
E-mail : sksharma@inflibent.ac.in



Dr. G. K. Sharma is professor and Head of Information Technology Group at Indian Institute of Information Technology & Management (IIITM), Gwalior. He has more than 20 years of work experience in both the research and academic world. Prior to IIITM, he was Professor & Head of the Department of Computer Science & Engineering at Thapar Institute of Engineering & Technology (TIET), Patiala. Dr. Sharma contributed more than 27 publications at the national and international levels. He did his Master's and Ph.D. in Electronics & Computer Engineering from University of Roorkee (now, Indian Institute of Technology), Roorkee.His area of interest is Distributed Computing.



Prof. P. N. Srivastava is Head of Dept. Mathematical Science and Computer Application and Director of Institute of Information Technology, Bundelkhand University, Jhansi. Dr. Srivastava has 40 years of rich teaching and 35 years research experience. He has guided 15 research scholars and contributed more than 50 research papers at National and International Level. His research areas are special functions, operations research, cryptography etc.
E-mail : pn_shrivastava@yahoo.com

.

Legal Text Retrieval and Information Services in Digital Era

Raj Kumar Bhardwaj

Abstract

India being a huge country by population has lots of court judgments available on print format to various law libraries of the country. Keeping in view the user demand into the library, this article touches the need for computerized legal text retrieval and information services for providing effective services to the users. Part of the discussion in this paper is the process of legal text retrieval that includes search file, search strategies and legal metadata, where search strategies include constructing a search request with the use of boolean operators. Problems of synonyms and homonyms are also covered in this article and in the end the legal information services in modern era are discussed which include online and offline databases.

Keywords : Legal text retrieval, Legal metadata, Information services

0. Introduction

Information play a significant role on all walks of life and lawyers are no exemption. A lawyer's success is more depends on the latest information about his case. He/she has to keep in touch with latest information. Information explosion in the field of law necessitates computerization of library and information centers so that legal professionals can get specific information expeditiously about a case.

1. Need of Computerized Legal Information System

- In a typical situation, a lawyer may be ignorant of the detailed applicable rules and these rules have to be dug out in some way according to legal method.
- Interpretation of statutory provision may yield one or more rules and one rule may be typically based on more than one legal source and lawyers have therefore to identify the relevant legal sources which when interpreted yield the applicable rules.
- Lawyers cannot cope up with sequentially reading all possible text.
- To handle multiple cases of various areas.

2. Process of Legal Text Retrieval

It provides the framework for text retrieval; helps together with retrieval tool, search request, problem identification, secondary search request and antecedent/ consequent.

3. Problem With Legal text Retrieval

The fundamental problem derives from the combination of free-text and Boolean processing as a means of retrieval. Usually all the words in the full text or abstract are indexed, given the open text nature of law. This means that while relevant documents may be returned on a keywords search, they shall be subsumed in a wealth of irrelevant material too i.e. levels of recall and more importantly relevance are deceptively low and the users must sift through these to get useful materials. Given that the volume of legal method

is rapidly increasing, the number of random and meaningless association made on a keywords search is likely to increase (despite moves towards allowing users to limit their search to a particular legal area), thus exacerbating this problems.

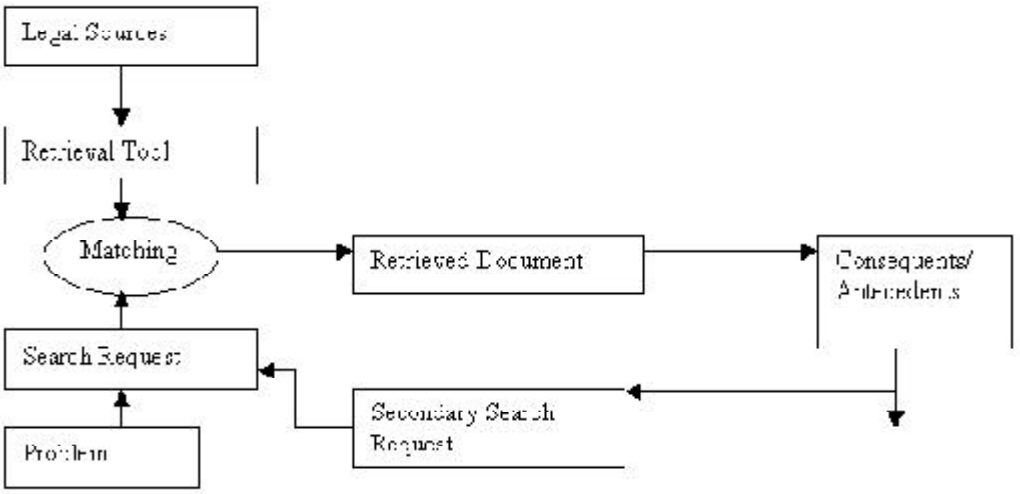


Figure 1. Legal text retrieval process

3.1 Text file

Text file is a copy of the input text; exactly it fed into the program, the program will however process the text for at least two purposes;

a. It will strip the text of codes which may be residue of a word processing publishing system and which governs layout and similar properties of the text.

b. Text Retrieval explicit it will identify the beginning of a document, a word, a sentence and a paragraph. And file indexed in such a way that program easily may retrieve a document by its address and display the text on the screen.

3.2 Search File

Search file establish the program that will discard a small number of frequently occurring words, typically conjunction, pronouns, adverbs, words like and well it and so on, a large total number of words in any text is made up of those commonly used words the function will vary with the language but typically, fixed field can retrieve document easily; it defined as records in any conventional database system and these field are as under:

- Party Name
- Judge Name
- Date of Judgment
- Subject

- ✍ Court
- ✍ Lawyers Name
- ✍ Evidence Persons
- ✍ Place of Parties
- ✍ Citation of Statutes
- ✍ Case CR No / RSA No
- ✍ Journals Bibliographical Detail

3.3 Search strategies

Search strategies we construct for better search results but before its construction we must have some prerequisites, which include:

a. Problems Partition

b. Set of concepts with understanding of Synonyms and Homonyms.

c. Proper Use of Boolean Curse

d. Spelling

First we should understand the problem and partition of it, and is very important because its partition can close to the solution.

Secondly, set of concepts and synonyms understanding. We sometimes make up few synonyms of a problem for a single idea. An example often cited is danger weapon. An idea encountered in the context of criminal law where provision of several jurisdiction have increased penalties if the criminal act is carried out using a dangerous weapon. In the context of a problem the idea may be expressed by obvious example gun or knife. Synonyms comes in several categories, a common diction is between is context independent and context sensitive synonyms.

4. Legal Information Services

Presently more and more legal firms are launching their websites to provide legal information services to lawyers so that lawyers can get maximum information about a case. Following are some examples for some famous legal reference tools;

4.1 manupatra.com

It has large number of online databases, the sets provides comprehensive information on wide range of issues:

a. 1000 forms

b. 1000 central acts

c. Stamp duty act of 9 states

d. Court Rules

e. Cause list

f. Court fees

g. Tribunals

h. Court Calendars

i. Stamp Duty

j. Notification and circulars

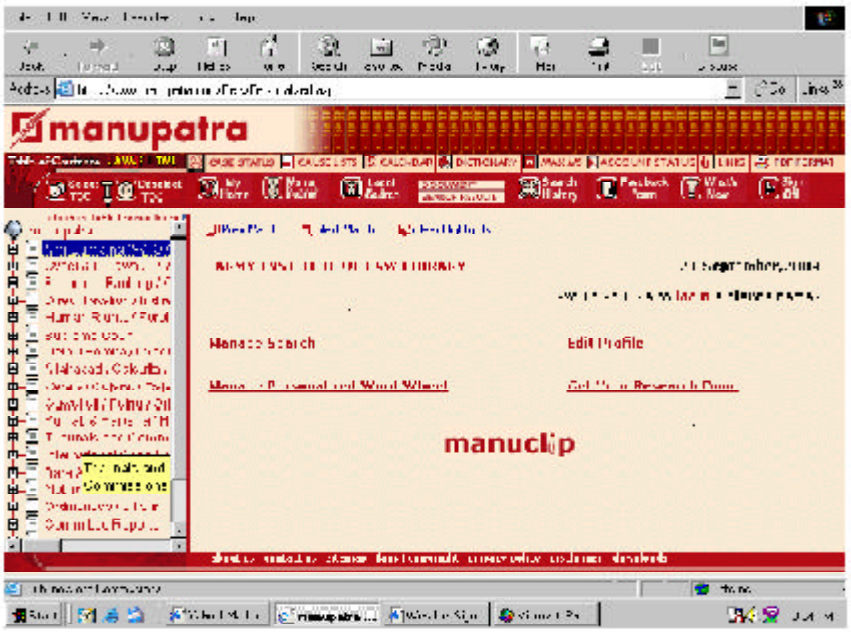


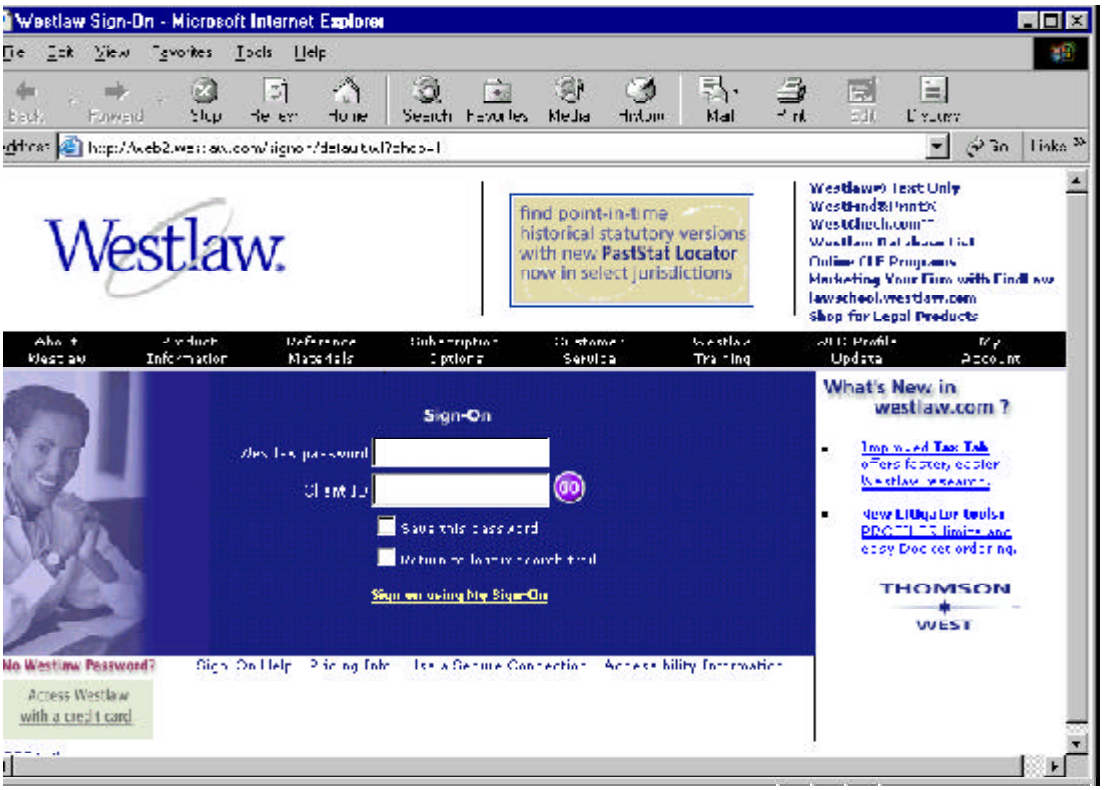
Figure 2. Manupatra a legal reference tool

4.2 SC Case finder (Offline Database)

This database provides comprehensive details of Supreme Court (SC) Cases from 1950 to till date, with search facility - first search, word assist, browsing through search request, going through the each result, parts of case notes, skipping to any case note in the result list at random, rearranging the search result in different order, search with in search, search through topic statutes, case name approach, print the case, adding your own annotation, advance search.

4.3 West law

This site gives comprehensive details of legal research service, instant access to statutes, new business information, public record, form, career at west and other information of federal circuit courts.

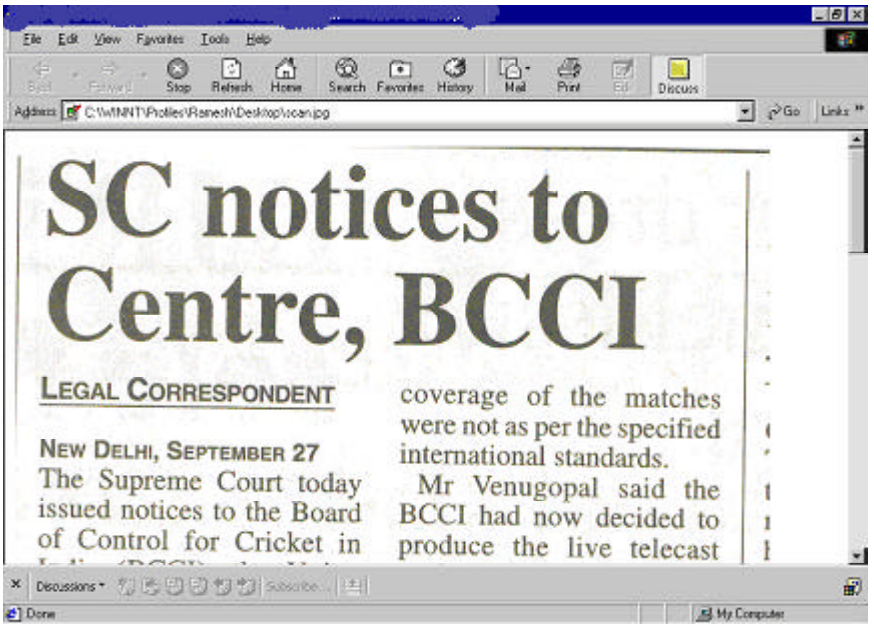


5. In-house Database

In-house database can be developed using WINISIS software, fields for FDT are mentioned above in the legal metadata, this type of database are very useful for users of law library.

5.1 News clipping Service

Newspapers clippings are very much useful for legal professionals, and in library we can develop a Newspaper Clipping Database for legal information service with WINISIS software. One example of this type of database has given below:



SC notices to Centre, BCCI

LEGAL CORRESPONDENT

NEW DELHI, SEPTEMBER 27

The Supreme Court today issued notices to the Board of Control for Cricket in

coverage of the matches were not as per the specified international standards.

Mr Venugopal said the BCCI had now decided to produce the live telecast

Fig. 3. Full text of news

6. Conclusion

Information is the lifeblood of a knowledge-based economy. By collecting data from many different sources and translating them into meaningful information, databases are indispensable to legal professionals, business people, scientists, scholars and consumers. Most recent legal information is published in electronic databases. New information technologies and electronic communication facilities provide opportunities for libraries to play an even more prominent role in the support of teaching, learning and research than before. The use of information technology in libraries will helps in processing and providing the legal information to the users. It is becoming clear that the future in providing legal information will be in electronic format. Therefore, it is critically important to establish from the outset clear standards for publication over the Internet.

7. References

1. J.Bing and T. Harvold. Legal Decision and Information System, Norwegian University Press , Oslo, 1977

2. Ibid pp185-186

3. Bing,J. Performance of legal text retrieval system: curse of boole , Law Library Journal, 79, 187-202

4. Allen, Layman/Saxon, Charles (1985) Computer-aided normalization and unpacking: Some interesting machine process able transformations of legal rules; Charles Walter (ed)Computer Power and Legal Reasoning. West, St Paul 1985:495-572.

5. Attar, R/Fraenkel, AS (1980) “Experiments in local metrical feedback in full-text retrieval systems”: In~om~alion Processing & Management. I 15-126.

6. http:/www.manupatra.com

7. http:/www.westlaw.com

8. SC Case finder manual

About Author

Raj Kumar Bhardwaj is working as Librarian in the S.D College (Lahore) Ambala Cantt. Haryana.
E-mail : rkbhardwaj123@yahoo.com

Building the German Digital Library Vascoda: Status Quo and Future Prospects

Tamara Pianos

Abstract

vascoda (www.vascoda.de) is a new portal for scientific information – it is the nucleus of the German digital library. vascoda is a central access point for interdisciplinary searches ranging from humanities and social sciences to medical studies, engineering, and more. Access to all types of documents is made possible: born-digital as well as digitised and print materials can be obtained either free of charge or through pay-per-view options. The service already includes full-texts, link-collections, databases, subject-specific search engines and more.

More than forty German institutions – mostly libraries and other information specialists are working together to offer users an actual one-stop-shop for all scientific information.

The first release of vascoda was realized through the co-operation of the network of subject-based Virtual Libraries sponsored by the DFG (German Research Society) as well as the Information Alliance and the Electronic Journals Library sponsored by the BMBF (Federal Ministry for Education and Research). Together, these institutions make it possible to combine the search for information and the access to full-text documents. Some of the journal articles and other documents can be accessed free of charge from anywhere in the world, the licence information for Germany is provided by the Electronic Journals Library and other partners.

In this paper, the services of vascoda and examples of how to find quality information for different fields of study will be presented. In spring 2005, a number of new services will be provided. This paper will also look into some of the future developments of vascoda.

Keywords *:* Digital Libraries, Portal, Information Services

0. Introduction

A number of libraries and subject information providers as well as learned societies provide a variety of subject portals with high-quality information. Academic researchers often know the important websites for their specific subjects but it is very hard to keep track of all the information that is available through the internet. There is a need for orientation, harmonization, and the possibility of simultaneous searches to enable researchers to find important information quickly. The search for information should also lead directly to the desired document to speed up the process of information retrieval.

In many countries different providers – state funded or commercial - try to offer their academic researchers and students portals or subjects hubs which provide easy access to quality information. Most of these projects specialize in specific subjects or in specific types of media (e.g. journals, Internet resources, e-publications etc.). vascoda's aim is very ambitious and there are many obstacles along the way, but the vision is to actually provide academic researchers with everything they need from simultaneous searching in many high-quality databases, a ranked presentation of the results, and the possibility of accessing the

desired documents either directly or by ordering them directly through document delivery services like subito (http://www.subito-doc.com/) or interlibrary loan.

Easy access to academic information through an interdisciplinary search combined with easy access to full-texts and navigation to individual subject-specific services is already provided by vascoda. Many problems still need to be solved before vascoda can be an actual one-stop-shop though. One major obstacle is the complicated licence situation in Germany. Due to the federal structure there are no national licences for important databases, so the solutions will have to work with this structure. Centralized solutions are despised in Germany, so the creation of one central point of access and the construction of a national infrastructure are important - yet difficult to achieve - goals. In this respect, vascoda is an important joint-venture of over forty German institutions (libraries, database-providers, etc.), sponsored by the Federal Ministry for Education and Research (BMBF) and the German Research Foundation (DFG). A list of members and associated members can be found at: http://www.vascoda.de/en/wir.html

This paper will present the aims of vascoda and the idea behind it. An overview over the already existing services and a glimpse at future prospects will be given.

1. Aims and Mission of vascoda

“Information overload” poses a problem that many researchers complain about. The information is theoretically available but since so much (mostly useless) information is presented through the internet, most people just do not know where to look for the information that is relevant to them. Also, it is widely believed that almost everything can be found through Google or other search engines. Many people do not know that a lot of high-quality information – even though it may be available free of charge – can not be found with search engines, because it is hidden in the deep web. The so-called STEFI-Studie showed that in the year 2000, 64.1 % of German students used search engines to look for information, while only 5 % used subject specific information gateways. (Klatt, p. 11) Information specialists know that search engines are not an effectual means for a comprehensive search, because the contents of most high-quality databases are not indexed by Google. Highly specialized databases can still offer a better service than a huge index of websites but many people do not know where to find this quality information. Thus, there is need for a service that provides orientation in the face of billions of documents that are available.

Google and other search engines have had a huge impact on how people want to search for information. The search has to be easy, the answers have to come fast, and the service has to be comfortable. Computer and Internet technology offer a wealth of opportunities for the search and delivery of information. By far, not all the possibilities have been used yet. However, there is a vision to provide users with all relevant information as quickly and as easy as possible. The BMBF (Federal Ministry for Education and Research) published its strategy in a position paper “Informationen vernetzen – Wissen aktivieren” (http://www.dl-forum.de/dateien/Strategisches_Positionspapier.pdf) According to these aims, vascoda tries to offer a comprehensive service for end-users who are not information specialists. Users who already use one of the individual portals (which make up vascoda) may use vascoda if they have interdisciplinary query. If, for example, a user wants to have information on “shamanism,” he or she can either search in the subject portal for Ethnology or search in vascoda and discover additional relevant answers from Psychology, Social Science, Economics, Sports Science, Education, History or Medicine. More complex searches will have to be done in the subject portals. Because of the bottleneck-phenomenon, vascoda will never be able to offer searches that are as complex as those that can be done in an individual subject portal.

The most important mission of vascoda is to create a homogeneous point of access to heterogeneous resources and to standardize the services wherever possible to provide the users with comfortable solutions.

This paper gives an overview over the services that are already available and on the plans for the near future and the next few years.

2. Status Quo

vascoda has been online since August 2003; a second release of vascoda was presented at the end of April 2004. At the end of 2004, twenty-one different subjects are included in the meta-search. The meta-search includes a number of specialized databases (e.g. Medline, FIS Bildung, PSYNDEX, SOLIS) and a number of databases which include metadata of quality Internet resources for specific subjects. Library OPACs are already partially included. More OPACs and Online Contents databases will also be included in the near future. Not all relevant content is included in the search yet and not all of the database providers manage to direct the user to the desired copy immediately. A great deal of work still needs to be done to find solutions for the different licence situations for local libraries. Since Germany does not yet have national licences for specific databases, it will be of greatest importance to find an easy-to-use system of providing licence information and the possibility of pay-per-use options for databases which are not available free of charge.

vascoda itself is a free service, but access to some of the partner's services may be restricted by licences. Access to full-texts is already provided in different ways, depending on license situations. The goals that vascoda wants to reach have been set very high, which means that much still needs to be done in order to create a comprehensive service.

3. A Search in vascoda

At the moment, a search in vascoda is a meta-search of a number of different databases.

It is a web-services structure using SOAP based on XML. The web-services structure of vascoda is described in detail in "Einsatz von Web-Services bei vascoda" (Steidl).

A search in vascoda will give a user a number of results from different subject areas. After a successful search the results are sorted by subject/provider.

The future prospects will be discussed in the Future Prospects-section.

4. Licences and Link Resolving: Retrieving the Full-Text

If the results generated by the vascoda-search are freely available Internet resources, the user can click on the link and get the information he or she desires. If, however, the information is an article from a book or a journal, the user will need information on how to access the material. For articles from electronic journals, the Electronic Journals Library (EZB) holds the licence information. It offers a comprehensive service for checking local licence situations. (http://www.bibliothek.uni-regensburg.de/ezeit/?lang=en) The EZB architecture is widely used in Germany, Austria, Switzerland and some other European countries and also by the Library of Congress in Washington. The EZB works with traffic lights which indicate the accessibility of the text. Green lights mean that the article is from a free web-journal and can be accessed from any computer in the world; yellow means that the respective journal is not generally a free journal but that the users' institution has licensed it, so that the user can access the text directly and free of charge; red means that the user does not have the license to access it. So, if the lights are green or yellow, the user can go ahead and read the full-text. This service depends on the delivery of the complete set of metadata to the EZB, so they can lead the user directly to the article. (This is not always the case yet.) It also means that the providers of the full-text will have to offer an open-linking-structure. In the future, a generic link resolver will make it easier to access journals which are freely available on the internet but which do not support an open-linking-structure yet. If the lights for a specific journal are red, the user might use a pay-per-view option, if it is available for the respective text.

vascoda is connected to the EZB via Open URL-technology, so that a user of vascoda can get the information on licences and availability through the EZB. The EZB leads the user to the full-text, via two or three clicks, or in many cases already directly.

The following example shows how a search in vascoda and the delivery of the full-text to the desktop ideally work. A user searches for “malaria” in vascoda and gets the following results:

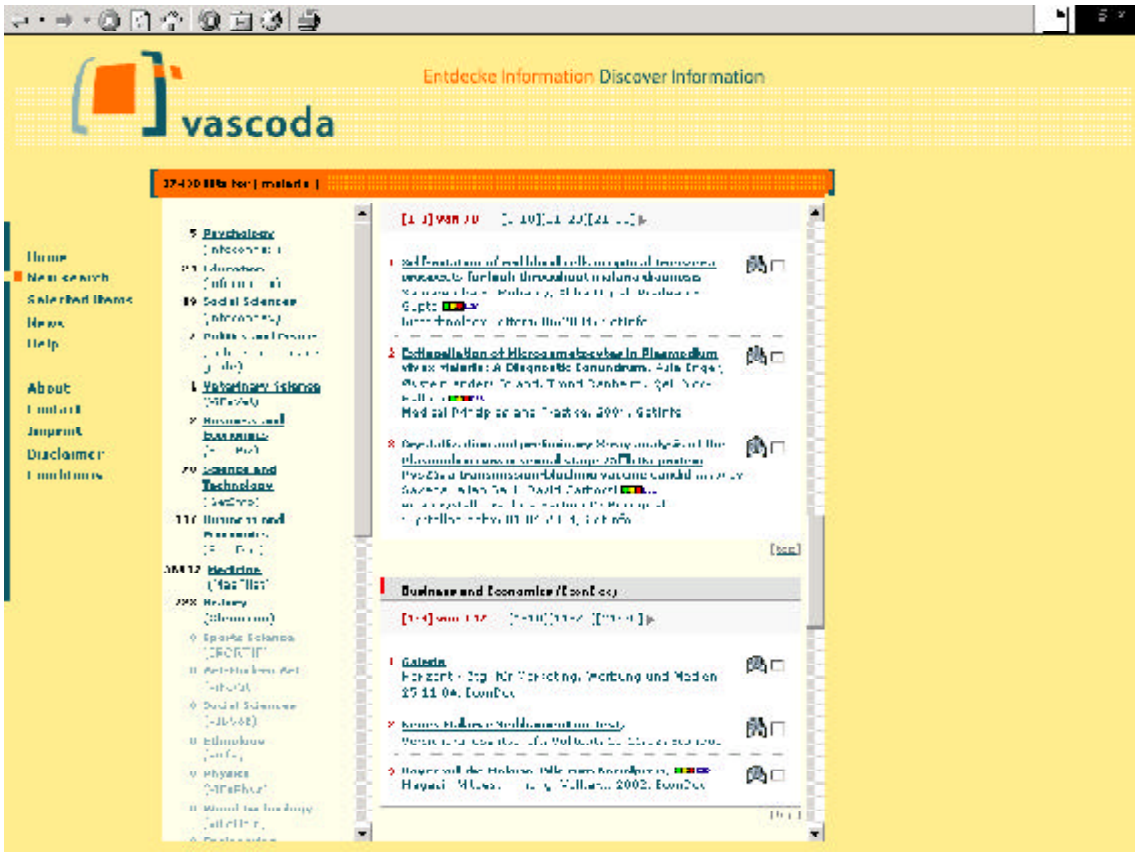


Fig. 1. Search Results for “malaria”

A search for “malaria” generates 37,430 hits from 10 different subject portals. One would expect results from the subject portal for Medicine but there are also useful hits from Psychology, Education and Economics etc. The small traffic lights indicate articles from electronic journals.

A click on the traffic lights-symbol reveals the licence situation at the users institution. Yellow traffic lights can only be presented at institutions which participate in the Electronic Journals Library, since the system needs to know the specific conditions, whereas green lights will be presented for the respective journals everywhere in the world.

If you click on the traffic lights symbol in the list of results, this will lead you directly to the Electronic Journals Library.

For the following search result, the traffic light is yellow, indicating that the user's institution – in this case the German National Library of Science and Technology – holds a licence for this electronic journal. This is why a "view full text"-button can be inserted. A click on the "view full text" button gives the user direct access to the desired full-text. If the full set of metadata is not made available to the EZB or if the publisher of the journal does not support an open linking structure, the user will have to take a few more steps to get the desired text. In this case, the caption on the link will read e.g. "view journal homepage". The user then has to follow the structure on the journal homepage to get to the full-text. This option is not as comfortable as the first one but it is still useful. Meanwhile a number of institutions try to increase the number of full-texts that can be directly linked to. In any case, the idea is to provide users with easy access to relevant full-texts, directly on the user's screen and free of charge wherever possible or through different versions of document delivery and pay-per-view-options wherever necessary.

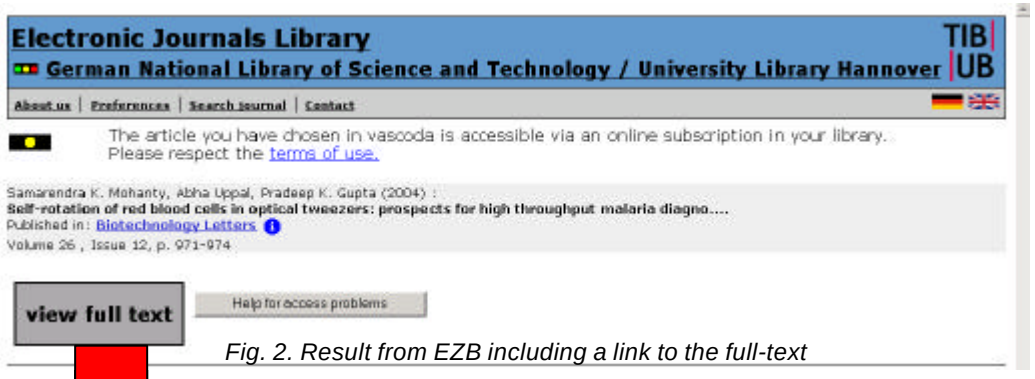


Fig. 2. Result from EZB including a link to the full-text

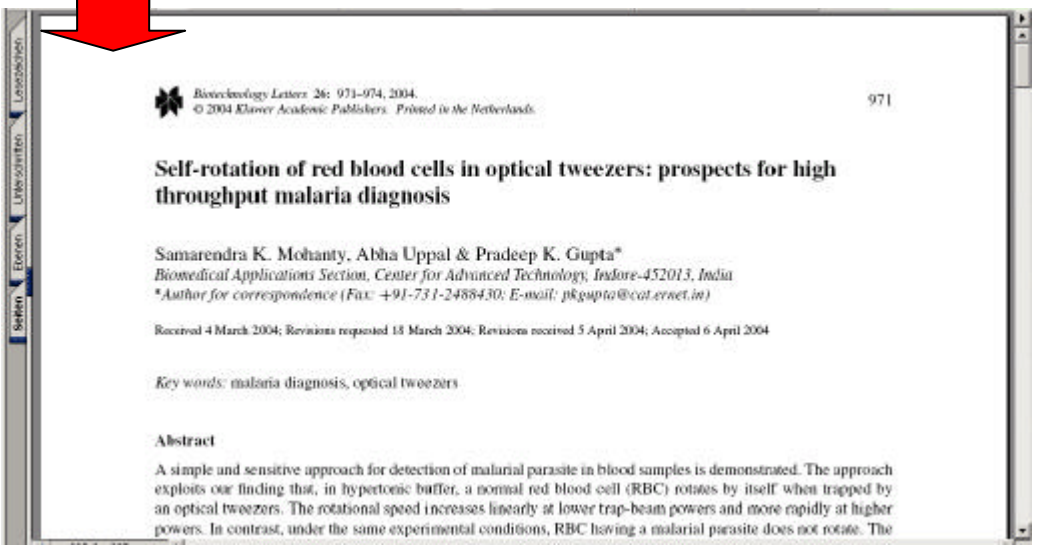


Fig. 3. Full-text from Kluwer Journal Biotechnology Letters delivered directly to screen

5. The Subject Portals

The individual subject portals offer a wealth of information in their respective subject areas. Not all of the databases that can be found in the individual portals are included in the vascoda meta-search yet, but they can be accessed through the subject portals already.

One example for a subject portal is MedPilot (www.medpilot.de), the portal for medicine and related sciences. It allows for a simultaneous search in 35 different databases. More than half of these databases can be used free of charge. The databases that can be searched range from Medline and Cochrane reviews to a number of library catalogues and publisher's journal databases to EMBASE and BIOSIS. There is a pre-selection of 15 freely available databases but the user can choose from the whole range of databases under the expert-search option.



Fig. 4 The expert search of MedPilot (www.medpilot.de) with pre-selected databases

A search for Malaria generates 33,410 hits in total. Many of the results have the EZB-traffic-light-symbol to show the licence situation of the journal.



Fig. 5 Search results in MedPilot with EZB traffic lights for licence situation

After clicking on the green traffic light, the user is directed to the journal's homepage, in some cases even to the full-text and can access the desired full-text without charge from anywhere in the world.

Other subject-based Virtual Libraries offer access to a wealth of subject specific databases and services as well, most of them not yet bundled under a meta-search. Up to now the following 21 subjects are covered: Cultural Anthropology, Earth Sciences, Economic Sciences, Education Science, English Studies, Engineering Sciences, Forestry, History, Mathematics, Medicine, Middle East including North Africa, Modern Art, Natural Sciences and Technology, Pharmacy, Physics, Political Science and Peace Research, Psychology, Social Science, Sports Science, Veterinary Science, Wood-Technology.

One example for a subject portal is the Virtual Library for Political Science ViFaPol.

Through this website the user finds high-quality internet-resources, library catalogues, databases and a number of services like an online-tutorial for Political Science.

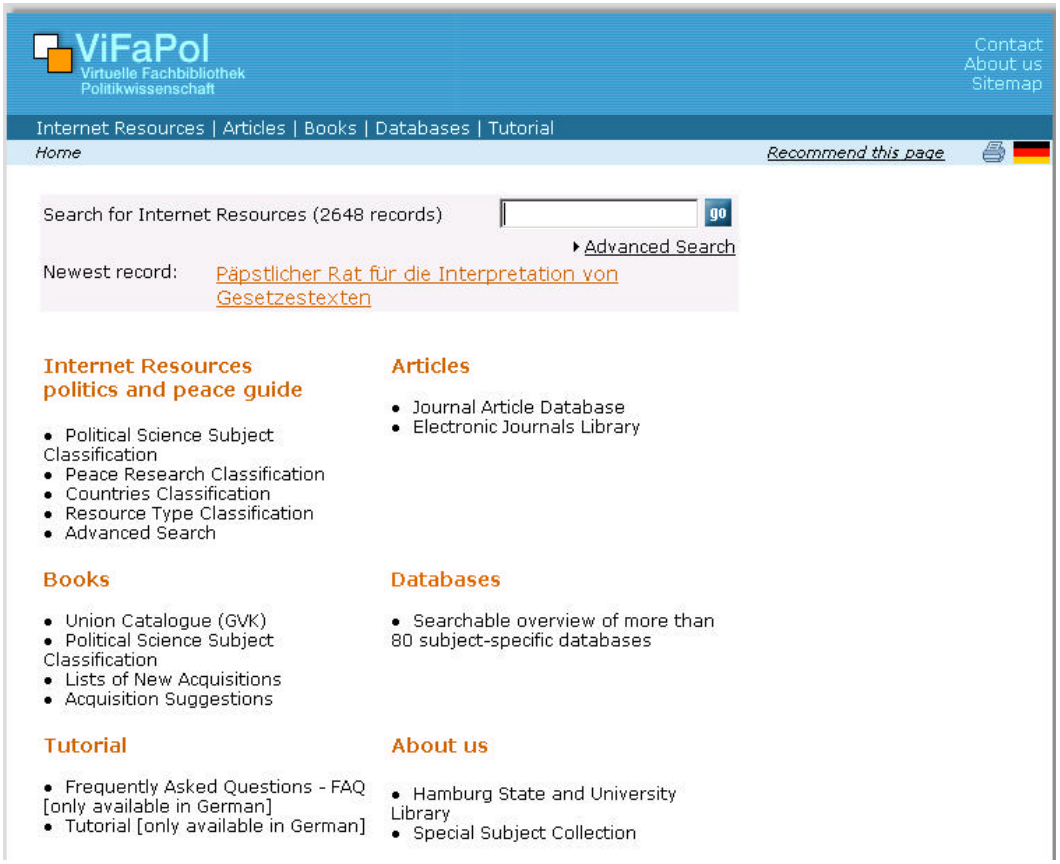


Fig. 6. The ViFaPol and its services (www.vifapol.de)

The individual subject portals give users the opportunity to search for subject specific information – no matter what physical form the information has. It could be genuinely in electronic format, or it could be printed materials that can be found via online-catalogues or databases or even microforms or videos. (Video material is quite an important resource for veterinary sciences for example.)

One of the many useful services that the ViFaPol provides is an overview over databases which are relevant to Political Science. The databases are marked as either available freely on the internet or licensed. If they are licensed the user has to find an institution that has access to the database but if the are freely available, the search can be started directly. For the presentation of the license situation the ViFaPol uses DBIS, a system that works quite similar to the EZB, only that it is used for databases instead of electronic journals. (http://www.bibliothek.uni-regensburg.de/dbinfo/)

6. Future Prospects

The software structure of vascoda will be changed in spring 2005. The search will be based on an architecture provided by the Information Portal Suite (IPS) from April 2005. (http://www.i-portalsuite.de/index_eng.html) The HBZ in Cologne will then be responsible for the operation of vascoda. The HBZ (Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen http://www.hbz-nrw.de/) is the centre of a the regional library network of North Rhine-Westfalia. The IPS-software combined with the know-how and tools that the HBZ has implemented in its own digital library can be used to improve the vascoda services immensely in 2005. At the moment, there are still a number of problems in the vascoda services which need to be dealt with in the next few months. Once the HBZ operates the vascoda service, a number of different services will quickly be implemented, like e.g. an availability check and the inclusion of document delivery through subito and interlibrary loan. In the near future, the user should be able to choose an “appropriate copy” from a range of different options. The German Union Catalogue of Serials (ZDB) – containing more than one million serial titles, including over 400,000 current periodicals and 5.7 million holding records of about 4,300 German and a number of non-German European libraries – is also a part of vascoda. In a joint venture, the EZB, ZDB, and other partners will make it possible to have an availability check and find out in which ways the desired text could be obtained. This could be through direct access, through information on where to find a print-version of the journal, interlibrary loan, document delivery, or pay-per-view.

The holdings of monographs can be found in the local and regional library catalogues. The information that is held by the regional Library Networks will soon be included in the presentation of the vascoda results. This means that interlibrary loans etc. can be initiated after the successful search in vascoda.

Search-engine-technology might also be used to improve the service, e.g. to make the service faster and give more search and service options to the user, to include more content, to implement proper ranking options to sort the results and to implement a search tool that can deal with different spellings of words or typing mistakes.

It will also be important to offer vascoda as a background service that will allow a seamless navigation in the content that is available locally, regionally and nationally, so that users can start their search in the local library portal and that vascoda can offer specific information wherever the local content does not provide this information.

There are a number of libraries in Germany that hold special collections for special subject fields. Webis (http://webis.sub.uni-hamburg.de/ssg/text/was_ist_webis.html#english) offers a good overview over the special collections system in Germany. The content of Webis is supposed to be integrated into vascoda in the near future.

If many of the important goals can be achieved in the next months, many of the obvious bugs that make vascoda sometimes difficult to use can be eliminated by the end of the year 2005. A number of additional services can then be implemented and the already existing ones will be improved steadily, so that by 2007 vascoda will hopefully by a really comprehensive service that provides users with orientation in the face of extremely heterogeneous information providers. By then, vascoda will be a means to get a comprehensive collection of high-quality information from every subject quickly and wherever possible directly to the desktop (or alternatively to the doorstep).

7. References

1. Hutzler, E. Die Elektronische Zeitschriftenbibliothek im Netzwerk Digitaler Bibliothek. (2003). Competence in Content, pp. 381-390, edited by Ralph Schmidt. Frankfurt am Main: DGI.
2. Klatt R. et al. Nutzung elektronischer wissenschaftlicher Information in der Hochschulausbildung: Barrieren und Potenziale der innovativen Mediennutzung im Lernalltag der Hochschulen (Dortmund: 2001). STEFI-Studie, http://www.stefi.de/download/kurzfas.pdf (Accessed on 09/12/2004).
3. Steidl N. Einsatz von Web-Services bei vascoda. Version 1.0, 2003http://www.dl-forum.de/Initiativen/vascoda_Praesentationen/Volltextdokumente/Web_Service_vascoda.pdf (Accessed on 09/12/2004)

About Author



Tamara Pianos studied English Philology and Geography at the University of Kiel. After finishing her dissertation in Canadian Studies, she started her traineeship in Osnabrueck and Cologne to become an academic librarian. In April 2002 she started working as the co-ordinator of the German Subject-Based Virtual Library which by now has become a part of the portal vascoda.

Digital Knowledge Resources for Agribusiness Development

J P S Ahuja M R Rawtani

Abstract

The article attempts to describe the concept of Agribusiness and the role of mapping the appropriate knowledge for Agribusiness Development. Also listed are some of the important digital information sources in Agribusiness with a special reference to the India specific resources.

Keywords: Agribusiness, Directory, Subject Gateway, Agricultural Resources

0. Introduction

Agribusiness, as a concept, was born in Harvard University in 1957, with the publication of a book, A Concept of Agribusiness, under the joint authorship of J. Davis and R. Goldberg. The authors believe agribusiness is the sum total of all operations involved in the manufacture and distribution of farm supplies; production activities on the farm; and the storage, processing and distribution of farm commodities and items made from them.

During the past four decades, this concept has received increased attention. Various definitions of agribusiness have evolved, but they are still based on the original proposal by Davis and Goldberg. In recognizing the interpretation of agribusiness by the two pioneers, however, from systems perspective we should understand that agribusiness also includes related activities in government service provision, rural education and an effective system for knowledge management and its dissemination.

1. Agribusiness: Definitions

Agribusiness in the Anglophone context (e.g. in American universities) is understood as “the production operation of farms, the manufacture and distribution of farm equipment and supplies, and the processing, storage, and distribution of farm commodities.” Here, the term agribusiness is oriented towards the business of agriculture.

Agribusiness in the context of German agricultural administration and science is understood as “the way or mode of managing agricultural enterprises at the production, (input/output) distribution and processing levels”. Here, the term agribusiness is oriented towards agribusiness management, with a bias in favour of microeconomics and entrepreneurship. In accordance with recent development trends, the**GTZ** defines agribusiness as “all market and private business-oriented entities involved in the production, storage, distribution, and processing of agro-based products; in the supply of production inputs; and in the provision of services, such as extension, research etc.” In general this represents a more holistic approach to market-oriented entities in the agro-food system. In any case, the precise orientation is still very much determined by the actual situation in each of our partner countries, and is influenced by a wide range of environmental conditions and their different status. Agribusiness support considers it to be an integral part of a country’s economic development concept, and is targeted towards the creation of jobs and income in mainly rural areas. In line with a common business concept, the guiding principle is always the market orientation of all support activities. Agribusiness is an integral component of rural development, and forms part of a strategy to improve regional economic development and ensure a safe food supply.

- It aims to: address market and private business-oriented entities directly
- Stimulate business opportunities through improved frame conditions in primarily agricultural rural areas
- Ensure a safe and high-quality food supply for the consumer

2. Role of Knowledge for promoting Agribusiness

Agribusiness places increasing emphasis on knowledge as a competitive factor. On the one hand this means an interest in co-operation on fundamental research in pre-competitive phases, but on the other hand more contract research and secrecy in the application of knowledge.

As international forces increasingly make they felt, the national government will have to adopt new positions that regard the reinforcement of national qualities as a beneficial strategy. This includes support for a high quality knowledge infrastructure and provision of world-class research and training opportunities.

Knowledge institutions will be more market-oriented and therefore take a more international approach. This requires important changes in attitude on the part of researchers as well as institutions.

Knowledge (in the broadest sense and not just technology) is treated more and more as a crucial competition factor. ICT will contribute to the rapid development of worldwide networks. At the same time, international competition will intensify and will be felt in the national market as well. There will be an increase in scale and international consolidation of market positions, and investments and activity in local markets will increase all over the world. This means that in the future it will not be a matter of simply going along to local markets and selling products, but of establishing an ongoing presence in those markets; a matter not of disposing of standardized products worldwide, but of benefiting from the differences between all the local markets.

2.1 Challenges for agribusiness, government and knowledge institutions

An international perspective leads to a number of challenges for *agribusiness entrepreneurs*.

2.1.1 Challenges for agribusiness

- using knowledge as the crucial competition factor
- establishing links with sources of knowledge and co-innovators anywhere in the world
- operating in flexible, worldwide networks supported by ICT
- investing, manufacturing and distributing in widespread international markets
- taking advantage of raw material flows shifting to new markets

The challenges facing agribusiness also have a significant influence on strategic positioning of the national knowledge institutions:

2.1.2 Challenges for knowledge institutions

- making knowledge valuable in a competitive international environment
- vigorously developing their function as co-innovators for clients
- ‘brokering and creating links’ for bodies requesting international knowledge

3. New emphases for knowledge policy and knowledge management

Knowledge, knowledge policy as well as knowledge management from important components of strategies which companies, the government and knowledge institutions can use to respond to the changes in operating environment mentioned above.

From this study of globalization and internationalization, three areas of activity are put forward which require strengthening. Firstly, knowledge about changes in the world agro-food market should be thoroughly surveyed. Secondly, international recognition and acknowledgement of the quality of the agro-cluster should be reinforced. In short, the power of companies and knowledge institutions in the agro-cluster to attract interest should be consolidated. Thirdly, knowledge institutions should take advantage of the trend for companies to seek out interesting sources of knowledge and co-innovators worldwide.

This presents knowledge institutions with an enormous challenge: to build up international positions and to make knowledge into a valuable commodity.

These themes for knowledge development and new conditions for the knowledge infrastructure are explained below.

3.1 Themes for knowledge development

a) The world agro-food market: new dimensions and configurations

This theme's key question is how, in global terms, the world market may change in the coming decades under the influence of technological revolutions (including ICT) and economic-political decision-making processes. What do the conceivable changes mean for the various sectors and functions (such as production, trade and distribution) in the context of Indian agribusiness? What are the consequences and possibilities of the flexible worldwide networks now taking shape for the position, function, operation and organisation of agribusiness companies?

3.2 New conditions for the knowledge infrastructure

a) Creating an international professional training centre for agribusiness top management in India

From the point of view of strengthening the international position of the Indian agribusiness knowledge base, it is necessary to assume a leading position in a number of selected and well-defined areas. Apart from the strong ICT work force that already exists, consolidation of knowledge base is required in the field of international enterprise, focusing on agriculture and world food needs.

Agribusiness is the *key player* in the arena of international business. There are also two other important players: the government and the public knowledge institutions. Internationalisation and globalisation will change the environment in which each of these players operates. Moreover there will be both interaction and differences in action resulting from the players' different positions.

4. Digital Knowledge Resources for Agribusiness

Agribusiness knowledge resources can be classified as :

1. Agribusiness sites on the Internet (http://lisweb.curtin.edu.au/web/agbustpts.html" # "agribussites)

2. Agribusiness dictionaries, handbooks, etc.(http://lisweb.curtin.edu.au/web/agbustpts.html" #"agribusreadyref)
3. Electronic journals (http://lisweb.curtin.edu.au/web/agbustpts.html"# "agribusejnls)
4. Annual Reports (http://lisweb.curtin.edu.au/web/agbustpts.html" # "annual reports)
5. Conferences (http://lisweb.curtin.edu.au/web/agbustpts.html" # "agribusconferences) Up and coming conferences, Papers available on the Internet
6. Discussion lists/Newsgroups (http://lisweb.curtin.edu.au/web/agbustpts.html" # "agribuselists) Groups which you can join via electronic mail, Newsgroups (similar to bulletin boards)
7. General Resources (http://lisweb.curtin.edu.au/web/agbustpts.html"#"aggen)
8. Specific Resources(http://lisweb.curtin.edu.au/web/agbustpts.html"# agspec)
9. Agriculture (http://lisweb.curtin.edu.au/web/agbustpts.html"# agriculture
10. Aquaculture (http://lisweb.curtin.edu.au/web/agbustpts.html" # "aquaculture)
11. Crop Protection (http://lisweb.curtin.edu.au/web/agbustpts.html" # "crop protection)
12. Crops http://lisweb.curtin.edu.au/web/agbustpts.html" # "crops)
13. Economics and Finance (http://lisweb.curtin.edu.au/web/agbustpts.html" # economics)
14. Horticulture (http://lisweb.curtin.edu.au/web/agbustpts.html" #"horticulture)
15. Legal Resources (http://lisweb.curtin.edu.au/web/agbustpts.html" # law)
16. Livestock (http://lisweb.curtin.edu.au/web/agbustpts.html" # "livestock)
17. Management (http://lisweb.curtin.edu.au/web/agbustpts.html" # "management)
18. Marketing (http://lisweb.curtin.edu.au/web/agbustpts.html" # "marketing)
19. Viticulture (http://lisweb.curtin.edu.au/web/agbustpts.html" # viticulture)

5. Examaples for Each type of Agribusiness Categories

5.1 General Resources

1. Agriculture@Internets (http://www.internets.com/agri.htm) Huge listing of agricultural websites and databases
2. Agrisurf (http://www.agrisurf.com/agrisurfscripts/agrisurf.asp?index=_25)

5.2 Specific Resources

5.2.1 Agriculture

1. agLINKS (http://www.agpr.com/agpr_htmls/aglinks.html)
2. AgNIC - Agriculture Network Information Cente (http://www.agnic.org/)
3. Agricola (http://www.nal.usda.gov/ag98/) A bibliographic database of citations to the agricultural literature created by the National Agricultural Library and its cooperators.
4. Agrisurf - the Farmers Search Engine (http://www.agrisurf.com/)
5. Agriculture Western Australia (http://www.agric.wa.gov.au/)
6. Agricultural Pests and Feral Animals (http://www.affa.gov.au/content/output.cfm?ObjectID=D2C48F86-BA1A-11A1-A2200060B0A06275)
7. Agrigate (http://www.agrigate.edu.au/) An Agricultural Information Gateway for Australian Researchers.

8. Agripedia. (http://www.ca.uky.edu/agripedia/) College of Agriculture Kentucky University Agricultural Links.
9. AgriWeb (http://www.ruralnet.com.au/AgriWeb/)
10. AgView (http://www.agview.com/)
11. American Society of Agricultural Engineers (http://199.97.51.12/)
12. Australian Centre For International Agricultural Research (http://www.aciar.gov.au/)
13. Australian Collaborative Land Evaluation Program (http://www.cbr.clw.csiro.au/aclep/)
14. Australian Farming Virtual Library (http://farrer.csu.edu.au/AFVL/)
15. Australian Surveying and Land Information Group (http://www.auslig.gov.au/)
16. Canada Agriculture Online (http://www.agcanada.com/)
17. The Farm Shed (http://www.thefarmshed.com.au/index.jhtml)
18. Food & Fertilizer Technology Center. (http://www.fftc.agnet.org/)
19. Natural Resource Management (http://dpie.gov.au/content/output.cfm?ObjectID=3E48F86-AA1A-11A1-B6300060B0AA00011)
20. USDA Economic Research Service.(http://www.ers.usda.gov//)The United States Department of Agriculture official source of economic analysis and information on agriculture in rural America.
21. WWW Virtual Library - Agriculture (http://cipm.ncsu.edu/agvl/)

5.2.2 Aquaculture

1. Australian Aquaculture Centre (http://www.aquaculture.com.au/)
2. Australian Fisheries Management Authority (http://www.afma.gov.au/)
3. Australian Seafood Industry Council (http://www.asic.org.au/)
4. Fisheries and Oceans Canada Pacific Region (http://www.pac.dfo-mpo.gc.ca/)
5. Fisheries Research and Development Corporation (http://www.frdc.com.au/)
6. Fisheries Western Australia (http://www.wa.gov.au/westfish/)
7. Florida Bureau of Seafood and Aquaculture.(http://www.fl-aquaculture.com/)
8. Harbor Branch Oceanographic Institution (http://www.hboi.edu/aquaculture.html)
9. NetVet Aquaculture Sites(http://netvet.wustl.edu/fish.htm" \l "aquaculture/)
10. NOAA Fisheries: U.S Dept. of Commerce National Oceanic and Atmospheric Administration (http://www.nmfs.noaa.gov/)
11. Ozefish Info:Australian Aquaculture and Fisheries Information (http://www.ozefish.com.au/)
12. Purdue University Aquaculture Links (http://www.fnr.purdue.edu/fi/mason/links.html)
13. Tasmanian Salmonid Production (http://www.utas.edu.au/docs/aquaculture/salmon/)
14. World Aquaculture Society (http://www.was.org/)

5.2.3 Crop Protection

1. Biological Control Virtual Information Centre (http://ipmwww.ncsu.edu/biocontrol/) An excellent site providing access to a wide range of resources relevant to researchers,students and professionals.
2. The British Society for Plant Pathology.(http://www.bspp.org.uk/)
3. Entomology on World Wide Web.(http://insects.tamu.edu/entoweb/)

4. Global Crop Pests. (http://www.nysaes.cornell.edu/ent/hortcrops/english/) This Cornell International Institute for Food, Agriculture and Development site provides links regarding crop pest diagnosis and IPM information capability among extensionists and farmers of developing countries.
5. International Survey of Herbicide-Resistant Weeds(http://weedscience.com/)
6. North Carolina National IPM Network - Pest Identification (http://ipmwww.ncsu.edu/PEST_ID/pestid.html)
7. Which Weed? (http://www.tassie.net.au/TVWS.Weeds) Tamar Valley Weed Strategy Working Group.

5.2.4 Crops

1. Agricultural Production Systems Research Unit (http://www.apsru.gov.au/)
2. Alberta Agriculture, Food and Rural Development (http://www.agric.gov.ab.ca/index.html)
3. Australasian Tree Crops Sourcebook on-line (http://www.aoi.com.au/atcros/)
4. Australian Cotton Cooperative Research Centre. (http://www.cotton.pi.csiro.au/)
5. Australian New Crops Project (http://www.newcrops.uq.edu.au/)
6. Canola Information Service (http://www.canolainfo.org/html/links.html) This Canadian site provides links to statistics, processing, biotechnology and organisations.
7. Dilmah Tea. (http://www.dilmahtea.com/plant-pot/plant-pot.htm) This site provides information regarding the processes involved from tea plant to tea pot.
8. Grain Zone - Grains Research and Development Corporation (http://www.grdc.com.au/)
9. International Crops Research Institute for the Semi-Arid Tropics (http://www.cgiar.org/icrisat)
10. International Rice Research Institute (http://www.cgiar.org./irri/)
11. Irrigation sites (http://au.yahoo.com/Science/Agriculture/Crops_and_Soil/Irrigation/)
12. Listing of Useful Plants of the World (http://www.newcrops.uq.edu.au/listing/listingindex.htm)
13. NewCROP (http://www.hort.purdue.edu/newcrop/home)
14. Pulse Australia (http://www.pulseaus.com.au/)
15. Rice Web : a compendium of facts and figures from the world of rice.(http://www.riceweb.org/)
16. The Tea Council Online.(http://www.teacouncil.co.uk/)

5.2.5 Economics and Finance

1. Agribusiness and Agrieconomics (http://www.lib.lsu.edu/bus/agbus.html) Louisiana State University
2. Austrade Online (http://www.austrade.gov.au/)
3. Australian Agriculture, Fisheries and Forestry - Economics(http://www.affa.gov.au/content output.cfm?ObjectID=3E48F86-AB1C-11A1-B6300060B0AA00003)
4. Australian Bureau of Statistics (http://www.abs.gov.au/)
5. Australian Stock Exchange (http://www.asx.com.au/)
6. Australian Dept of Foreign Affairs & Trade (http://www.dfat.gov.au/)
7. The Chicago Board of Trade (http://www.cbot.com/cbot/www/main/0,1394,,00.html)
8. Chicago Mercantile Exchange (http://www.cme.com/)
9. Cotstat and Community Cycle Analytics (http://www.cotstat.com/)
10. FAO Statistics (http://apps.fao.org/lim500/Agri_db.pl) This sites provides international statistical data on crops and livestock primary and processed.

11. FinancialWeb (FINWeb) (http://www.finweb.com/) A financial economics WWW server with links to journals, working papers, databases and other Internet resources.
12. Reserve Bank Bulletins (http://www.rba.gov.au/PublicationsAndResearch/Bulletin/index.html)
13. Sydney Futures Exchange (http://www.sfe.com.au/)
14. Teague Australia Seed and Grain Brokers - Production Statistics - Grain (http://www.tjt.com.au/tjt/statisticsp.html?ptype=Grain)
15. Teague Australia Seed and Grain Brokers - Production Statistics - Seed (http://www.tjt.com.au/tjt/statisticsp.html?ptype=Seed)
16. United States Department of Agriculture Economics Research Service (http://www.econ.ag.gov/)
17. The Universal Currency Converter (http://www.xe.net/currency/)
18. The World Bank Group. (http://www.worldbank.org/)

5.2.6 Horticulture

1. Aggie Horticulture.(http://aggie-horticulture.tamu.edu/) Information Server of the Texas Horticulture Program at Texas A&M University.
2. Apple Information Manager (http://orchard.uvm.edu/aim/default.html) This University of Vermont site includes links to various sites including Pest Management.
3. Australian Citrus Growers Association (http://www.farmwide.com.au/nff/acg/acg.htm)
4. Australian Macadamia Society(http://www.macadamias.org/)
5. Australian Olive Association (http://www.australianolives.com.au/)
6. Bureau of Postharvest Research and Extension - Phillipines (http://www.bphre.com/)
7. Chile Pepper Institute (http://www.chilepepperinstitute.org/)
8. Floriculture.com (http://www.floriculture.com/)
9. Fruits of Warm Climates (www.hort.purdue.edu/newcrop/morton/index.html) The full text of this book, by Julia Morton, originally published in 1987.
10. Horticulture Australia.(http://www.horticulture.com.au/)
11. Horticulture & Crop Science in Virtual Perspective (http://www.hcs.ohio-state.edu/)
12. Plant Breeder's Rights (http://www.dpie.gov.au/content/output.cfm?ObjectID=D2C48F86-BA1A-11A1-A2200060B0A05727)
13. Purdue University Horticultural Web Sites. (http://www.hort.purdue.edu/hort/other)
14. Sydney Postharvest Laboratory. (http://www.postharvest.com.au/)
15. The Ukexnet Horticultural Index (http://www.ukexnet.co.uk/hort/)

5.2.7 Legal Resources

1. Agriculture Law (www.agriculturelaw.com/) A United States site that provides insight to current issues for American farmers.
2. Australasian Legal Information Institute (AUSTLII) (http://www.austlii.edu.au/) Access to Australian primary and secondary legal materials including legislation and cases.
3. Australian Bills Net (http://www.aph.gov.au/parlinfo/billsnet/main.htm)
4. Australian Federal Government Transcripts (http://www.aph.gov.au/hansard/)
5. UniServe Law (http://uniserve.edu.au/law/Welcome.html) An electronic clearinghouse facilitating access to resources for teachers and students in law.
6. SCALEplus (http://scaleplus.law.gov.au) A legal information retrieval system owned by the Australian Attorney General's Dept.

5.2.8 Livestock

1. Animal Health Australia (www.aahc.com.au/) Animal Health Australia - provides statistics and information about animal health in Australia.
2. Agricultural Business Research Institute - BREEDPLAN. (http://abri.une.edu.au/bplan.htm) This is a beef cattle genetic evaluation system. It covers a wide range of traits including birth weight, calving ease, growth, milking ability, fertility and carcase information.
3. AgriOne Cattle Links (http://www.agrione.com/cattle.html)
4. AgriOne Pork Links (http://www.agrione.com/porklinks.html)
5. AgriOne Sheep Links (http://www.agrione.com/sheep.html)
6. The Angus Society of Australia (http://www.angusaustralia.com.au)
7. Association for the Advancement of Animal Breeding and Genetics Inc.(http://agbu.une.edu.au/~aaabg/)
8. Australian Dairy Corporation (http://www.dairycorp.com.au/)
9. Australian Limousin Breeders Homepage (http://www.northnet.com.au/~limo/)
10. Australian Ostrich Association (http://www.aoa.asn.au/general.htm)
11. Australian Poll Hereford Society (http://www.pollhereford.com.au/)
12. Breeders World (http://www.breedersworld.com/) A Livestock Directory with excellent links to Beef, Sheep and Swine Websites.
13. Cattle Council of Australia (http://bioag.byu.edu/zoology/crandall_lab/crayfish/crayhome.htm)
14. Deer Industry Association of Australia (http://www.diaa.org/)
15. Farmwide Cattle Links (http://www.farmwide.com.au/links/search.asp?query=CATTLE)
16. Meat and Livestock Australia (http://www.mla.com.au/)
17. Meat Net (http://www.aginfo.com.au/htms/meatlinx.htm)
18. NSW Meat Industry Authority (http://www.meat.nsw.gov.au/) This site includes saleyard statistics.
19. NetVet Cow Sites (http://netvet.wustl.edu/cows.htm)
20. NetVet Dog Sites (http://netvet.wustl.edu/dogs.htm)
21. NetVet Horse Sites (http://netvet.wustl.edu/horses.htm)
22. NetVet Pig Sites(http://netvet.wustl.edu/pigs.htm)
23. NetVet Poultry Sites (http://netvet.wustl.edu/birds.htm" \l "poultry)
24. NetVet Sheep Sites (http://netvet.wustl.edu/smrum.htm" \l "sheep)
25. NetVet Small Ruminants Sites (http://netvet.wustl.edu/smrum.htm)
26. The Pork Council of Australia (http://www.pca.org.au/)
27. Product Integrity and Chemical Usage (http://www.xtra.com.au/picu)
28. Sheep'o - Australian Sheep and Wool Industries on the Web (www.aussiesheep.com/)
29. Sheep Resources (http://www.ansi.okstate.edu/library/sheep.html)
30. Simmental Australia (http://www.simmental.com.au/home.html)
31. Swine Net (http://www.swine.net/)
32. Wool.com - Everything Wool (http://www.wool.com/)
33. WorldMeat (http://www.worldmeat.com.au/)

5.2.9 Management

1. Australian Institute of Management (http://www.aim.com.au/)

2. Department of Agriculture, Fisheries and Forestry - Operating Environment (http://www.affa.gov.au/taxreform/)
3. Knowledge Inc.(http://www.webcom.com/quantera/)
4. Knowledge Management for the New World of Business (http://www.brint.com/km/whatis.htm/)
5. Knowledge Management. A WWW Virtual Library on Knowledge Management. (http://www.brint.com/km/)
6. Yahoo - Business and Economy (http://www.yahoo.com/Business_and_Economy/)
7. Yahoo-Farm Management ("http://search.yahoo.com/bin/search?p=farm+ management&y=y&e=578870&f=0%3A2766678%3A2718086% 3A159860%3A159869%3A160068%3A578870&r=Regional% 02Countries %02Australia %02Science/)

5.2.10 Marketing

1. Agriculture online - Markets (http://www.agriculture.com/markets/)
2. Australian World Wide Wool Information and Marketing Service (http://www.wool.net.au/)
3. Business Entry Point (http://www.business.gov.au/) This Australian Commonwealth Government site has links to starting a business, drafting business plans, employing staff and taxation.
4. Market Asia (http://www.fintrac.com/rap/)
5. Today's Market Prices (http://www.todaymarket.com/) This sites provides prices of agricultural products from around the world.
6. World Bank Commodity Prices (http://www.worldbank.org/prospects/pinksheets/)

5.2.11 Viticulture

1. Australian National Wine & Grape Industry Centre. (http://www.csu.edu.au/research/rpcgwr/nwgic.htm)
2. Australian Wine Research Institute (http://www.awri.com.au/flash.html)
3. Australian Wine Online (http://www.winetitles.com.au/wineonline.html)
4. AusVit Vineyard Management System.(http://www.csu.edu.au/research/rpcgwr/ausvit/)
5. Co-operative Research Centre for Viticulture (http://www.crcv.com.au/)
6. Eco-Rating International:Eco Survey of California Vineyards and Wineries (http://www.eco-rating.com/sces.html)
7. University of California Davis Department of Viticulture & Enology. Wine and Grape Information (http://wineserver.ucdavis.edu/winegrape/index.htm)
8. Wine of Australia (http://www.wineaustralia.com.au/)

5.3 Agribusiness Dictionaries, Handbooks etc.

1. Australian Financial Review Dictionary of Investment Terms. (http://www.county.com.au/web/webdict.nsf/pages/index)
2. Campbell R. Harvey's Hypertextual Finance Glossary - Duke University (http://www.duke.edu/~charvey/Classes/wpg/glossary.htm)
3. New York Mercantile Exchange Glossary (http://www.nymex.com/refernce/glossary.htm)
4. Ohio State University Plant Dictionary (http://www.hcs.ohio-state.edu/plants.html)
5. Typesetting & Publishing Glossary (http://www.sos.com.au/files/glosray.html)

6. Soil Science Society of America - Soil Glossary (http://www.soils.org/sssagloss/)
7. Wisconsin Department of Natural Resources Glossary of Lake and Water Terms (http:/ www.dnr.state.wi.us/org/water/fhp/lakes/laketerm.htm)

5.4 Agribusiness Electronic Journals

Electronic Journals vary considerably in accessibility, format and scope. Some will only be available by subscription others will be free; some will be full of hyperlinks others will be a replica of the print version; some will only give contents pages. This is still a developing area in the publishing world and uneven quality will be with us for some time. The suggestions below may prove useful in finding appropriate electronic journals in your area of interest.

5.4.1 General Electronic Journal Sites

1. The World-Wide Web Virtual Library: Electronic Journals (http://vlib.org/)
2. Elsevier Science Agricultural and Biological Sciences Home Page (http://www.elsevier.com/ homepage/browse.htt?mode=basic&key=SSAN) This site provides access to information on print and electronic publications.
3. Contents Direct (http://www.elsevier.nl/locate/ContentsDirect) This is a free e-mail alerting service for Elsevier Science Journals.
4. Australian PC Magazine (http://apcmag.com/)
5. Reserve Bank Bulletins (http://www.rba.gov.au/PublicationsAndResearch/Bulletin/index.html")
6. Science Komm (http://www.sciencekomm.at/journals/agric.html) A large number of agribusiness journals can be sourced via this service.

5.4.2 Specific Agribusiness Electronic Journal Sites

1. Agricultural Outlook (http://www.ers.usda.gov/epubs/pdf/agout/ao.htm/)
2. Agricultural Research Magazine (http://www.ars.usda.gov/is/AR/)
3. Agriculture Online (http://www.agriculture.com/)
4. Agricultural Economics (http://www.elsevier.com/homepage/sae/econbase/agecon/menu.sht)
5. Agricultural Economics Virtual Library: Journals & Research (http://www.agecon.com/index.htm)
6. American Journal of Enology and Viticulture (http://www.ajev.com/)
7. Australian Agribusiness Review (http://www.agribusiness.asn.au/review/index.htm)
8. Australian and New Zealand Wine Industry Journal (http://www.winetitles.com.au/wij/index.html)
9. Australian Journal of Grape and Wine Research Online (http://www.asvo.com.au/front_ajgwr.html)
10. Australian National University Department of Economics Working papers in Economics and Econometrics (http://ecocomm.anu.edu.au/economics/misc/working.html)
11. Better Farming (http://www.betterfarming.com/index.htm)
12. BRW (http://www.brw.com.au/)Business Review Weekly.
13. CAE working paper series (http://www.farmlandinfo.org/cae/wp/caewpabs.html) Centre for Agriculture in the Environment
14. Canola Guide MagazineOnline (http://www.agcanada.com/cn/cn.htm)
15. Co-operative Research Centre for Viticulture Newsletter (http://www.crcv.com.au/index_n.html)
16. CSIRO Journals (http://www.publish.csiro.au/journals/samples.cfm)

es

17. Farm Journal Today (http://www.farmjournal.com/)
18. Farmers Weekly Interactive (http://www.fwi.co.uk/)
19. Food and Fertilizer Technology Centre Newsletter (http://www.fftc.agnet.org/library/list/pub/nl.html)
20. Journal of Agribusiness (http://www.agecon.uga.edu/~jab/jabs.htm)
21. Journal of Animal Science. (http://www.asas.org/jas/)
22. Journal of the International Association of Agricultural Economists (http://www.ag.iastate.edu/journals/agecon/jpage/home.html)
23. Olives Australia (http://www.oliveaustralia.aust.com/)
24. Poultry Science Journal (http://www.psa.uiuc.edu/toc.html)
25. Progressive Farmer (http://www.progressivefarmer.com/)
26. RIRDC Research Reports (http://www.rirdc.gov.au/reports/)
27. Wine Business Monthly (http://winebusiness.com/html/monthlycover.cfm

5.5 Annual Reports

A useful way of accessing company information is accessing company annual reports over the Internet.

1. Agriculture Western Australia (http://www.agric.wa.gov.au/segments.asp?cid=453&pid=1&pg=1)
2. Tropical Savannas CRC - Annual Reports (http://savanna.ntu.edu.au/centre/annrep.html)

5.6 Agribusiness Conferences

While there is somewhat of a trend for conferences to be held via the Internet it is most common for conferences to be held in the traditional style. The suggestions below are some of the ways of finding out what conferences are happening. It is now very common for a conference to have a web page giving considerable pre-conference information including pre-prints in some cases.

5.6.1 General Conference Sites

1. Internet Conferences and Events (http://www.loc.gov/global/internet/conference.html)

5.6.2 Specific Conference Sites for Agribusiness

1. Yahoo - Agribusiness Conferences and Trade Shows (http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Agriculture/Conventions_and_Trade_Shows/)
2. Agricultural Conferences, Meetings and Seminars Calendar (http://www.agnic.org/mtg/) The Agricultural Conferences, Meetings, Seminars Calendar (AgCal) provides a central repository for information and links to information concerning agricultural conferences of scientific significance.
3. All Conferences.Net (http://all-conferences.net/)
4. The International Association of Agricultural Economists (IAAE)Upcoming Conferences. (http://www.iaae-agecon.org/conferences.asp)

5.7 Agribusiness Discussion Lists

To learn how to subscribe to Discussion Lists you may want to try a guide first:

1. E-Mail Discussion Groups and Lists - Resources (http://www.webcom.com/impulse/list.html)

5.8 General Directories of Email Discussion Lists

1. Topica: Email newsletters and discussion groups (http://www.liszt.com/)

5.9.1 Some Indian Web-Sites on Agribusiness

1. Agriwatch
   Description: Agricultural and food commodity-oriented market tracking service; offers consulting and subscription...
   http://www.agriwatch.com/
2. **Agribusiness** to **India**
   http://www.austrade.gov.au/ci_template/0,1114,MetaRID%253DPWB193022,00.html
3. **India's** Premier web site on **Agribusiness**
   Your gateway to Indian Agribusiness with online bulletin board for trade enquiries and crop and market information
   http://www.agroindia.org/
4. Indian Society of **Agribusiness** Professionals
   Description: Indian Society of **Agribusiness** Professionals (ISAP) aims to satisfy the needs of the farming community...
   http://www.isapindia.org/
5. Agriculture Gateway for **India** - Economics
   http://web.aces.uiuc.edu/aim/diglib/india/economics.htm
6. Agro Tech - conferences
   These Conferences have been structured as a Global Forum for Food and Agribusiness n India with significant international participation to position India
   http://www.agrotech-india.com/cnfrnce.htm

5.9.2 Online Indian Agribusiness Publications

1. http://www.indiancommodity.com
2. http://www.ikisan.com
3. http://www.krishiworld.com
4. http://www.krishiudyog.com
5. http://www.soyachaupal.com
6. http://www.indiaagronet.com
7. http://www.agroconnect.com

5.10 List of some e-journals on Agribusiness

Title

1. AgExporter - online (ProQuest)
2. Alaska journal of commerce - online (ProQuest)
3. American Economic Association Quarterly - online (JSTOR)
4. American economic review - online (JSTOR)

5. American economist - online (Ebsco Publishing)

These journals are available through the content providers such as Ebsco,ProQuest, Eco, Catchword, JSTOR as indicated above.

6. References

1. agLINKS (http://www.agpr.com/agpr_htmls/aglinks.html)
2. AgNIC - Agriculture Network Information Cente (http://www.agnic.org/)
3. Agricola (http://www.nal.usda.gov/ag98/) A bibliographic database of citations to the agricultural literature created by the National Agricultural Library and its cooperators.
4. Agrisurf - the Farmers Search Engine (http://www.agrisurf.com/)

About Authors

Mr. J P S Ahuja is Manager, Department of Information Systems and Computer Services at NABARD, Head Office, Bandra-Kurla Complex, Mumbai and Research Scholar at the Department of Library and Information science, University of Rajasthan, Jaipur.
E mail : ahujajps@sify.com or jpsahuja@indiatimes.com

Dr. M R Rawtani is Associate Professor and Head of the Department of Library and Information Science, University of Rajas than, Jaipur. He has presented number of papers in seminar, conferences and journals. He is also a member of many professional bodies.
Email :rawtanimr_jaipur@sancharnet.in

Networking and Security Aspects in Libraries

P Balasubramanian K Paulraj
S Kanthimathi

Abstract

Now a day the libraries are made Digital using our modern computers. While implementing these types of tasks there is a keen notification towards the security of the network we are using, which means that the data and information present in our system should not be destructed or damaged by the users as well as the professional hackers. So security in Digital libraries has become a major task. This article explains you the various security aspects in detail.

Keywords : Modern Library, Library Networking, Library security, Information and data security

0. Introduction

The securing of conformations and other technologies has been highly beneficial to libraries. Libraries need to have policies. There should be highly protective and secure systems to be placed in Libraries so that the hackers could not destroy the contents of the Libraries.

The libraries are to be networked so that a distant client of that particular library can access the details of the Library. These can be done by having shared cataloguing, union cataloguing, document delivery services, inter library loan, e-mail, bulletin board, current awareness services and online public access services. Network security is the balanced act for Libraries. The data, which are in the local network, should be distributed throughout the system. The data should be accessed flexibly. The flexibility of assessment has made the system security a weak one. So the system security should be enabled. That is in other words the system should be avoided from the access of unauthorized persons who may hack the local library system. Not only the crackers but there may be butter fingered persons by which a small mistake can destroy or damage the data. The later problem is quite common and it cannot be driven out easily because the data are un intension ally corrupted. The past discussion about these facts tells us that in web enabled in house library and information services; each user should be given a user id and a password to browse the databases and downloading information. As most of the librarians were not much aware of the security problems which may engage the network there are so many causes occurs which may damage the network. Even most of the college courses even not depict the security constraints in libraries. Minimally, effective information security in libraries should include staff assigned to information security tasks, Training all personnel in information security issues and procedures, specific policies dealing with information privacy, physical security of equipment and computer security procedures, physical security plans, level of access of data and monitoring of different types of access.

The above points can be implemented for academic, corporate, special libraries and collections. They can be intended for libraries of all sizes, with all type of patrons, funding models and organized structures. The need of security should be of the same rate whether the data present is huge or small.

1. Information Security

Information security is not the same as that of the computer security. It is a part of the computer security. Computer security relates to securing computer systems against unwanted access and use; information security also includes issues such as information management, information privacy and data integrity. Having a proper security for all the data inside the library or else by having proper backups for all the data can do the information security. The physical integrity should also be well fit for a secure system. It can also be considered as the protection of data against accidental or intentional disclosure to unauthorized persons or unauthorized modifications and destructions. The holding of user name and a password as an important tool for securing data in a networked environment. This may be called to be as the Digital signature. It includes functions such as data integrity, authentication and non-repudiation. The security system is needed for the following reasons:

- Damage should be less. This is that the authorized person's carelessness may lead to the destruction of the whole setup or the destruction of a particular login
- There should be a confidential protection.
- Malicious damage should be prevented. some person who may have the authorization for the whole system may change or modify the performance which may lead to corruption or misleading of the system.

2. Data Security

The information stored in a computer system is more valuable than the equipment itself. The file information may be unique for a library and the information centre that created it. But the commercial hardware and software can be replaced or damaged by outside vendors. It is possible for the data to be either private to a particular user or be capable of being shares among a number of users in a way, which can be flexibly controlled. The operating system should provide various safe guards. They may be

- Granting a particular user permission to access the network
- Allowing the particular user to use it in restricted ways.
- Safety from accidental and malicious users.
- Safety from damages from the information centers.
- Privacy if needed with access by either the owner of the file or a specified user.
- Safety from malfunctioning.

3. Access Rights

All information networks allow establishing access rights to the network information. They differ in their degrees of sophistication and complexity. The access rights are based on two issues: one is keeping the people away from the data, and preventing errors that can occur when several people access the same data. Methods of ensuring data:

There are different types of attacks:

- a professional hacker with destructive mind could cause severe damage to a network.
- To come in lime light for recognition and appraise
- A company's competitor may cause such a attack

- In some cases it is noticed that the internal employees are also responsible for network attacks.

Thus the various attacks include interruption, interception and modification and fabrication. The security aspect can be done in two angles one is by logical access limit and other is the physical factor.

Some networks may posse's logical access to their system by using dial up capabilities, account numbers and other networking arrangements. Unfortunately this arrangement also creates security problems since they expose the system to unauthorized users. These unauthorized users may also be difficult to identify. There are various security measures and processes to be discussed they are as follows:

3.1 User account

The first level of network security is the use of user account to allow only authorized users access to the network. Without an account, an computer user can't log in and there by cannot use the network. Each user will be provided an User ID and a password. The user must enter them when he likes to log in. Some networks uses wildcards in the user ID. The problem here is that all the persons of the particular wild card has the same password. Because of the effectiveness of security system depends on the privacy of user password using wild card accounts defeat the purpose of using separate user accounts.

3.2 Intrusion detection system:

Intrusion detection system (IDS) is those systems, which tells about snooping element around the premises for security holes. It can be considered as proactive defense for damage prevention. It monitors network traffic or hosts access attempts. It notifies suspicious traffic or hosts access attempts. The notification is through alarming like bells. It prevents unauthorized accessing of data through intelligent surveillance, and intrusion and attack detection, inappropriate URL detection and blocking and alerting, logging and real time response.

3.3 Location based Network security solutions

It is a new approach to authentication that utilizes space geodetic methods to form a time dependant location signature that is virtually impossible to forge .It is service that restricts network access to unauthorized users who gain entry only from pre-authorized locations. This solution may be beneficial for those whose network security may benefit from a network access control mechanism featuring constantly varying location signature. Time and location signatures can generate one time security password that could be a helpful addition to network protection toolkits.

This system may avoid the interruption of various anonymous afforded persons. High valued networks can be benefited by this system.

3.4 Switches

A series of modules have been introduced by CISCO that integrate security into its catalyst 6500 series switches. It has a multi layer intelligent storage switching products for used in storage area network (SAN). It has high performance firewall, VPN, secured sockets and network analysis module into the core. These switches deliver a number of storage networking innovations to information customers and

will drive SAN consolidation, increase data availability and allow customer to be more efficiently manage their storage resources.

3.5 Firewalls

While considering a network there are some parts, which can be accessed by everyone and are some parts, which could not be accesses by the foreigners. When there is a huge rise in the demand of remote access there occurs a problem over the above said problem. An object called the firewall can do this. This can be explained as firewall is a security system that selectively denies all access to designated portions of the network, based on how the network is accesses. In library it acts as a separator between the intranet and the Internet.

3.6 Virtual private networks

Security experts say that WLAN can be allowed on the enterprise network if there is a VPN.VPN place authentication and encryption tools on the top of the WLANS. Preventing any unauthorized party from intercepting wireless traffic. It ensures that the person accessing the network is the right person with the proper access levels and provides integrity to a network connection.

4. Conclusion

As we all know that any kind of information is saved in computer for further use and in today's world the use of computer and network has been pervasive for the flow and access of data. Internet is being used for accessing the information from a remote computer system. But security is essential and many a tools and concepts are available with easy solutions. So it becomes necessary to be secure. In the coming time various new featured security means will be provided to secure the information network.

5. References

1. Sharma.T.R.B. Modern Trends in Library Resource Sharing Network. In Herald of Library Science, 33. (1-2) P. 28-34, 1994.
2. Sangam S.L. and Leena. V. Digital Library Services Chennai Caliber Feb 2000
3. Arora (Jagdish), Web Based Digital is Sources and Services: Trends and innovation. Paper Presented at Caliber-2001, held at University of pune during March 15-16, 2001, P.185-212.
4. D-LibMagazine, Amonthly Electronic Journal.(URL:http://www.dlib.org)
5. http://www.reference.com
6. http://www.infotoday.com
7. http://www.library.com
8. http://www.loc.gov.
9. New technique for 500 times faster date speed over Net/ANI, Hindustan Times, 14 Oct, 2002, P.11

About Authors



Mr. K Paulraj presently working as Librarian in P M T College, Melaneelithanallur, Tamil Nadu. He has done MCom, MLIS, PGDLAN and doing PhD from Manaonmanium Sundaranar University, Tirunelveli. He is a life member of various professional associations. He has contributed 11 papers in seminars, conferences & journals.
E-Mail : paulsuresh2005@yahoo.co.in



Mr. P Balasubramanian presently working as Librarian in SCAD College of Engineering & Technology, Cheranmahadevi, Tamil Nadu. He has done MA, MLIS, PGDCA, PGDPR and doing PhD from Manaonmanium Sundaranar University, Tirunelveli. He is a life member of various professional associations. He has contributed 15 papers in seminars, conferences & journals.
E-mail: bala_phd2000@yahoo.co.in



Dr. S Kanthi Mathi is working as Librarian (Senior Grade) at Rani Anna Constituent College for Women, Tirunelveli, Tamil Nadu. She has done MLIS and PhD. She is a Research Guide for MPhil and PhD scholars at Manaonmanium Sundaranar University, Tirunelveli. She is a life member of various professional associations. She has contributed 15 papers in seminars, conferences & journals.

Role of the Library Homepage as a New Platform for Library Services: A Case Study

Hemant Kumar Sahu

Abstract

There is an enormous range of available information in the world. In electronic environment many libraries have created a presence on the Web, but have we really thought about why we want to be there? Should library Web sites be grounded in the past or look forward to the future, or both? This paper has tried to focus on various issues related to the object and importance of the library homepage, how to create, what is requirement etc for library homepage. The role of the library is to select, acquire, organize and make available an appropriate subset of this information. Proving electronic information to its users has become a common feature of many special libraries through their homepages. A case study of IUCAA library's homepage is presented. It discusses how IUCAA Library has taken challenges of the new emerging technologies and increasing demands of its users by adopting electronic information sources and services and how it generates value added electronic information for its users. This paper also describes the important electronic sources in astronomy & astrophysics, which is used in IUCAA library.

Keywords : Library Services, Library Homepage, Portal

0. Introduction

In the Internet era, for a library to be recognized and its services to be made available globally creation of a homepage is indispensable. Internet is bringing sweeping changes in most of our daily activities. It is offering access to news, banking services, business opportunities, mails, educational facilities and technology online at our desktops, laptops, and mobile phones. This has changed the overall ways and means, mode and methods of information dissemination. Here the only mantra is how fast the information can be sent to the end users. Like other fields, in the field of library and information science too it has made a great impact. It has changed the overall concept of libraries, role of librarians and has shown the library professionals how the information can be disseminated to their clientele at minimum cost, effort and time. With the fast growth and easy accessibility to Internet, libraries either now or in the near future have to develop their own homepages to meet the wide information requirements of their clientele. Considering the huge costs involved in developing the necessary infrastructure such as owning and maintaining the web servers, high bandwidth Internet connection, the required software, and the technical know-how many libraries, with little knowledge of HTML, librarians can develop homepages utilizing the services of web space providers like Geocities, Tripod, etc. (Pujar, and Manjunath, 2000)

1. Role/Importance of Library Homepage

It should be clear to all of us that the object of a library Web site is connected to the type of library represented. My own context is a special library, so the mission of special library's homepage is provides library service to scientists/Associates/Visitors etc of the Institutions for their information needs. The research library's homepage can support research in higher education through providing access to Internet research tools and full text databases, e-resources virtual observatory etc. It can support user's information requirement through online full text reserves and other means as OPAC, Online journals, e-archive, databases etc. Special libraries generally need to service their parent organization, and the library homepage will reflect this through focusing almost exclusively on the parent institution's users and visitors. (Stover, Mark 1997)

1.1 Role of librarian/information scientists for library homepage

Librarians seeking to apply their traditional role of selection, organization, and dissemination to the Web environment can use some of the following examples in their work. Selection of information resources can be reflected on the homepage through creating links to other relevant sites as well as creating links to full text electronic resources. In fact, many librarians are beginning to view Web "collection development" as a task equally important to traditional (print-based) collection building. It is in some ways more challenging, given the changing nature of Web resources.

Providing access to information can be reflected on the Web through the following: internal search engines, online reference service, stable links to other Internet sites, access to the online catalog and other databases, basic information about the library (hours, staff, collections, etc.), and timely updates. Perhaps the most important of these is access to the online catalog of the library's local collection(s). While many library Web sites provide a telnet-based connection to their online catalog, a growing number are transitioning to a Web-based interface. A Web-based searchable online catalog is preferable in several respects: it provides a consistent and standardized interface for the user, it avoids the necessity of a helper application on the client side, and (in many cases) it allows more flexibility for the user in manipulating data retrieved from the online catalog (Stover, Mark, 1997).

2. Basic Requirements to Make/Maintain for Library Homepage

To start designing and developing a homepage, one must have the following hardware and software:

1. PC (Minimum Pentium series- I, II, III, IV)
2. Modem
3. Internet connection (Dialup/Leased line/VSAT)
4. Internet browser (Netscape/Internet Explorer)
5. Basic knowledge of HTML

Telephone connection if you have does not have leased line connection

2.1 Creation of Homepage

The creation of homepage/website and its subsequent hosting involves certain major steps such as, signing with the web space provider if you are planning to put your page through web space provider, creating contents, uploading the site by using FTP (File Transfer Protocol) and so on. Each one of these has to be dealt carefully to make the website live and interesting. (Pujar, and Manjunath, 2000)

3. Characteristics & Contents of Library Homepage

Initially one gets puzzled what to put on it. It is better to do some homework before actually creating the contents. This can be done by going through WebPages of other libraries, which are already available on the net. To start with, one can include the following information under four major categories:

A. Information

- ✍ Brief introduction about library
- ✍ Details about library collection

- ✍ Library working hours
- ✍ About library staff
- ✍ List of Current journals
- ✍ Journal Holdings list (Back volumes)
- ✍ Monthly additions of documents

B. Services

C. Databases available etc

D. Links to other sites (with a disclaimer)

To create a homepage with above-mentioned contents, one needs to have an HTML editor. This editor is basically a helper program that lets the developer to manipulate the codes. These editors are popularly called as WYSIWYG editors. The term WYSIWYG stands for 'What You See Is What You Get'. These editors basically hide HTML codes and make creation of HTML pages a simple process just like word processing documents. But there is a limitation with these editors as it is not possible to do everything with HTML. Netscape composer or Front Page Express can also be used to create a HTML page. These come along with Netscape and Internet Explorer respectively. The HTML pages can also be created using MS-WORD. People who are familiar with HTML are recommended to use text editors such Norton Editor, EDIT or NOTEPAD and then insert the HTML commands to get the desired output (Pujar and Manjunath, 2000)

To set up a good website/homepage, one need not be a good designer. One can start with a simple design, without harsh colors and graphics. Usually harsh colors and too much of graphics make the web page unattractive and takes lot of time to get loaded. It is suggested not to put all the information on one page, instead try to make multiple pages and provide links to each other so that visitors can easily move from one page to another. Always it is desirable to name the first page as index.htm or index.html, as most of the web servers treat it as the first page. It is something like an index to the site. At the end of the first page, it is desirable to provide a link to the e-mail address, so those visitors can send their suggestions, comments or query instantly.

4. Electronic Information & User's Need

E-information may be broadly defined as " The information stored in a medium, which requires an electronic device to read /access its contents. Information stored in different electronic media such as CD-ROMs, Floppies, Magnetic Tape, Video, Hard – disk itself of PC which can be retrieved with help of Personal Computer, CD/VIDEO player, OPAC catalogue, online Journals, Archive, databases such as ADS, PROLA, VizieR catalogue etc.

The development of computer technology and the communication technology and mainly INTERNET has revolutionized the information provision process. The users are required to be well informed about the latest technical developments and their information requirement has become very complex. The users require information very quickly and they also need the information in readily usable format. In the fields of Astronomy and astrophysics, amount of information that is generates is tremendous and it is very difficult to cater to all the information needs of our users. Exploring the information technical developments and proving electronic information needs of the users can satisfy the extensive satisfied to a large extent. Handling information electronically has helped in providing fast and efficient information.

4.1 Users services through library homepage

The following services can be provided through well-developed/mentioned library homepage.

1. Information about library, its collection, services to its users.
2. More information can be provided though linking of databases related to their interest for library users.
3. Current Awareness Services
4. OPAC: Online Public Access Catalogue
5. Online / Electronic Journals:
6. Contents page Service to remote users
7. Others web-based services to its users.

5. About IUCAA & Its Library

The Inter University Center for Astronomy and Astrophysics (IUCAA) is an institution set up on 1989 by the University Grants Commission to promote nucleation and growth of active groups in Astronomy and Astrophysics in Indian Universities. IUCAA is a national autonomous institution and aims at being a center for excellence within the University sector for teaching, research and development in Astronomy and Astrophysics. IUCAA is a premier Scientific Institute engaged in promote nucleation and growth of active groups in Astronomy and Astrophysics in Indian Universities as well as engaged in research in frontier areas of Astronomy and Astrophysics. Its holds a prominent position among the top few institutions in the World. The Institute boasts of more than 60 scientists from IUCAA as well as from Indian Universities, who rate among the top scientists in India and many of them are held high esteem worldwide.



Slide No. 01: IUCAA Home page (http://www.iucaa.ernet.in)

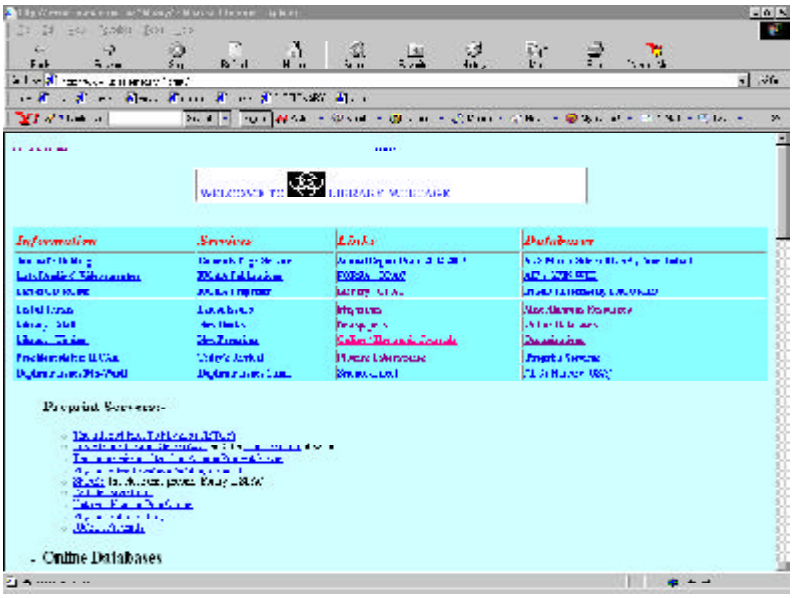
5.1 IUCAA Library

IUCAA library was established in the year 1989, as part of institution with following objects:

- To act as information/ reference center in Astronomy & Astrophysics and allied subjects for researcher and associates/visitors who is coming across the country for their research works.
- To collect, process, store and disseminate information in the field of Astronomy and astrophysics and allied subjects

The IUCAA Library is one of the most advanced modern libraries specializing in astronomy and astrophysics in India. It was the first library in the country to dispense with the card index in favor of a computerized database. Serving as the main resource library in astronomy and astrophysics in the university sector, it is extensively used both on and off campus. In the latter mode it provides references, copies etc. of relevant literature to users from all over India. The object of the library is to support the main objectives and to achieve the goal of parent organization. The books include major collections of astronomy and astrophysics, physics, mathematics, statistics, computer science and electronics, and also a representative collection of books from other branches of science. The library subscribes to important journals in the field of astronomy and astrophysics, many of which are received by airmail. The library is providing excellent service to the in-house faculty as well as to the associates, visitors, students, amateur astronomers, and teachers coming to IUCAA. It is extensively used both on and off campus. In the latter mode it provides references, copies etc. of relevant literature to users from all over India. Contents pages Service being provided to University community by IUCAA Library. The main purpose of IUCAA library is to promote the economical and efficient delivery of information within the university-sector for teaching, research and development in astronomy and astrophysics. It also encourages co-operative efforts for research resources, computing and communications network. Library also strengthens communication and collaboration between research and educational communities. Library takes a national leadership role in the generation and dissemination of knowledge in areas of strategic importance to India in the field of astronomy and astrophysics and also contributes to the lifelong learning opportunities to of all users of the community.

5.2 IUCAA Library Home Page:



Slide No. 2 : IUCCA Library Home Page (http://www.iucca.ernet.in/~library

5.3 Features of IUCAA Library Homepage:

1. Library related information to users as about library, holding of journals, list of thesis, list of audio/ video material, list of available CD-ROMS, list of available duplicate issues of journals, which may required by others library, information of library staff, library timing etc.

2. Services: This part of homepage has been given information about Current Awareness Services, which are given to its users, such as list of today's arrival, list of latest issues available in library, list of IUCAA preprints/Research papers, list of IUCAA publications, List of new books which is recently added in the library, information about contents page services which is given to remote users such as IUCAA Associates/Visitors.

3. Links: In the IUCAA library homepage various types of links has been linked to users as IUCAA OPAC, FORSA OPAC, list of Online journals subscribed by IUCAA library, list of magazines, list of newspapers etc.

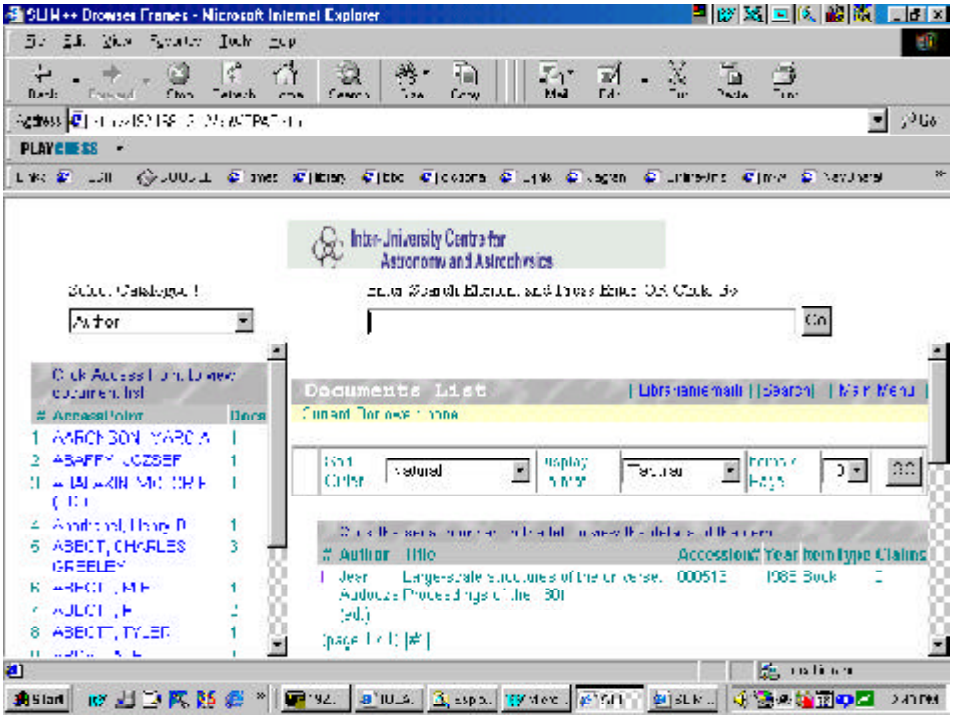
4. Databases: In library homepage various databases also has been links which is frequently used by its users such as: ADS (Astronomical Data System of NASA), AIPs SPIN WEB, e-archive from lan.archive.org, PROLA, Astronomy & Astrophysics Abstracts, Annual Review of A & A, E-print archive, SISSAT.IT Preprint server, SPIRES HE-E-Server, ADS digital library, list of FORSA members with its subscribes journals, science-direct and many others importance archive databases related to A & A.

The homepage of library helps to develop interaction with other libraries. Let to know users about its collection, services, rules, and procedures other libraries such as dealing with same interest as FORSA groups are getting enough interactions among themselves by their homepage.

In the process of providing right information to the right reader at the right time to keep open mind it adapt the new technologies mainly Computers & communications. Libraries have often been among the first departments within an organization to use computers to automate housekeeping activities and were able to use the potential of information technology to access remote databases.

5.4 OPAC: Online Public Access Catalogue

The complete database of the library documents is available electronically. IUCAA library's OPAC can be accessed through INTRANET as well as through INTERNET (http://libibm.iucaa.ernet.in) other than library itself.



Slide no. 03: OPAC at IUCAA Library (http://libibm.iucaa.ernet.in/)

5.5 Online / Electronic Journals

Libraries having online access to electronic journals are suggested to link such journals through their homepage so that the individual user need not remember addresses of all the sites. IUCAA library has subscribes 130 journals. Out of 130, around 100 are available online in the full-text format in addition to the print subscription. IUCAA library's E-journals homepage provides link to about 110 journals, both subscribed and free. Till 1997 there were only traditional print subscriptions. At present around 110 journals can be accessed online. Though the access is through IP authentication for most of the journals, a few of them require password for full-text access. All the subscribed journals are available to the members in IUCAA domain; where as the journal in different IRC's can be accessed based on username and password. The library homepage facilitates the users by giving access to the individual journal homepages, well-maintained links, thus avoiding the hassle of remembering the password or the URL required. Further, some journals can be referred online at four more stations of IUCAA by username & password authentication, namely IRC's Cochin, Raipur, and New Delhi & Darjeeling. The dynamic IP addressing and access has given us the benefit of reaching these electronic journals, from different physical locations, just by subscribing a single copy of the print journal. IUCAA library homepage can access by following site: http://www.iucaa.ernet.in/~library/Online.html (Online/Electronic Journal's home page at IUCAA Library)



Slide no. 04: Online Journal's Homepage at IUCAA Library (http://www.iucaa.ernet.in/~library/Online.html

Library homepage has to be update on regular basis and information has to be given to its valuable users for their update.

6. Conclusion

Users requirement is basically depended on their research areas he/she has selected. User expectations have changed in the context of time, space & effort. Users are beginning to expect document Delivery rather than bibliographic pointers. New information technologies and electronic communication facilities provide opportunities for libraries to play an even more prominent role in the support of teaching, learning and research than before. The object of the library homepage will depend on its parent organization and its clientele. Academic, public, and special libraries will all have different objects, and sometimes-local considerations will impact the nature of a library's object. In any case, library Web site designers must have a clear understanding of the library's mission before embarking on construction of the site. Here an effort has been made to describe how a basic library homepage can be created at minimum cost without investing on web servers and related software. This may not be an exhaustive article where- in one may find every thing on web page development under one roof. Using other programming languages and software such as JAVA, PERL, etc can do future development of homepage.

7. References

1. Inter-University Centre for Astronomy and Astrophysics (2004) http://www.iucaa.ernet.in/ (accessed on 01/10/2004)
2. Kidger, M. et. al., (1999), Internet Resources for Professional Astronomy. In Pro. IX Canary Islands Winter School, Cambridge, Cambridge Universities Press.
3. Library, Inter-University Centre for Astronomy and Astrophysics (2004) http://www.iucaa.ernet.in/~ibrary (accessed on 01/10/2004)
4. Louis, Christina & Vagiswari, A.V. (2000) Information Metamorphosis in Physics and astronomy Libraries. In DRTC Annual Seminar on Electronic, Sources of Information, Bangalore, India 1-3 March 2000.
5. Online Journal Homepage, Library, Inter-University Centre for Astronomy and Astrophysics (2004) http://www.iucaa.ernet.in/~library/Online.html (accessed on 01/10/2004)
6. Pathak, P.J. & Das, Saroj (2000) Utilization of Electronic Information at IPR Library: A case study. In DRTC Annual Seminar on Electronic Source of Information, 1-3 March 2000.
7. Pujar, S. M. and Manjunath, G. K. (2000) Developing a Library Homepage: A Low-Cost Solution; In 'XIX IASLIC' Conference, Bhopal, India. Nov 13-16, 2000.
8. Sahu, H. K. (2003) Online Journals: How To Get Up-To User's Desktop - A Case Study, In Proceedings Volume of 48th Indian Library Association, pages Pp. 63-69.
9. Sahu, H. K. (2004) Use and access of Electronic Information at IUCAA Library: a case study, In SIS Proceedings volume, IIT Madras, India. 21-23 January 2004, Pp. 58-70.
10. Sarah Stevens-Rayburn, and Ellen Bouton (1998), if it's not on the Web, it doesn't exist at all: Electronic Information resources – Myth and Reality. In Pro. Astro. Soc. Pacific Con., Vol. 153, p. 195.
11. Stover, Mark (1997); The Mission and Role of the Library Web Site (http://www.library.ucsb.edu/universe/stover.html) (Accessed on 01/10/2004)

About Author



Mr. Hemant Kumar Sahu, holds B Sc, M L I Sc and doing PhD from University of Pune. He is working as Scientific/Technical Assistant –III at Inter-University Centre for Astronomy and Astrophysics, Pune, since April 1997, before joining IUCAA, Pune, He has taken training in Centre for Advanced Technology, Indore. He has to his credit about 7-8 articles that are published in international and national conference / seminar proceedings.
E-mail: hksahu@iucaa.ernet.in, URL: http://www.iucaa.ernet.in/~hksahu

Wireless Network Connections Policies & Standards

Atul M. Gonsai N N Jani
Nilesh N Soni

Abstract

This paper presents wireless networking scenario with standards & policies for its implementation. Wireless LAN is needed requirement for the organizations & institutions for unlimited access to their wired LAN. This WLAN set up provides Internet access too. Internet & WLAN access needs security to protect data. The paper also discusses security aspect of WLAN; cost consideration, different types of network setup for different need. The paper's focus is WLAN & Internet connectivity & its implementation. The benefited group for this paper is network administrators & management people of any organization.

Keyword : Wireless Network, Access Point, Wireless Security, Wireless Adapter Card, Network

0. Introduction

The term wireless networking refers to technology that enables two or more computers to communicate using standard network protocols, but without network cabling. Strictly speaking, any technology that does this could be called wireless networking. The current buzzword however generally refers to wireless LANs.[1] This technology, fuelled by the emergence of cross-vendor industry standards such as IEEE 802.11, has produced a number of affordable wireless solutions that are growing in popularity with business and schools as well as sophisticated applications where network wiring is impossible, such as in warehousing or point-of-sale handheld equipment.

1. Components of Wireless Network

There are two kinds of wireless networks:

1. An ad-hoc or peer-to-peer wireless network consists of a number of computers each equipped with a wireless networking interface card. Each computer can communicate directly with all of the other wireless enabled computers. They can share files and printers this way, but may not be able to access wired LAN resources, unless one of the computers acts as a bridge to the wired LAN using special software. This is called "bridging". [2]



Figure 1: Ad-Hoc or Peer-to Peer Networking.

Each computer with a wireless interface can communicate directly with all of the others.

2. A wireless network can also use an access point, or base station. In this type of network the access point acts like a hub, providing connectivity for the wireless computers. It can connect (or "bridge") the wireless LAN to a wired LAN, allowing wireless computer access to LAN resources, such as file servers or existing Internet Connectivity.

There are two types of access points:

- Dedicated hardware access points (HAP) such as Lucent's Wave LAN, Apple's Airport Base Station or Web Gear's Aviator PRO. The Figure 2 shows Hardware access points offer comprehensive support of most wireless features.

  Wireless connected computers using a Hardware Access Point.

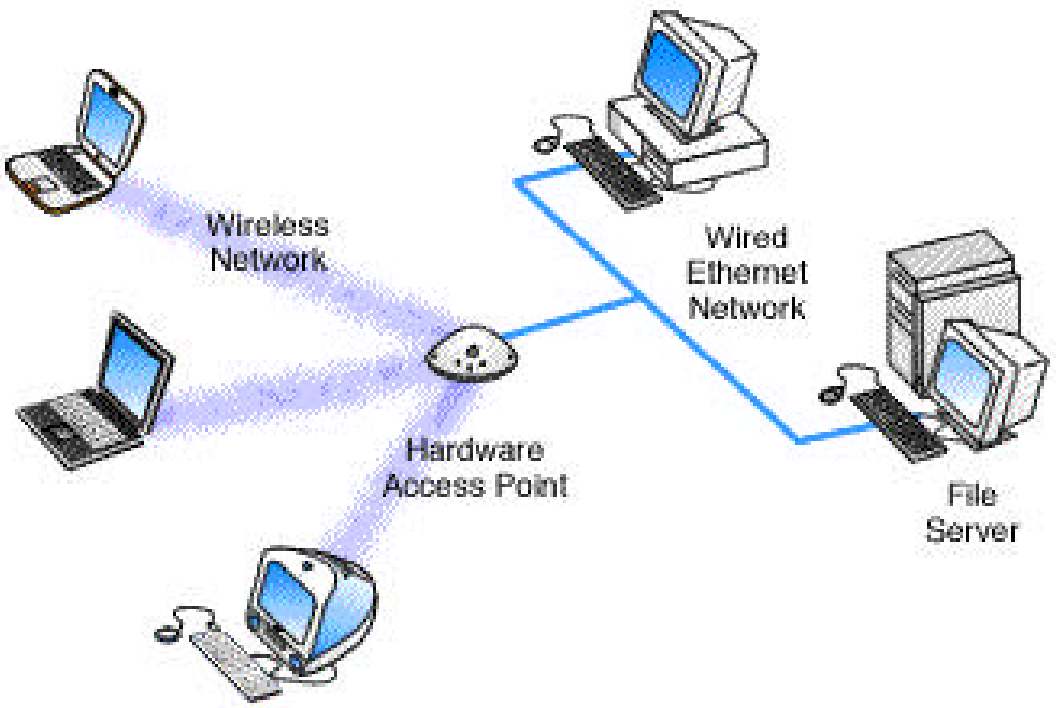


Figure 2: Hardware Access Point.

- Software Access Points which run on a computer equipped with a wireless network interface card as used in an ad-hoc or peer-to-peer wireless network which is shown in Figure 3. The Vicomsoft Internet Gateway suites are software routers that can be used as a basic Software Access Point, and include features not commonly found in hardware solutions, such as Direct PPPoE support and extensive configuration flexibility, but may not offer the full range of wireless features defined in the 802.11 standard. With appropriate networking software support, users on the wireless LAN can share files and printers located on the wired LAN and vice versa.

Wireless connected computers using a Software Access Point.

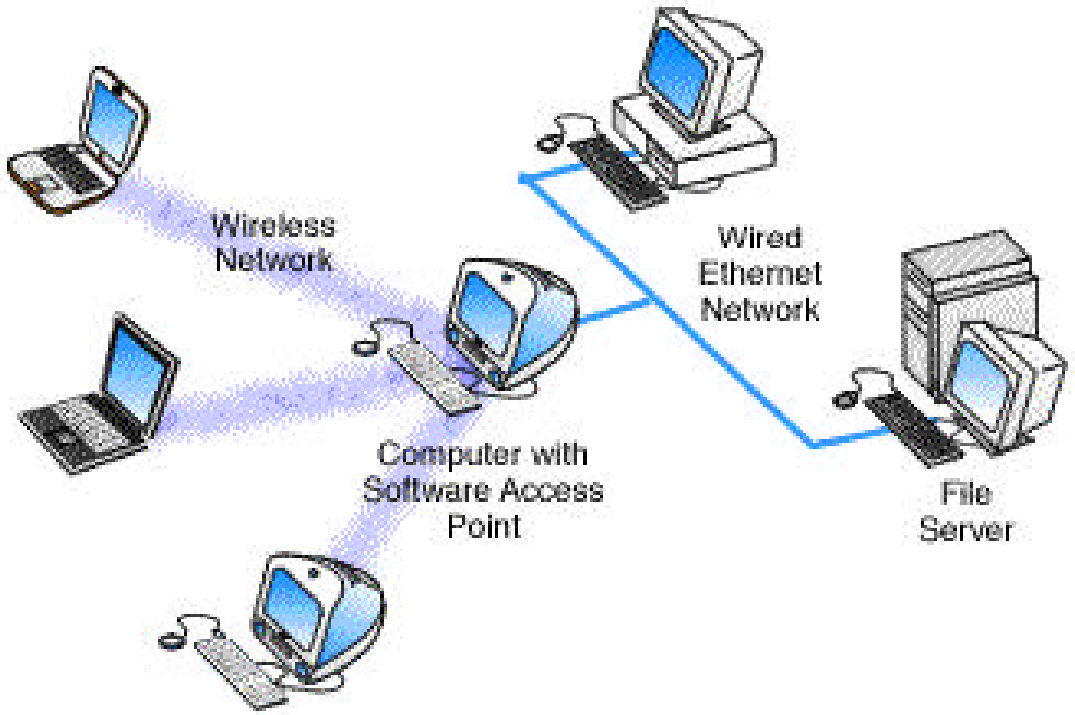


Figure 3: Software Access Point.

2. IEEE 802.11 Wireless standards

Wireless networking hardware requires the use of underlying technology that deals with radio frequencies as well as data transmission. The most widely used standard is 802.11 produced by the Institute of Electrical and Electronic Engineers (IEEE). [5] This is a standard defining all aspects of Radio Frequency Wireless networking. [6]

The following standards are approved:

Standard	Description	Approved
IEEE 802.11	Standard for WLAN operations at data rates up to 2 Mbps in the 2.4-GHz ISM (Industrial, Scientific and Medical) band.	July 1997.
IEEE 802.11a	Standard for WLAN operations at data rates up to 54 Mbps in the 5-GHz Unlicensed National Information Infrastructure (UNII) band.	Sept 1999. End-user products began shipping in early 2002
IEEE 802.11b	Standard for WLAN operations at data rates up to 11 Mbps in the 2.4-GHz ISM (Industrial, Scientific and Medical) band.	Sept 1999. End-user products began shipping in early 2000

The following standards are in draft or conditional approval stage.

Standard	Description	Current Status
IEEE 802.11g	High-rate extension to 802.11b allowing for data rates up to 54 Mbps in the 2.4-GHz ISM band.	Draft standard adopted Nov 2001. Full ratification expected late 2002 or early 2003.
IEEE 802.15.1	Wireless Personal Area Network standard based on the Bluetooth™ specification, operating in the 2.4-GHz ISM band.	802.15.1-2002 conditionally approved on March 21, 2002.

The following standards are still in development, i.e., in the task group (TG) stage.

Task Group	Project Scope
IEEE 802.11e	Enhance the 802.11 Medium Access Control (MAC) to improve and manage Quality of Service, provide classes of service, and enhanced security and authentication mechanisms. These enhancements should provide the quality required for services such as IP telephony and video streaming.
IEEE 802.11f	Develop recommended practices for an Inter-Access Point Protocol (IAPP) which provides the necessary capabilities to achieve multi-vendor Access Point interoperability across a Distribution System supporting IEEE P802.11 Wireless LAN Links.
IEEE 802.11h	Enhance the 802.11 Medium Access Control (MAC) standard and 802.11a High Speed Physical Layer (PHY) in the 5GHz band. Objective is to make IEEE 802.11a/h products compliant with European regulatory requirements
IEEE 802.11i	Enhance the 802.11 Medium Access Control (MAC) to enhance security and authentication mechanisms
IEEE 802.15 TG2	Developing Recommended Practices to facilitate coexistence of Wireless Personal Area Networks™ (802.15) and Wireless Local Area Networks (802.11).
IEEE 802.15 TG3	Draft and publish a new standard for high-rate (20Mbit/s or greater) WPANs™.
IEEE 802.15 TG4	Investigate a low data rate WPAN solution with multi-month to multi-year battery life and very low complexity/

The following standards are in draft or conditional approval stage.

Standard	Description	Current Status
IEEE 802.11g	High-rate extension to 802.11b allowing for data rates up to 54 Mbps in the 2.4-GHz ISM band.	Draft standard adopted Nov 2001. Full ratification expected late 2002 or early 2003.
IEEE 802.15.1	Wireless Personal Area Network standard based on the Bluetooth™ specification, operating in the 2.4-GHz ISM band.	802.15.1-2002 conditionally approved on March 21, 2002.

The following standards are in still development, i.e. in the task group (tg) stage.

Task Group	Project Scope
IEEE 802.11e	Enhance the 802.11 Medium Access Control (MAC) to improve and manage Quality of Service, provide classes of service, and enhanced security and authentication mechanisms. These enhancements should provide the quality required for services such as IP telephony and video streaming.
IEEE 802.11f	Develop recommended practices for an Inter-Access Point Protocol (IAPP) which provides the necessary capabilities to achieve multi-vendor Access Point interoperability across a Distribution System supporting IEEE P802.11 Wireless LAN Links.
IEEE 802.11h	Enhance the 802.11 Medium Access Control (MAC) standard and 802.11a High Speed Physical Layer (PHY) in the 5GHz band. Objective is to make IEEE 802.11ah products compliant with European regulatory requirements.
IEEE 802.11i	Enhance the 802.11 Medium Access Control (MAC) to enhance security and authentication mechanisms
IEEE 802.15 TG2	Developing Recommended Practices to facilitate coexistence of Wireless Personal Area Networks™ (802.15) and Wireless Local Area Networks (802.11).
IEEE 802.15 TG3	Draft and publish a new standard for high-rate (20Mbit/s or greater) WPANs™.
IEEE 802.15 TG4	Investigate a low data rate WPAN solution with multi-month to multi-year battery life and very low complexity/

3. Interconnecting Wireless LAN with wired LAN

To do this we will need some sort of bridge between the wireless and wired network. This can be accomplished either with a hardware access point or a software access point. Hardware access points are available with various types of network interfaces, such as Ethernet or Token Ring, but typically require extra hardware to be purchased if our networking requirements change.

If networking requirements go beyond just interconnecting a wired network to a small wireless network, a software access point may be the best solution. A software access point does not limit the type or number of network interfaces you use. It may also allow considerable flexibility in providing access to different network types, such as different types of Ethernet, Wireless and Token Ring networks. Such connections are only limited by the number of slots or interfaces in the computer used for this task. [4]

4. Access points in wireless network

This depends upon the manufacturer. Some hardware access points have a recommended limit of 10, with other more expensive access points supporting up to 100 wireless connections. Using more computers than recommended will cause performance and reliability to suffer.

Software access points may also impose user limitations, but this depends upon the specific software, and the host computer's ability to process the required information.

4.1 More than one access point

Multiple access points can be connected to a wired LAN, or sometimes even to a second wireless LAN if the access point supports this. In most cases, separate access points are interconnected via a wired LAN, providing wireless connectivity in specific areas such as offices or classrooms, but connected to a main wired LAN for access to network resources, such as file servers as shown in Figure 4.

Wireless connected computers using Multiple Access Points.

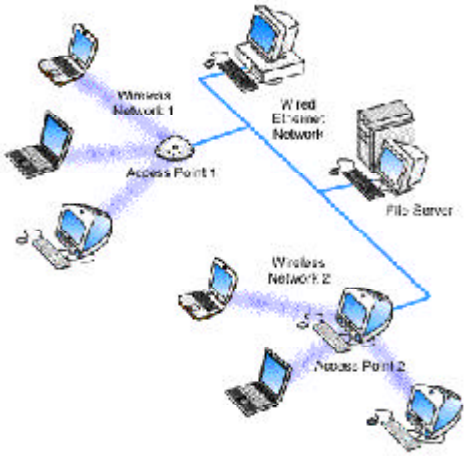


Figure 4: Multiple Access Points.

If a single area is too large to be covered by a single access point, then multiple access points or extension points can be used. The extension points are not defined in the wireless standard, but have been developed by some manufacturers. When using multiple access points, each access point wireless area should overlap its neighbors. This provides a seamless area for users to move around in using a feature called "roaming." Some manufacturers produce extension points, which act as wireless relays, extending the range of a single access point. Multiple extension points can be strung together to provide wireless access to far away locations from the central access point which is shown in Figure 5.Wireless connected computers using an Access Point with an Extension Point.

Wireless Connected Computers using an Access Point with an Extension Point

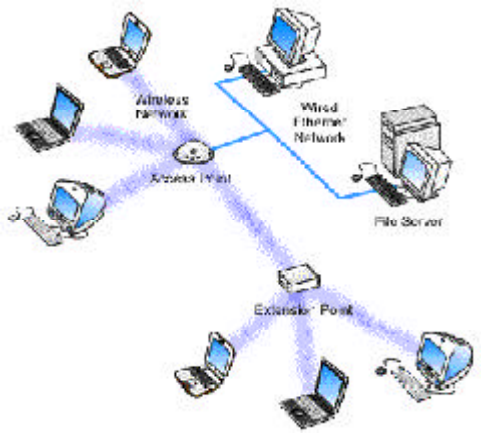


Figure 5: Extension Point.

4.1 Roaming

A wireless computer can “roam” from one access point to another, with the software and hardware maintaining a steady network connection by monitoring the signal strength from in-range access points and locking on to the one with the superior quality. [9] Usually this is completely transparent to the user; they are not aware that a different access point is being used from area to area. Some access point configurations require security authentication when swapping access points, usually in the form of a password dialog box. Access points are required to have overlapping wireless areas to achieve this as can be seen in the following diagram:

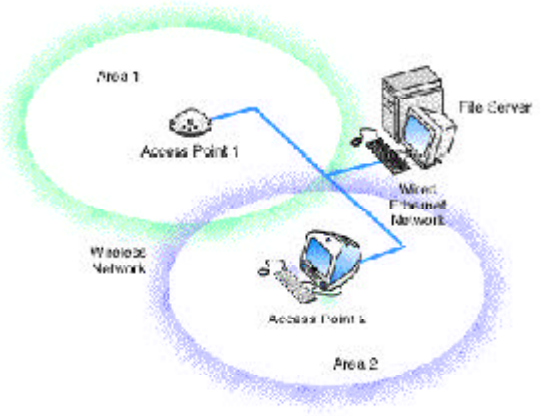


Figure 6: Roaming.

A user can move from Area 1 to Area 2 transparently. The Wireless networking hardware automatically swaps to the Access Point with the superior signal.

Not all access points are capable of being configured to support roaming. Also of note is that any access points for a single vendor should be used when implementing roaming, as there is no official standard for this feature.

5. Wireless Network to Interconnect Two LANs

Wireless networking offers a cost-effective solution to users with difficult physical installations such as campuses, hospitals or businesses with more than one location in immediate proximity but separated by public thoroughfare. This type of installation requires two access points. Each access point acts as a bridge or router connecting its own LAN to the wireless connection. The wireless connection allows the two access points to communicate with each other, and therefore interconnect the two LAN's.

A Hardware Access Point providing wireless connectivity to local computers and a software access point. The software access point provides Wired Ethernet network 2 computers access to Wired Network 1.

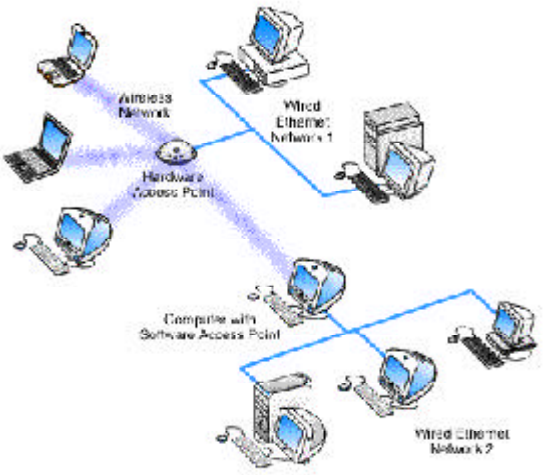


Figure 7: LAN to LAN Wireless Communications

Note that not all hardware access points have the ability to directly interconnect to another hardware access point, and that the subject of interconnecting LAN's over wireless connections is a large and complex one.

6. Security Scenario

Wireless communications obviously provide potential security issues, as an intruder does not need physical access to the traditional wired network in order to gain access to data communications. However, 802.11 wireless communications cannot be received —much less decoded— by simple scanners, short

wave receivers etc. This has led to the common misconception that wireless communications cannot be eavesdropped at all. However, eavesdropping is possible using specialist equipment.

To protect against any potential security issues, 802.11 wireless communications have a function called WEP (Wired Equivalent Privacy), a form of encryption which provides privacy comparable to that of a traditional wired network. If the wireless network has information that should be secure then WEP should be used, ensuring the data is protected at traditional wired network levels. Also it should be noted that traditional Virtual Private Networking (VPN) techniques will work over wireless networks in the same way as traditional wired networks. [3]

A recent survey highlighted that 25% of organizations not using wireless LANs were held back by security concerns**.** No organization wishes a user to walk into the building and gain access to the private staff network or any module of the management system. Restrictions need to be made on who can access the network and from what access point or building. However, security provisions can be built into wireless LANs making them as secure as most standard LANs. [8]

6.1 Unauthorized access

Unauthorized users accessing network through the WLAN/LAN are a major security concern. We have to provide some mechanism, which will, denies unauthorized users access or limits their access to public network segments such as the Internet.

6.2 Unauthorized devices

Some devices, such as unauthorized laptops or PDAs, can leave you wide open to attack. It is needed to provide some mechanism to automatically discover any new devices on your network and immediately alerts to administrator.

7. Costing

Although running costs can be comparable to traditional wired networks, wireless transmission and reception equipment is generally much more expensive than the cost of comparable wired components. For Example the cost for DWL-520+ PCI Adapter and DWL 1000AP+ Access point of D-Link product is approximately Rs. 11500 and Rs. 18000 respectively while many other wireless instruments are available in the market but the cost differs depending on the requirements and manufacturing company.[7]

7.1 Wireless Networking and the Internet

Once you realize that wireless cards are analogous to Ethernet cards and that empty space is analogous to Ethernet cabling, the answer to this question becomes clear. To share an Internet connection across a LAN you need two things:

- An Internet sharing hardware device or software program
- A LAN

If your LAN is wireless, the same criteria apply. You need hardware or software access point and a wireless LAN. Any computer equipped with a wireless network card running suitable Internet sharing software can be used as a software access point. (See Figure 8) A number of vendors offer hardware access points.

A hardware access point may provide Internet Sharing capabilities to Wired LAN computers, but does not usually provide much flexibility beyond very simple configurations. (See Figure 9)

Wireless connected computers using a Software Access Point for shared Internet access.

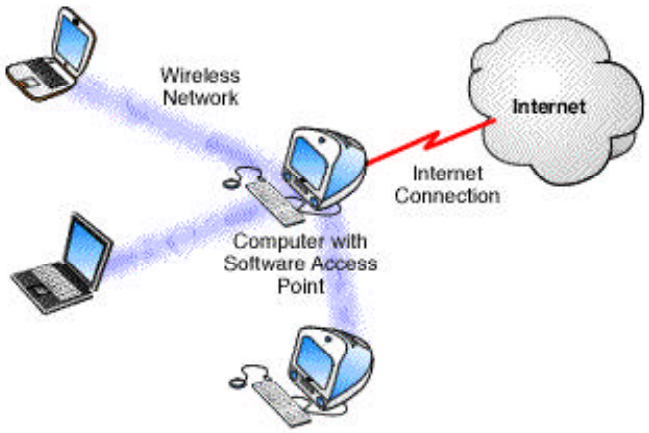


Figure 8: Software Access Point.

Wireless connected computers using a Hardware Access Point for shared Internet access.

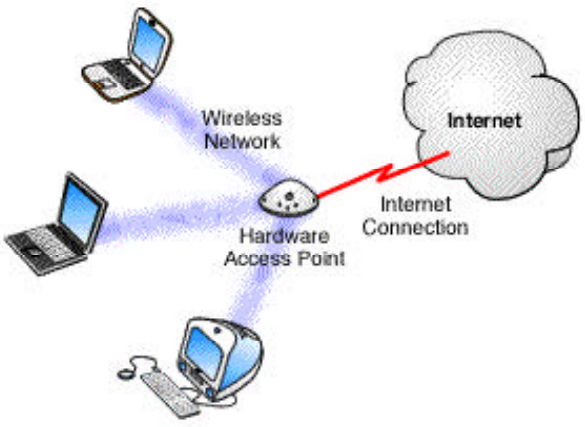


Figure 9: Hardware Access Point

If an existing wired LAN already has an Internet connection, then the hardware access points simply connect to your LAN and allow wireless computers to access the existing Internet connection in the same way as wired LAN computers.

Wireless connected computers using Multiple Access Points.

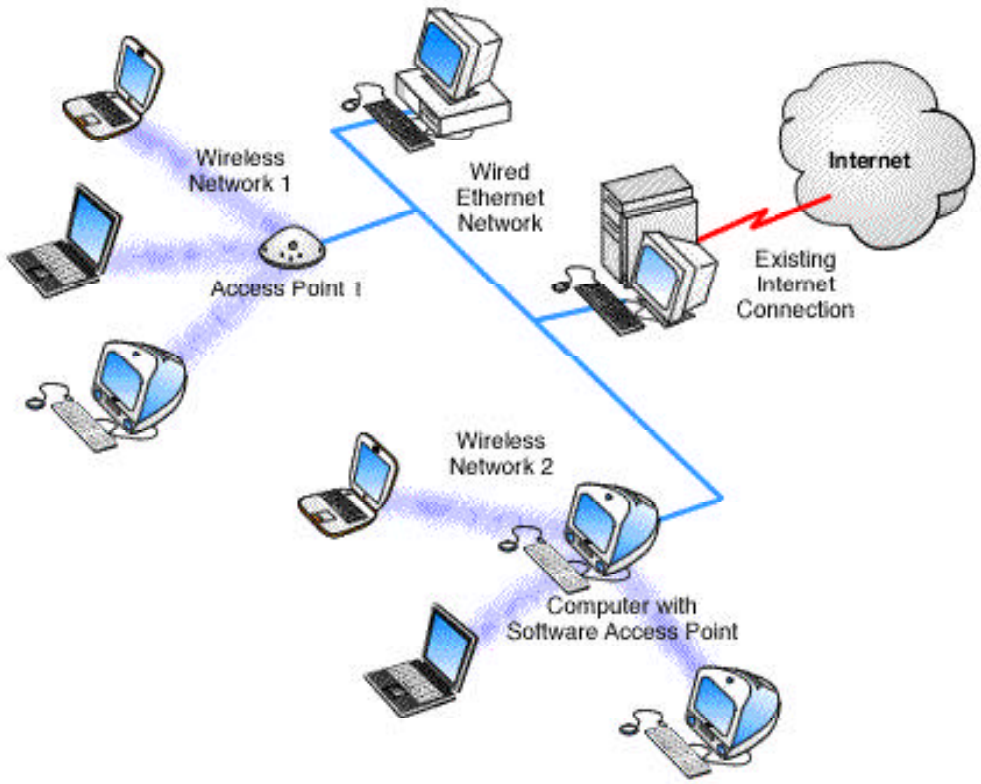


Figure 10: Multiple Access Points.

If there is no existing Internet connection, then this depends on the access point

Wireless connected computers using Multiple Access Points. All wired and wireless computers access the Internet through a single software access point.

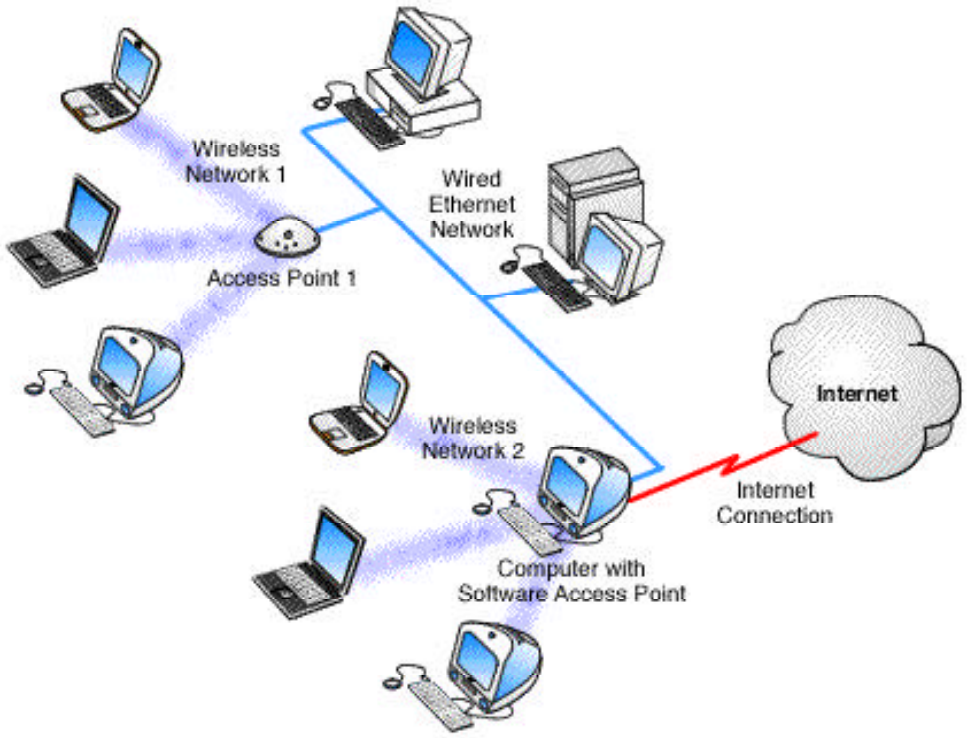


Figure 11: Software Access Point sharing one Internet connection.

If an access point provides some form of Internet sharing itself, then having multiple such access points connected to a wired LAN may require some special configuration, or possibly may require an additional Internet sharing device or software program.

8. Conclusion

The extension of the domain of the LAN and its integration with another network makes it convenient to adopt the technology of wireless LAN. This extension open ups an easy connectivity path towards hand held devices. Which is infecting the tomorrows need? Whatever is invested in network technology remains fruitful but an additional investment on wireless connectivity for internetworking and global connectivity creates an environment unconstraint information access even when an individual is on the move. A futuristic extension of sensors network willl also brings comparative ease. The modeled implementation is first step towards WLAN / LAN and Internet connectivity with considerable implementation of needed information security.

9. References

1. CSI Communications ISSN 0970-647x January-2004 page No.10-12
2. Wireless products and its applications www.winncom.com accessed on (15-5-2004)

3. IEEE 802.11b Wireless LANs http://www.wlana.com , www.3com.com Accessed on (20-5-2004)
4. Wirelesslan.com answer page, 2000. What is a Wireless LAN?http://www.wireless.com accessed on (19-4-2004)
5. IEEE standards. www.ieee.org accessed on (23-6-2004)
6. Stewart S. Miller (2003) Wi-Fi Security McGraw-Hill Companies, Inc.USA pp-75-80
7. For wireless products www.dlink-india.com Accessed on (18-11-2004)
8. Still secures A Guide to Wireless Network Security http://www.stillsecure.com Accessed on (21-1-2004)
9. Eric Ouellet, Robert Padjen (2002) Building a Cisco Wireless LAN Syngress Publishing, Inc. USA pp- 69-81, 183-185

About Authors



Mr. Atul Gonsai has completed his Graduation in Bachelor of Business Administration (BBA) and Master degree in Master of Computer Applications (MCA) then recently in completed his Ph.D thesis work in Computer Networking. He is working as Assistant Professor in Department of Computer Science Saurashtra University Rajkot after completing his MCA from the same Institute in April 2000. He has written 13 research paper and 3 paper submitted for conference for acceptance. His research interest encompasses performance tuning of computer networks, wireless networks, High performance networking etc. He has taken part in 9 national and 3 international conferences and seminars. He is DCNI D – link Certified Network Integrator from D-link India Goa. He has published one book and is also writing one book on Computer Networking, which is not yet published. He is maintaining and handling MCA computer Labs and Saurashtra University Library Networking. He is life Member of ISTE New Delhi.
Email : atulgosai@yahoo.com or amgosai@sauuni.ernet.in



Dr. N N Jani is currently acting as Prof. & Head, Department of Computer science, Saurashtra university Rajkot. He obtained his masters degree of MSc then Ph.D in the area of material science. He has 60+ research contributions to his credit along with 10 books. He is currently guiding research scholar leading to Ph.D degree in the area of High performance networking, High performance, Biometric technology, data ware housing, data mining application and web services applications. His current research is on Embedded systems, MEMS and in near future in the area of nanotechnology.
Email : nnjani@sauuni.ernet.in



Mr. Nilesh N Soni has completed his Graduation in Bachelor of Commerce and also Library & Information Sc. with Master degree in 1992. He is working as I/c University Librarian, Saurashtra University Library Rajkot. He has written 7-research paper and 1 paper submitted for international conference for acceptance. His research interest encompasses computerization of Library, Digitization of Library Collection, networks, wireless networks etc. he has already computerized Nirma and Saurashtra university library. He has taken part in 5 national and international conference and seminars. He is life member of ISTE – New Delhi, ILA- New Delhi, Gandhralaya Sewa Sangh- Gujarat. He is engaged in the project to create a database of Rajkot Library Networking.
Email : sulnilesh@yahoo.co.uk

Library Consortia Model for Country Wide Access of Electronic Journals and Databases

A T Francis

Abstract

The present approach towards partnership, networking, consortia and resource sharing adopted by Indian libraries need radical changes to evolve responsive partnerships in order to achieve best performance in service. The current practices of journal acquisition in most of the libraries in the colleges and universities in India are print based, in which each library is an island with regard to access of information. Moreover, there is wide disparity in the availability and use of information among different universities and colleges. But, consortia based acquisition and electronic desktop delivery of information can eliminate this gulf and increase the access and use considerably.

Thus the difficulty now faced by the students, teachers and scientists in getting academic and research information will be eliminated on achieving full bibliographical control on the information documents available world over.

This paper depicts the benefits of library consortia, analyses the present trend in the formation of consortia in India and suggests a new model of library consortia in which all academic institutions and government research organizations could participate. The formation of such a unique consortium under the direction and full support by the Government of India is stressed. The role that can be played by the INFLIBNET Centre of the University Grants Commission in the formation and management of such a consortium is also depicted. The areas of re-defining and re-engineering the operations of the university libraries necessitated by the consortia based electronic information and document delivery services are also discussed.

Keywords: Library Networking, Consortia, Library Consortia Model, Indian Library Consortium, University Libraries, Re-Engineering

0. Introduction

The Library and Information Systems(LIS) and Services are being transformed by the modern Information and Communication Technologies(ICTs). The ICTs have the potential to transform both the processes and products of the entire economic sphere - as well as all types of market transactions, institutional linkages, and human interactions and learning. Since nearly all economic activities rely on information acquisition, processing, and transmission, the scope for the use of this technology is unbound. Information is the "lifeblood" of competitive markets, and improvements in information technologies are transforming the whole economies into fast-moving information-intensive economies and globalising production and competition in many industries and services. The libraries and information centres are one of the major supporting agencies involved in the process of information transfer and finally the diffusion of information and information technology. Consequently, both the Information Systems and the Information Professionals are adapting to meet the changing needs and growing expectations of the users (Jalloh, 2000). The initiatives and development in the areas of automation, networking, resource sharing, consortia, digital libraries, electronic document delivery, etc. have causes to emerge new practices in the operations and management practices of the Library and Information Systems world over. As compared to the

research libraries, the university and college libraries in India have shown a slow rate of progress in this direction. Now, these libraries also have shown increased interest in the inevitable refinement of their management processes.

1. Library Consortia: Bridging the Gulf in the Availability of Information

The phenomenon of information revolution has posed several problems and this has far reaching implications in the society. The nation or society which possesses more information will lead the world. This is also true in the case of individuals. The persons have more information will guide a group or society and they will be superior to others. This power of information has induced the nations and individuals to acquire and control more and more quantities of information. But, in this race, the poor nations, societies, institutions or individuals will be back as compared to others. This has created a big gulf in the availability and use of information. A study by the author in 1997 revealed that while 77 per cent (total no. 2,31,510) of the higher education faculty in India were engaged in the affiliated colleges, only 23 per cent (total no. 69,283) of the faculty were engaged in the Universities (Francis, 1997). Though the salary and other emoluments given for the faculty in colleges and universities are same, there is much difference in the scientific productivity and research out contributed by these two classes of the faculty. A major reason for this difference is the lack of availability and use of research information.

As compared to the university libraries, only a meager amount was allotted for college libraries in India. Over the time, academic institutions typically have spent a decreasing percentage of their educational and general budgets on their libraries. Nonetheless, academic institutions and library clients expect their libraries to obtain new electronic resources while simultaneously maintaining or growing traditional print collections until the electronic resources are fully stable. Libraries also are expected to do this with no additional funding. Academic libraries and information providers must use information technologies to facilitate increased information delivery and to make e-information more generally, readily, and flexibly accessible than its print counterpart. The current practices of journal acquisition in most of the libraries in the colleges and universities are print based, in which each library is an island with regard to access of information. The digital libraries expected to increase information access. They will allow for greater standardization of data, multiple and remote access to information resources, easy sharing, etc. (Aregu, 2001).

The Library Consortia can be an ideal solution in this context, if that has been established and managed at the wider interests of the society and the mankind in total. The activities and operations of the library and information centres are being influenced and drastically changed with this new approach to information management. The pattern of common acquisition, subscription or licensing for access by the consortia will benefit more to the poorer group.

The conventional practices of journal acquisition are grounded in the legacy of a print-bound world in which each library is an island of access for its own patrons. But with electronic desktop delivery of information, the increased ease of access allows far greater information use than previously possible (Sanville, 1999). Experiences of the several libraries show that the improved ease of access has demonstrated the high elasticity in information usage. Libraries in consortia model of purchase of electronic journals can seek this desirable outcome that provide for expanded journal access. A study in the Ohio University Library System reveals that the use of journal titles has increased by three times than they previously held in print. Even the small and new colleges were also the beneficiaries through access to scholarly journals. As the evolution to broad scale electronic access continues, libraries and consortia must take advantage of the opportunities to adopt sustainable economic model of information purchase that maximizes information use.

There are several efforts to operate library consortia at regional, national or international levels. The Washington Research Library Consortium (WRLC, 2004)) is a good regional resource-sharing organization established by several universities in the Washington, D.C. metropolitan area to expand and enhance the information resources available to their students and faculty. The International Coalition of Library Consortia (ICOLC, 2004) is an informal organization that began in 1997. Comprising about sixty library consortia in the United States, Canada, the United Kingdom, the Netherlands, Germany, Israel, and Australia, the Coalition represents over 5,000 member libraries worldwide. The Coalition serves primarily higher education institutions by facilitating discussion among its members on issues of common interest.

2. Present System of Consortia in India

Several working models of library consortia are prevailing in different nations and among different groups of institutions. In India, we have seen some library consortia such as UGC Infonet of the University Grants Commission, the INDEST (The Indian National Digital Library in Engineering and Science and Technology) of the Indian Institute of Technologies (IIT) and similar institutions, Consortium of the Council of Scientific and Industrial Research (CSIR), etc. Several other organizations such as the Indian Council of Agricultural Research (ICAR), State Agricultural Universities, Indian Space Research Organisation (ISRO), Defence Research and Development Organisation (DRDO), All India Council of Technical Education (AICTE) and some other individual groups of institutions have started working to form different consortia.

3. Why Separate Consortia?

All consortia are formed on the basis of memorandum of understanding and agreements between the publishers, database vendors, user institutions, libraries, etc. Since the basic responsibility of providing infrastructure for education and research in India is vested with the Union and State Governments and the fund required for this is meeting from the government exchequer, the present system of formation and maintenance of different library consortia for each group of academic and research institutions is unscientific. The contradiction is that all educational and most of the research institutions in the country are getting full or partial financial support from the government. All universities, colleges and poly techniques in India are funded either by UGC, ICAR, AICTE, Indian Medical Council (IMC) or similar government agencies. All major research institutions directly run by the Government of India. There are some such institutions under the state governments also. This situation has posed a major question, "Why separate consortia for each group?". Potter (1997) argues as follows. "The fact that a group of libraries shares a common funding source, be it directly through elected officials or through a board of regents or oversight agency, is an important reason to build statewide cooperative systems. There is great appeal in efforts to pool resources and in cooperating to control costs". GALILEO in Georgia, the Louisiana Library Network, OhioLINK, TexShare in Texas, and VIVA in Virginia were some of the consortia functioning in USA in 1997 with state wide operation.

4. Consortia for Countrywide Access: a future model

The present system of consortia has lot of merits over the earlier pattern of individual subscription to journals and databases. Such benefits can be maximized by establishing consortia for nation wide access. Instead of establishing separate library consortia by different groups of educational and research institutions, it is better to form one consortium for all educational and government research institutions with country wide access to all online journals and databases. Such a consortium in India may be named as, "Indian Library Consortium" (ILC). Here, the unnecessary duplication of efforts for the establishment and maintenance of separate consortia within a country can be avoided. Moreover, wide disparity existing

in the availability, accessibility and use of information among education and research institutions can be eliminated to a large extent by this system. This disparity will be minimized on establishing high speed Internet connectivity with uniform per user bandwidth to all academic and research institutions.

The publishers and database vendors also will be benefited in this venture because their efforts for marketing and providing services and technical support will be less. Access to electronic journals and online databases are now controlled by the database vendors either by authenticated I. P. (Internet Protocol) Numbers of the Proxy Servers or by User Name and Password. In a single consortia, they can adopt the Gateway based Access Control. The vendors can establish Mirror Server Centres in each country. This has several benefits in database maintenance, backup and service. So, it is high time to consider the model for a single consortium and a real network at country level. Several experts and users have felt the need for such a network or consortium of all LIS. But, it is difficult to include the commercial organizations and industrial houses in the consortia visualized for academic and government research institutions.

5. Role of INFLIBNET

The established task of the INFLIBNET (Information and Library Network) of the UGC is to interconnect all educational and research institutions in the India for pooling and sharing of library and information resources and there by ensure easy availability and speed access of information resources for all academic and research programmes in the country (UGC, 1988). The service purview of the INFLIBNET covers all types of educational and research institutions and government organizations and hence, it has the primary responsibility to plan and establish a single consortium for the whole country. This should not be limited to serve only the conventional universities and colleges, but include agricultural, veterinary, medical and engineering universities, all types of colleges, government research institutions engaged in the fields of defence, space, industry, agriculture, medicine, atomic energy, etc. The fund required for establishing and maintaining the proposed Indian Library Consortium should be set apart either centrally in the Union Budget or by individual controlling institutions such as UGC, ICAR, AICTE, DRDO, ISRO, IMC, etc. In order to establish an integrated information system in India, increased administrative, financial and technical support by the government is essential (Francis, 1998). Though the role of the INLIBNET is clearly defined, more dynamic and strong actions are needed to achieve target.

6. Need for Re-engineering University Libraries

In reaping benefits of the modern ICTs, the library and information centres in India are far behind as compared to those in the developed nations. Inadequate fund, infrastructure, manpower training, unscientific rules and management policies, etc., are the reasons for this situation. Uneven distribution of the available resources causes wide disparity in information availability, access and use. Gaur (2003) indicated the need for proper management models suitable for the modern ICT environment as follows. “It is important to find out why Indian Libraries and Information Centres have not been able to benefit to the extent expected by the computer revolution in spite of huge investments, and with so much of hue and cry. … But, in reality all these effort have gone as waste. Why is it so?. Are these efforts not in proper direction ?. Or is there something wrong in our planning?. In this process there will be a need for models and frameworks that help us to understand and identify specific problems”.

The advent of electronic journals and online databases coupled with high speed data communication facilities has paved the way for the present form of library consortia. The model of single library consortia, proposed for the whole country, can bring an ideal situation of information availability and use, which provide maximum economy and service efficiency.

Consortia based information acquisition, processing and servicing warrant a total re-structuring of the entire processes of the university libraries. The principles of centralized and co-operative operations as advocated by Dr. S.R. Ranganathan can be effectively implemented in the consortia mode. Maximum benefits of such a system can be reaped in a single consortium for country wide access of information resources.

7. Major Areas of Re-engineering

A study conducted mainly in the Kerala Agricultural University reveals that a thorough re-structuring of the University Library and Information System is needed to suit the requirements warranted by the modern ICTs. In has felt that the modern ICTs warrants total changes in the processes of acquisition, technical processing, user education, services, human resource development (HRD), financial management, etc. of the university libraries. It is visualized that, the consortia based operations and electronic document delivery services of the libraries will enhance this need for thorough re-structuring. In comparison, it is revealed that similar situation prevails in almost all universities in India and the findings are relevant to them also.

- Re-engineering the Acquisition : The activities of document suggestion, selection, approval, order placing, passing of bills for payment, releasing payment, etc. can be done through the local area network and or Internet. The work of accessioning can be done using computers. Due to some audit regulations, it may not be possible initially to replace the paper form of Accession Register by computer form. So, till the audit regulations are getting revised, the libraries can prepare Accession Register by printing the document details from the OPAC (Online Public Access Catalogue) in Loose Sheaf. This will avoid a lot of duplicate and unnecessary work of accessioning. Moreover, multiple copies of accession registers can also be printed easily.
- Re-engineering the Classification and Cataloguing : It is ideal to centralize globally the work of classification and cataloguing. The Library of Congress (LC) is already doing such work efficiently. Few libraries in India are availing the service of LC for downloading the classification and catalogue data. The ILC can either subscribe the service of the LC or the ILC itself can start such service for all libraries in India. It is ideal to collaborate with the National Library, Calcutta in this venture. Here, the work of preparation of Indian National Bibliography also will be automatically supplemented.
- Re-engineering the User Services : The work of user services has to be thoroughly re-engineered on the lines of consortium and electronic document delivery. The procedures for all user services such as compilation of bibliographies, current awareness service, selective dissemination of information, photocopying, other documentation services, CD based and online database services, etc. has to be re-defined and need total revision.
- Re-engineering the User Education : The user education programmes need to be more technology oriented. Such programmes should be able to impart minimum theoretical practical knowledge on computers, printers, scanners, network based systems, search software of different database vendors, Internet search engines and services, web browsing, downloading, e-mail, data copying to CDs and DVDs, multimedia applications, computer viruses, etc. The orientation and training should be designed in such a way to provide confidence in locating and using the required and authentic information easily.
- Re-engineering the Human Resource Development : The aspect of re-engineering with regard to the human resource development is mainly concerned with the areas of staff selection, orientation, training, technology adoption, work study, change management, motivation, etc. The technologies, methods and procedures, used in the library and information systems are more dynamic and

changing as compared to many of the other professions. But, on the contrary, it is experienced that a large number of library and information professionals were reluctant to change. This also contributed to the low performance in the profession. This situation stresses the need for a total revision of the HRD policies of the LIS.

In order to prepare professionals and other staff in the libraries, Continues Education Programmes (CEP) should be conducted regularly. The course contents and conduct should be evaluated and reviewed frequently. The compulsory orientation and refresher courses conducted by the UGC for the library professionals / teachers in the UGC Cadre have helped to improve the situation. But, it is evaluated that the conduct of the UGC refresher courses in many of the universities is downgraded to that of a routine process. Lack of infrastructure and latest technology systems, non-availability of competent and experienced faculty, inadequate interest in participants, etc. are some reasons for this situation.

The CEP has also to be provided for the professionals below the UGC cadre, that is, semi-professionals and the non-professionals, because, every staff in a library has important role in winning performance.

- Re-engineering the Financial Process : In modern libraries, the stress will be on electronic resources. So, more money has to be set apart for purchase and maintenance of hardware and software systems, Internet connectivity, etc. than the conventional pattern of spending for large buildings, furniture, books, journals, binding, etc. Moreover, in consortia system, lot of work of acquisition, processing and services will be centralized. Hence, a substantial portion of the fund will have to be either deviated or pooled for consortium. This will need revision in budget process and financial management.

8. Re-structuring the Divisions of the University Libraries

The conventional classification of work into various departments or divisions or sections of the university libraries and their nomenclature such as Acquisition, Classification, Cataloguing, Circulation, Documentation, Serials Control, Inter Library Loan, etc. are not relevant in the modern environment of ICTs and the consortia based services. Close analysis of the present pattern of operations and services revealed that a re-definition of the functional divisions of the university libraries as follows is essential:

1. Information Acquisition Division : This include the work of acquisition of all types of documents such as books, journals, CDs, cassettes, etc.
2. Information Processing Division : The classification, cataloguing, development and up-dation of all types of in-house databases, OPAC or Web OPAC management, etc. The process of classification and cataloguing will be centralized at national or global level and hence the main work will be downloading and uploading of data and doing modifications for local situations. Digitization of theses and dissertations will have to be done regularly.
3. Information Services Division : Membership management, User orientation and education, Reference, circulation, bibliographic and documentation services, Inter Library Loan, Electronic Document Delivery, CD and Internet based database services, downloading, printing and copying services and all other types of user services can be assigned to this division.
4. Information Technology Division:- Acquisition and maintenance of computers, printers, copying equipments, systems for Internet connectivity such as V-SAT, leased line, Modems, Switches, Routers, etc., will be the main work of the division. Purchase, installation, training and maintenance of general and library management software, firewalls, etc. are to be done by this division. Web page designing, web hosting, management of electronic marketing and user surveys, etc. can be entrusted to this division.

9. Conclusion

The possibilities of ICTs, Digital Information, Electronic Document Delivery, Library Consortia, Web based operations, etc. have helped to provide better services to the users. But, wide disparity in the availability and use of academic and research information still prevails among different universities and research institutions in India. Moreover, establishment and maintenance of separate library consortium for each group of government and government supported institutions in a country has lead to the duplication of efforts and additional investment. Since the present pattern of higher education and research is inter-disciplinary, clear cut demarcation of areas of subject interest and information requirement is difficult. That means, the information requirements are cross-disciplinary and also at micro-level. This underlines the need for providing access of information in all subject areas to the students, teachers and researchers in all branches. This justifies the establishment of National Library Consortium which automatically will bring economy, efficiency and equality in information availability and use.

10. References

1. Aregu, Raphael (2001). Digitizing agricultural data for rapid agricultural modernization in Uganda: strengthening Ugandan NARS issues. In Digital libraries: Dynamic Landscapes for Knowledge Creation, access and Management. The fourth International Conference of Asian Digital Libraries, Bangalore, India, December 10-12, 2001, organized by University of Mysore and Indian Institute of Information Technology, edited by Shalini R. Urs, TB Rajashekhar and KS Raghavan. Bangalore: ICADL. 404-415 pp.

2. Francis, AT (1997). Regional Information Networks: necessary thrust area for INFLIBNET to establish Integrated Information System in India. In Information Technology Applications in Academic Libraries in India with emphasis on Network Services and Information Sharing. Fourth National Convention of Libraries in Education and Research (CALIBER – 97), Patiala, India, March 6-8, 1997, organized by INFLIBNET Centre and Thapar Institute of Engineering and Technology, edited by AL Moorthy and PB Mangala. Ahmedabad: INFLIBNET Centre. 102-106 pp.

3. Francis, AT (1998). Integrated Agricultural and Rural Information System (IARIS): an evaluation of the existing Information System in Kerala. In Towards the new Information Society of Tomorrow: Innovations, Challenges and Impact. 49th FID Conference and Congress, New Delhi, India, October 11-17, 1998 edited by Malwad NM et al. New Delhi: INSDOC. III-153 pp.

4. Gaur, Ramesh C (2003). Reengineering library and information services: process, people and technology. Mumbai: Allied Publishers. 112-114 pp.

5. ICOLC (International Coalition of Library Consortia) (2004): Statement of current perspective and preferred practices for the selection and purchase of Electronic Information. http:// www.library.yale.edu/consortia/ icolcpr.htm. (Accessed on 04-10-2004).

6. Jalloh, Brimah (2000). A plan for the establishment of a library network or consortium for Swaziland: Preliminary Investigations and Formulations. Library Consortium Management: An International Journal. 2(8) 165-176.

7. Potter, William Gray (1997). Recent trends in Statewide Academic Library Consortia - consortia - consortia discussed include Georgia's GALILEO, Louisiana Library Network, OhioLINK, TexShare, and Virginia's VIVA - Resource Sharing in a Changing Environment. Library Trends. Winter. 45(3). Accessed through htth://www.findarticles.com).

8. Sanville, Tom (1999). Use levels and new models for consortial purchasing of electronic journals. Library Consortium Management: An International Journal 1(3). 47–58.

9. University Grants Commission (1988). Development of an Information and Library Network, Report of the Inter Agency Working Group. New Delhi: UGC. 488 p.

10. Washington Research Library Consortium (2004). WRLC Program Goals: Shared Digital Library (Library Information Technology Services). http://www.wrlc.org. (Accessed on 11-11-2004).

About Author



Mr. A.T. Francis holds MLIS and MCOM. Presently he is working as Assistant Librarian in the Kerala Agricultural University. He has worked as a Library Professional for more than 15 years in organizations, like DRDO, Lakshadweep Administration, etc. His areas of interest are Information and Communication Technologies, Library Automation, Networking and Resource Sharing, Re-Engineering and Management of Library and Information Systems, HRD and User Education and Training.
E-Mail : francisaloor@yahoo.com

Subject Gateways: An Overview

R T Yadav

Abstract

Libraries have been undergoing tremendous developments in this era, adapting to the advent of information technology in the day to day working. Electronic, Digital Libraries have arrived. OPAC, CDROM databases, Electronic journals and Internet access exist in most Libraries. Libraries are no longer storehouses but are becoming gateways of relevant information. Some progress has been made towards increasing the relevancy of the data with the induction of various search engines and subject directories. Despite these activities, information sources remain scattered, hared to find and difficult to access. With the Internet, WWW and Information explosion, identification and extraction of information resources is an essential function of all libraries and information centers. The Electronic information sources are rapidly growing and with a wide variety in form and content it takes a lot of time to get the required information. Using technology the institutions, associations and individuals build a kind if network resource discovery service, called "Subject Gateways" on the web, which is de-facto network use environment. These subject gateways evolved during the last five years among early digital library projects within the library communities of the various countries. Subject gateways allow libraries and related organizations to explore the usefulness of their subject expertise in the organization of knowledge in the word of network-based, Digital information. This paper gives an overview of the subject gateways, its definitions, feature and lastly some useful subject gateways and their services.

Keywords : Subject Gateways, Portal, Information Services

0. Introduction

Subject gateways do not exist in isolation. For the user they form part of the wider experience of resource discovery. On the one hand, the searcher whether child or professor is faced with the compelling option of using the global services, such as Yahoo, Altavista, and Google, as a first step. The Undifferentiated experience offered by such services can be compared with the specialist view offered by information gateways. Gateways offer the user an alternative to the generalist approach of the commercial global search engines, but in order to optimize the gateway service we need to gain a better understanding of users' requirements for particular types of search during the research and learning process. It would be instructive to compare information seeking behaviour and success rates for a variety of uses of global search engine as compared to gateways. Likewise one could analyse the differences in users' search strategies within the context of the traditional library, hybrid library and subject gateway. It may be helpful to liken the subject gateway approach to the traditional "departmental library'' as the user's first port of call, a place where the user feels comfortable in a known environment and is able to gain skills to navigate a limited area of information. It would be interesting to see how far we could draw parallels between the requirements of users of subject gateways and the users of ``subject based'' libraries.

The users of both services benefit from an understanding of the boundaries and content of the information space they are accessing. It is important to connect user behaviour regarding Internet resource discovery with wider issues relating to the use of information in the learning and research processes. What does the user want from the research experience? Understanding users' behaviour in relation to gateways will enable gateway managers to position themselves within the mesh of existing gateways and meet the needs of their target audiences.

1. Why Gateways?

- Innovative ways of providing information and services such as electronic resources, course specific, and library help pages and document delivery.
- Complex library searches due to the cross functions and links between the online catalog, journal aggregator databases, electronic resources etc..
- The need to identify and present high quality free information resources on the web and distinguish those from library licensed material.
- The increasing expectations of users for interfaces to lead directly, without undue hunting, to the information or service they need.
- Many library web sites get congested with ample content, general objectives and duplication of services. It is no longer simple and clears how to do research or projects with a tool that provides gateways for getting information. There are number of steps that a library researcher must perform to successfully get information. Thus Gateways are needed to improve the effectiveness of Internet searching and will serve as a source of information in specific areas and saves the time of the users.

2. What are Gateways?

Moffat describes the establishment of the gateways as “a process if identification, filtering description, classification and indexing before they are added to databases which is freely available via World Wide Web (WWW).

- an online service that provides links to numerous other sites or documents on the Internet
- selection of resources in an intellectual process according to published quality and scope criteria (this excludes e.g. selection according to automatically measured popularity)
- Intellectually produced content descriptions, in the spectrum between short annotation and review (this excludes automatically extracted so-called summaries). A good but not necessary criterion is the existence of intellectually assigned keywords or controlled terms.
- intellectually constructed browsing structure/classification (this excludes completely unstructured lists of links)
- at least partly, manually generated (bibliographic) metadata for the individual resources

So we can say the internet search tools to help people find resources on the internet, e.g electronic journals, software, data sets, e-books, mailing lists or discussion groups, articles or papers or reports, bibliographies, bibliographical databases, organizational home pages, educational materials, news, resource guides and more.

Gateways offer linked collection of internet resources via a database of resource description. This can be

- Browsed - according to broad classification
- Searched - through index
- Quality controlled - due to selection

3. What are Subject Gateways?

Subject Gateway is an organized collection of resources on a given subject along with a retrieval mechanism. This essentially means that the scope of the search domain is well defined and limited to a subset of what exists in general. In the simplest form, the resources may be made available as a structured hyper-linked directory as followed by some of the search engines sites that offer directory services. The structuring followed is mostly only by broad subject areas. But for arranging topics under research areas where the intension of the subject areas is narrow and specific, broad directory structuring would not serve the purpose. Representation of subjects to that extent involves deploying various skills in the area of secondary information work such as document description, classification and derivation of subject headings

Subject Gateways is allowing links amongst electronic resources stored on services dispersed geographically on distant locations. The gateways sites redirect a user to the holders of the original digital material. A subject gateway can be defined as facilities that allow easier access to web-based resources in a defined subject area. The simplest types of subject gateways are sets of web pages containing list of links to resources. Some gateways index their lists of links and provide a simple search facility. More advanced gateways offer a much enhanced service via a system consisting of a resource database and various indexes, which can e searched and / or browse throughout a web based interface.

Subject gateways are also known as:

- Subject -based information gateways (SBIGs)
- Subject based gateways
- Subject index gateways
- Virtual libraries
- Clearinghouse
- Subject trees
- Pathfinders
- Quality : controlled subject gateways etc.

3.1 Subject Gateways Definition:

According to Dempsey L and other that “Subject gateways are internet services which supports systematic resource discovery. They provide links to resources (documents, objects, sites or services) predominantly accessible via the internet. The service is base on resource description. Browsing access to the resource via subject structure is an important feature.”

3.2 Subject Gateways are characterized by key factors:

- They are selective, pointing only to internet resources that meet with quality selection criteria.
- They are built by subject and information specialists - often librarians.
- Generally limit to specific subjects
- Scope-policy declaring what subjects they are indexing
- Defined target group e.g. academics, researchers, etc.

- Quality criteria - there is an official set of quality criteria
- Classification system - used as underlying system for browsing possibilities
- Use of open standards - to support co-operation with other services e.g. cross-searching
- Manually created records - rich resource description containing relevant information

3.3 Historical Development of Subject Gateways:

The subject gateways emerged in response to the challenge of resource discovery in a fast developing internet environment in the early and mid 1990's. Due to the emergence of the network information retrieval systems (Gopher, WWW, Archive, netfirst, etc.) and access protocols (ftp,gopher,telnet, http, etc.) innovative information technologies and services emerge.

The electronic libraries programme (e-Lib) of JISC of the UK Higher Education Funding Council set up in 1995, which includes besides other things, access to network resources (ANR) and subject gateways, were funded as the part of ANR area and latter on it latter on it lead to the funding and establishment of eLib subject gateways.

Different between Search Engine and Subject Gateways

Search Engine	Subject Gateways
General resource is available	It is a gathering place of discipline specific resources
It totally depends on the powerfulness of the lected search-engines algorithms.	High level of human input is there, as the se-resources must meet a number. of criteria applied by a librarian or academic, who ensures that only high quality, relevant resources are included in the database.
The results can be overwhelming, unmanageable, and full of irrelevant references and are often too prolific to meet user needs.	The results are specific, precise and linked to relevant documents.
Records are created by an automatic process and typically consist of mixture of metadata offered by the author of the page (if this is available) and text picked up from the page itself.	Records are created by a cataloguer, which is designed to highlight the main feature of resource in an easily readable, concise fashion.
Entries are displayed more as "raw-data"	Entries are described in a more "human- read able fashion"
In indexes pages. It indexes resources.	

3.4 Why Libraries should go for Subject Gateways?

Libraries are the most suitable institution to undertaken this work due to the following reasons:

The natural metaphor

- Browsing, reference desk
- Expertise in releveant area
- Classification, acquisition, keyword
- Information, seeking behavior
- Guiding & helping users

Benefits of Gateways for Libraries and Users

- Leading the way into the information age
- Communicating with non-nerds
- Access to huge-high quality collection
- Integrate into existing structures on the internet
- Diverse resource brought together
- Research, learning, leisure, enrichment - all brought together
- Someone to ask - what's where?
- What's good?

4. Key Initiatives for building tools and standards in Subject Gateways:

4.1 ROADS (Resource Organisation and Discovery in subject-based Services)

It is being funded by the JISC (Joint Information System Committee) through e-lib programme. It is an open source set of software toolkit, which enable the set up and maintenance of web-based subject gateways. A ROAD based information gateways is based on a database that contains information about internet resources. The records in the database contain information such as description and keywords. The user is given access to this information while either browsing or searching the database. This is particularly important fir geographically distant resources that might require some time and effort to access. The software includes the database technology, required to set up a gateways. For downloading the free online software visit its site URL: http://www.ilrt.bris.ac.uk/road

4.2 DESIRE (Development of a European Service for Information on Research and Education)

This is one of the largest projects funded by the Telematics for Research Sector of the Fourth Framework Programme funded by the European Union. In particular, DESIRE intends to provide:

- Tools for indexing and cataloguing information servers
- Tools for management and maintenance if information servers
- Demonstration and evaluation of tools and techniques for information catching and secure access to information servers
- Background information for developers of networked information systems
- Training materials

DESIRE published the “Information Gateways Handbook” a guide for libraries interested in setting up large-scale subject gateways of their own. This handbook is freely available at the site: (http://www.desire.org) and describes all the methods and tools require to set up a large scale internet subject gateways.

5. Subject or Portal Gateways

Subject gateways or portal variably called subject-based information gateways (SBIGs), subject-based gateways, subject index gateways, virtual libraries, clearing houses, subject trees, pathfinders, guide to Internet resources, and a few more variations thereof, provides an organized and structured guide to Internet-based electronic information resources that are carefully selected after a predefined process of evaluation and filtration in a given subject area or specialty. Subject gateways redirect a user to the holders of the original digital material. The subject gateways restrict their operation to providing linkages to independent third party sources.

Some of the important subject gateways are as follows:

- LibrarySpot.com: (http://www.libraryspot.com/)
- Librarians' Index to the Internet (LII) (http://lii.org/)
- Argus Clearing House (http://www.clearinghouse.net/)
- Galaxy (http://galaxy.einet.net/)
- Direct Search (http://gwis2.circ.gwu.edu/~gprice/direct.htm)
- Vlib: The Virtual Library (http://www.vlib.org/)
- Academic Info (http://www.academicinfo.com/)
- BUBL (http://bubl.ac.uk/)
- BIOME (http://biome.ac.uk/)
- The Scout Report (http://scout.cs.wisc.edu/report/sr/current/)
- LivingInternet.com (http://www.livinginternet.com/)
- Edinburgh Engineering Virtual Library (EEVL) (http://www.eevl.ac.uk)
- Social Science Information Gateway (SOSIG) (http://sosig.ac.uk/)
- Digital Librarian (http://www.digital-librarian.com/)
- QUEST.net (http://www.re-quest.net/) Internet Public Library (http://www.ipl.org/) BioMedNet (http://www.bmn.com/)

5.1 The Virtual Library (http://www.vlib.org/):

The Virtual Library is the oldest catalogue of the web, started by Tim Berners-Lee, the creator of html and the Web itself. Unlike commercial catalogues, it is run by a loose confederation of volunteers, who compile pages of key links for particular areas in which they are expert; even though it isn't the biggest index of the Web. The Virtual Library pages are widely recognized as being amongst the highest-quality guides to particular sections of the Web. Individual indexes live on hundreds of different servers around the world. A set of catalogue pages linking these pages is maintained at http://vlib.org. Mirrors of the catalogue are kept at East Anglia (UK), Geneva (Switzerland) and Argentina. Each maintainer is responsible for the content of their own pages, as long as they follow certain guidelines. The central affairs of the VL are now coordinated by a newly-elected Council.

5.2 Academic Info (http://www.academicinfo.com/):

Academic Info, online since 1998, began as an independent Internet subject directory owned by Michael Madin and maintained with the assistance of a quality group of subject specialists. In the spring of 2000 Michael left the University of Washington Gallagher Law Library to focus solely on Academic Info. In 2002 Academic Info became a registered non-profit organization of the State of Washington. Academic Info is now ad-free and relies on donations to remain online. Academic Info aims to be the premier educational gateway to online high school, college and research level Internet resources. The primary focus of the site is academic, with its intended audience at the upper high school level or above. A priority is adding digital collections from libraries, museums, and academic organizations and sites offering unique online content. The current focus is on English language resources but selectively sites in other languages will be considered. Users can search by subjects like The Arts, Biological Sciences, Business, Digital Library, Education, Engineering, Health & Medicine, History, Humanities, Law & Government, Library & Info Science, Religion, Sciences, and Social Sciences.

5.3 Librarians' Index to the Internet (LII) (http://lii.org/):

The Librarians' Index to the Internet (LII) consists of more than 8,600 Internet resources selected and evaluated by librarians for their usefulness to users of public libraries. Free e-mail subscription to the LII New This Week (http://www.lii.org/search/ntw) incorporates most recent resources added to the LII. It has close to 12,000 subscribers in 85 countries. ILL also offers co-branding service to the libraries that are members of the Library of California. The site provides both browsing and searching interfaces.

5.4 Argus Clearing House (http://www.clearinghouse.net/):

The Argus Clearing House is a guide to the meta resources. It provides a central access point for value-added topical guides that identify, describe, and evaluate Internet-based information resources. The Argus Clearinghouse is a non-profit venture run by a small group of dedicated individuals. The Argus Clearinghouse is intended to be a resource that brings together finding aids for students, researchers, educators, and others interested in locating authoritative information on the Internet.

5.5 LibrarySpot.com (http://www.libraryspot.com/):

LibrarySpot is a free virtual library resource centre for educators and students, librarians and their patrons, families, businesses and just about anyone exploring the Web for valuable research information. LibrarySpot.com aims at breaking through the information overload of the web and bring the best library and reference sites together. Sites featured on LibrarySpot.com are hand-selected and reviewed by an editorial team for their exceptional quality, content and utility. Published by StartSpot Mediaworks, Inc. in the Northwestern University / Evanston Research Park. LibrarySpot is the first in a family of vertical information portals designed to make finding the best topical information on the Internet a quick, easy and enjoyable experience. The LibrarySpot.com has received more than 30 awards and honours. Most recently, Forbes.com selected LibrarySpot.com as a "Forbes Favourite" site, the best in the reference category, and PC Magazine named it one of the Top 100 Web Sites. LibrarySpot.com has been featured on CNN, Good Morning America, CNBC and in many other media outlets.

5.6 Galaxy (http://galaxy.einet.net/):

Galaxy, originally a prototype associated with the DARPA-funded Manufacturing Automation and Design Engineering (MADE) program, is the oldest browsable / searchable web directory. It is a searchable Internet directory with a mission to provide contextually relevant information by integrating state-of-the-art

technology with the human touch. Galaxy employs the best of technology and human expertise to organize information in a way that makes it both understandable and highly relevant to users' needs. The information contents of the meta resource is compiled and organized by human Internet Librarians rather than by computer. The Galaxy hierarchy is built utilizing a vertical structure, i.e. the information on particular topics is very deep in content. While other search technologies may yield millions of pages per search (mostly extraneous), Galaxy provides concentrated, relevant results.

5.7 Direct Search (http://gwis2.circ.gwu.edu/~gprice/direct.htm):

Direct Search is a growing compilation of links to the Internet resources that contain data not easily or entirely searchable / accessible from general search tools like Alta Vista, Google, or Hotbot. Direct Search has its own search interface.

5.8 BUBL (http://bubl.ac.uk/):

BUBL LINK is the catalogue of selected Internet resources covering all academic subject areas and catalogued according to DDC (Dewey Decimal Classification). All items are selected, evaluated, catalogued and described. Links are checked and fixed each month. LINK stands for Libraries of Networked Knowledge. BUBL 5:15 provides an alternative interface to this catalogue, based on subject terms rather than DDC. The aim is to guarantee at least 5 relevant resources for every subject included, and a maximum of 15 resources for most subjects, hence the name 5:15. Big subject areas are broken down into smaller categories. However, the upper limit of 15 is not rigidly applied, so there may be up to 35 items for some subjects. The subject terms used in BUBL LINK / 5:15 were originally based on LCSH (Library of Congress Subject Headings) but have been heavily customized and expanded to suit the content of the service. The aim is to make it very easy to locate Internet information about a large number of subjects. The BUBL LINK catalogue currently holds over 11,000 resources. This is far smaller than the databases held by major search engines, but it can provide a more effective route to information for many subjects, across all disciplines.

5.9 LivingInternet.com (http://www.livinginternet.com/)

The mission of this web site is to make comprehensive, in-depth information about the Internet available around the world. The site was developed from 1996 through 1999, posted on January 7, 2000, and is updated weekly. The site is equivalent to a book of more than 600 pages, with more than 2,000 intra-site links and 2,000 external links woven into the text, making it the first Internet publication of a reference work fully integrated with the web on this scale. Google ranks the site number one in the Internet courses category, and Yahoo lists it as one of the top three sites on Internet history.

5.10 Edinburgh Engineering Virtual Library (EEVL) (http://www.eevl.ac.uk/)

Edinburgh Engineering Virtual Library (EEVL) is an award-winning free service, which provides quick and reliable access to the best engineering, mathematics, and computing information available on the Internet. It is created and run by a team of information specialists from a number of universities and institutions in the UK for students, staff and researchers in higher and further education, as well as anyone else working, studying or looking for information in Engineering, Mathematics and Computing. EEVL provides a central access point to networked engineering, mathematics and computing information. Resources being added to the catalogues are selected, catalogued, classified and subject-indexed by experts to ensure that only current, high-quality and useful resources are included. They include e-journals, databases, training materials, professional societies, university and college departments, research projects, bibliographic databases, software, information services and recruitment agencies. EEVL, in

addition to Internet Resource Catalogues, provides targeted engineering search engines - to UK engineering sites, to engineering e-journals and to engineering newsgroups, and to specialized information services, such as the Recent Advances in Manufacturing (RAM) bibliographic database, and the Offshore Engineering Information Service. MathGate at EEVL is involved in the Secondary Homepages Project for UK Mathematics Departments. EEVL's scope is limited to the three subjects, and is therefore more focused than the big search engines. Searching EEVL will retrieve high quality resources, but because EEVL's resources are handpicked, the numbers of sources covered in it are not comparable to the Internet search engines.

5.11 Social Science Information Gateway (SOSIG) (http://sosig.ac.uk/)

The Social Science Information Gateway (SOSIG) is a freely available Internet service which aims to provide a trusted source of selected, high quality Internet information for students, academics, researchers and practitioners in the social sciences, business and law. It is part of the UK Resource Discovery Network. The SOSIG Internet Catalogue is an online database of high quality Internet resources. It offers users the chance to read descriptions of resources available over the Internet and to access those resources directly. The Catalogue points to thousands of resources and each one has been selected and described by a librarian or academician. The catalogue is browsable or searchable by subject area. Social Science Search Engine is a database of over 50,000 Social Science Web pages. Whereas subject experts have selected the resources found in the SOSIG Internet Catalogue, those in the Social Science Search Engine have been collected by software called a 'harvester'(similar mechanisms may be referred to as 'robots' or 'Web crawlers'). All the pages collected stem from the main Internet catalogue this provides the equivalent of a social science search engine.

5.12 BIOME (http://biome.ac.uk/)

BIOME is a collection of gateways, which provide access to evaluated, quality Internet resources in the health and life sciences, aimed at students, researchers, academics and practitioners. A core team of information specialists and subject experts based at the University of Nottingham Greenfield Medical Library creates BIOME. The Internet resources are selected for their quality and relevance to a particular target audience. They are then reviewed and resource descriptions created, which are stored, generally with the associated metadata, and generally in a structured database. The consequence of this effort is to improve the recall and especially the precision, of Internet searches for a particular group of users. BIOME is a hub within the Resource Discovery Network (RDN) (http://www.rdn.ac.uk), and is funded by the Joint Information Systems Committee (JISC) (http://www.jisc.ac.uk/). There are five dedicated subject services (gateways) within BIOME, each covering a specific area within the health and life sciences. These gateways are AgriFor, VetGate, OMNI, Natural Selection and Bio Research.

5.13 The Scout Report (http://scout.cs.wisc.edu/report/sr/current/)

The Scout Report is the flagship publication of the Internet Scout Project. Published every Friday both on the web and by e-mail, it provides a fast, convenient way to stay informed of valuable resources on the Internet. A team of professional librarians and subject matter experts select, research, and annotate each resource. Published continuously since 1994, the Scout Report is one of the Internet's oldest and most respected publications. The Internet Scout Project is located in the Department of Computer Sciences at the University of Wisconsin-Madison, and is funded by a grant from the National Science Foundation.

5.14 Internet Public Library (http://www.ipl.org/)

The Internet Public Library is a product of the University of Michigan's School of Information and Library Studies. It includes extensive directories of online texts, newspapers, magazines and reference materials;

plus an exhibit hall and other special sections. At the moment, the home page features links to 4,699 critical and biographical sites dedicated to authors and their works, and an online history of the Harlem Renaissance in New York between 1900 and 1940.

5.15 Digital Librarian (http://www.digital-librarian.com/):

Maintained by Margaret Vail Anderson, a librarian in Cortland, New York. Internet information resources are catalogued according to subject categories and format-types. Digital Librarian does not have a search interface for the resources catalogued on the site. It has a browsing interface and see and see also references to related resources.

5.16 QUEST.net (http://www.re-quest.net/):

QUEST.net is a free online library offering substantive, fully annotated, links to valuable resources in both a unique frame version and a non-framed version. This web site serves to help students and professionals to locate day-to-day and much needed information and resources in a relatively quick and concise manner. It serves as a one-stop resource directory, providing the Internet community with thousands upon thousands of links, which it's committee of web surfers, have found to be the most useful, informative and productive. The meta resource provides fully annotated description of each link together with its URL allowing visitors to know what to expect from the Web site. Each link has been specially hand picked to provide with the best and most relevant links in each category. This web site is useful in an extraordinary way with it's devoted committee of web surfers work diligently, day-after-day, sorting through the vast galaxies of cyberspace to bring the best and most current resources available.

5.17 BioMedNet (http://www.bmn.com/):

BioMedNet is owned by Elsevier Science and is part of the Reed Elsevier group of companies. BioMedNet is the Web site for biological medical researchers. To date there are more than 800,000 members of BioMedNet with more than 20,000 people joining per month. Membership to BioMedNet is free and members can search all of BioMedNet without charge. However, viewing full-text articles from publishers often requires payment or a subscription. The site has links to more than 3500 reviewed information resources. The resource provides online access to more than 15,000 review articles. HMS Beagle: The BioMedNet Magazine is issued every fortnight. The Magazine can be subscribed by e-mail or can be accessed online.

5.18 Other Subject Gateways

- Biz/ed - Business and Economics Education on the Internet (http://www.bized.ac.uk/):

  Biz/ed (http://www.bized.ac.uk) is a free online service for students, teachers and lecturers of business, economics, accounting, leisure and recreation and travel and tourism. The gateway contains a ROADS based Internet catalogue with over 1400 Internet resources selected and described by subject experts.

  Biz/ed is targeted at students and teachers in the post-16 education sector, covering schools, FE colleges, universities and beyond. The site offers support for economics, business, accounting, leisure and recreation and travel and tourism at many different levels including AVCE, AS and A2 level, International Baccalaureate, HNC, HND and MBA.

  The Biz/ed site is a unique combination of primary and secondary teaching and learning resources. Resource discovery is integrated with simulations, worksheets, glossaries, spreadsheets, resource databases, online chat with examiners and a series of Virtual Worlds to give a rich package of support for teachers, lecturers and students.

- National Maritime Museum's Port (http://www.port.nmm.ac.uk/)

  Port is the National Maritime Museum's subject gateway to maritime information from the Internet. Subject gateways provide access to searchable and browseable catalogues of Internet based resources, all of which have been quality controlled or assessed before inclusion on the site. Librarians at the National Maritime Museum are actively involved in cataloguing and recording quality controlled online resources for Port.

  The information on Port has been grouped into twenty subject headings. To make Port easier to browse and to search, each subject heading divides into specific groups of information, improving accessibility for researchers. An example of this is the hierarchical structuring of Conflicts at sea, in which refined tiers of information on naval and military history lead users to subject-specific information on battles and wars, such as World War Two, followed by resources on D-Day.

  All websites included in the gateway are catalogued in records that assess the resource's content, origin, and nature. This means that users can access Port in the knowledge that they are looking at a quality-controlled collection of resources. When you use Port to locate online resources you know that it has the professional qualities and rigorous standards usually associated with such an institution as the National Maritime Museum.

- OMNI - Organising Medical Networked Information (http://www.omni.ac.uk/)

  OMNI (Organising Medical Networked Information) is a gateway to evaluated, quality Internet resources in health and medicine, aimed at students, researchers, academics and practitioners in the health and medical sciences. OMNI is created by a core team of information specialists and subject experts based at the University of Nottingham Greenfield Medical Library, in partnership with key organisations throughout the UK and further a field. OMNI, also provides training materials and workshops. Browsing can be done via either alphabetical topics, classified topics, or via MeSH headings. In addition, OMNI provides a range of biomedical value-added services, including a MEDLINE review section, mirrors of key NHS IT strategy documents, and the UK CME database.

  OMNI is one of the gateways within the BIOME service (http://biome.ac.uk/). BIOME is part of the Resource Discovery Network (RDN) http://www.rdn.ac.uk/, and is funded by the Joint Information Systems Committee (JISC).

- ICSU Navigator for Primary Scientific Publications (http://eos.wdcb.ru/icsu/navigator/navy.htm):

  The main goal of this project-to create representative source of information on the primary scientific publications in all disciplines covered by the International Council for Science (ICSU).

  The database basically will include descriptions of and links to materials, which are considered as primary scientific publications, i.e. scientific journals, serials and other relevant publications, which are published or approved by ICSU bodies as publications containing real science. The latter is the main criteria for including a publication into the ICSU Navigator database. The ICSU Press recommended focusing the proposal in more narrow scientific areas and soliciting support and advice from the relevant Unions. The recommended areas are Geophysics (with further transition to Earth Sciences) and Physics.

7. Conclusion

Gateways, in every discipline and subjects help the users in Academic Libraries where the thrust is on cutting edge technologies. Gateways serve as a ready reference tool. It is hoped that this compilation will be useful to the departmental needs. While studies of such nature can never be all comprehensive and complete, attempts should be made to narrow down on the resources and services, thus helping the users. Updating of resources in such compilations are absolutely necessary, only then will this gateway be of importance and relevant.

8. References

1. Desire - Development of a European Service for Information on Research and Education, http://www.desire.org/
2. DutchESS - Dutch Electronic Subject Service, http://www.konbib.nl/dutchess/
3. EEVL - The Edinburgh Engineering Virtual Library, http://www.eevl.ac.uk/
4. The Finnish Virtual Library Project, http://www.uku.fi/kirjasto/virtuaalikirjasto/
5. IMesh, http://www.desire.org/html/subjectgateways/community/imesh/
6. NMM Port, http://www.port.nmm.ac.uk/
7. OMNI - Organising Medical Networked Information, http://www.omni.ac.uk/
8. PINAKES - A Subject Launchpad, http://www.hw.ac.uk/libWWW/irn/pinakes/pinakes.html
9. SOSIG - The Social Science Information Gateway, http://www.sosig.ac.uk/
10. http://www.academicinfo.net/refdirectories.html

About Author



Mr. Radheshyam Yadav is Information Management Administrator in Cairn Energy India Pty. Ltd, Surat. He holds B.A (Econ), BLISc, MLISc, PGDCA, PGDLAN. He worked with Gujarat Vidyapeeth, PRL, INFLIBNET Centre, Ahmedabad as Library trainee. He worked as a librarian in HCPDPM, Ahmedabad, 2003. He has attended several Seminars, Conferences, Workshops, etc and published several papers in international and National Level. He is a lifemember of ILA, IASLIC and SIS. His area of interest is Library Automation and Digitization.
E-mail: shyamtyadav@yahoo.co.in

Students Attitudes Towards Digital Resources and Serivces in B.I.E.T., Davanagere : A Survey

M S Lohar Mallinatha Kumbar

Abstract

To evaluate the use of Digital Resources and services in Bapuji Institute of Engineering and Technology (BIET) college library in Davanagere (Karnataka). a survey of 100 Under Graduate, Post Graduate and Research Students of different branches was conducted by administering questionnaire. The analysis of the collected data covers the digital resources and services and how the digital resources improve carrier of the students and also deals with the problems that are faced in using the resources and services. Finally, it concludes that the main intention of the use of digital resources and services has been the academic interest of the students.

Keywords: Digital Resources, Survey, Academic Library

0. Introduction

With advancement of technology the libraries are moving towards digital resources, which are found to be less expensive and more helpful for easy access of information. In the digital era the commonly available digital resources like CD-ROMs, Online Databases, Library OPACs and Internet etc., which are replacing the print media.

At present the Karnataka state has more than 110 engineering colleges, which are affiliated, to Visveswaraiah Technological University. The Bapuji Institute of Engineering and Technology (BIET) is one of the prestigious engineering college established in the year 1979-80 with the initiation and foresight of the members of the Bapuji Educational Association Davanagere. This institute has 12 under graduate and 5 post-graduate programmes in engineering, technology and management. Research activities are also undertaken in all the departments in this institution. At present this institute is having more than 2192 students from different parts of the country and a few foreign students are studying here. The BIET Library has 31,217 volumes, 210 Indian & foreign Journals, DELNET databases searching and Internet facilities etc. are available in this library.

1. Scope and Limitation

The study is confined to the Under Graduate, Post-Graduate and few research students of BIET College Davanagere. The scope of the study is limited to the use of digital resources and services and its aim is to fulfill the academic needs of the students. It covers the available Digital Resources and services in BIET Library Davanagere.

2. Objectivs of the Study

To know the availability of different types of digital resources and services in BIET library.

- To know how the students are using the different types of digital resources.
- To know the purpose and utilization of the digital resources services by the students.

- To find out the hindrances faced by the students while accessing and using digital resources.
- To study the impact of digital resources over the traditional one.
- To suggest suitable recommendations to improve the digital resources and services for the benefit of students.

3. Methodology

A questionnaire was designed to elicit the opinion of the students. These were distributed among the different branches of under graduate, Post Graduate and few Research students. The collected data was further supplemented by informal discussions with the students. The analysis and interpretation of the data is presented in the subsequent sections.

4. Analysis

4.1 Semester wise Responses received from the students

Table 1 shows that, of the total 30.00% of the 7th & 8th semester followed by 25% of 5th & 6th semester and 15% of 3rd & 4th semester students are responded. Of the total 7.00% of each previous and final year P.G students and 4.00% of research scholars have responded the queries.

Table 1
Semester wise Responses received from the students

Semester	No of questionnaire issued	No of responses received	% of total responded
1st and 2nd	15	12	12.00
3rd and 4th	20	15	15.00
5th and 6th	30	25	25.00
7th and 8th	45	30	30.00
P.G Previous year	10	7	7.00
P.G Final year	10	7	7.00
Research Students	5	4	4.00
Total	135	100	100

5.1.1 Sex wise distribution of student's responses

Table 2 reveals that 65.00% of different branches of male and 35.00% of female students have studying in this college. This is a glaring example of sex-ratio imbalance of student's community.

Table 2
Sex wise distribution of respondents

Sex	No. of responses	Percentage
Male	65	65.00
Female	35	35.00
Total	100	100.00

5.1.2 Frequency of visit to Library by the students

The following table 3 shows that, of the total (100) respondents 30 (30.00%) of the students visit their library 'everyday' followed by 24.00% visit 'once in a week' and 15.00% of each student respondents visit '2-3 time a day' and 'once in a month' respectively. Whereas, 3.00% of the respondents 'never visit' the library.

Table 3
Frequency of visit to library

Frequency	**No. of Responses**	**Percentage**
Everyday	30	30.00
2-3 times in a day	15	15.00
Once in a week	24	24.00
Once in a month	15	15.00
Occasionally	13	13.00
Never	3	3.00
Total	100	100.00

Students opinion about availability of digital resources

Table 4 shows that, the availability of digital resources in the library. Of the total responses 55(39.28%) of the responses indicate the library has 'Internet' facility followed by 25.00% indicate 'CD-ROMs' are available in their library. 7.14% of each student's responses shows 'OPAC' and 'online databases' are available in the library respectively.

Table 4
Students opinion about the availability of digital resources

Type of digital resources	**No. of responses**	**% of responses**
CD-ROMs	35	25.00
Internet	55	39.28
OPAC	10	7.14
Online Database	10	7.14
E-journals	18	12.85
E-books	12	8.57
Any others	00	00
Total	140	100.00

*It is a multiple-choice question. *

5.2.1 Frequency of use of digital resources by the students

The frequency of using digital resources by the students is shown in the Table 5. Of the total respondents 30.00% use CD-ROMs 'everyday' followed by 23.00% use 'once in a week and 4.00% of the respondents 'never' use the CD-ROMs. Majority of the 43(43.00%) respondents use Internet 'Everyday' followed 25.00% use '2-3 times a day' and very negligible percentage (2.00%) of the respondents 'never' use Internet.

However, 30.00% of the students use OPAC 'once in a month' whereas, 27.00% use 'once in a week' and less percentage of (8.00%) the respondents 'never' use OPAC. Again 30.00% of the students are use online database 'once in a month' and 23.00% use everyday'. Less percentage of (7.00%) of respondents 'never' use online database.

Of the total 21.00% of the respondents use E-Journals 'once in a month' whereas, 20.00% use 'everyday' and 20(20.00%) respondents 'never' use e-journals. Majority of respondents 30(30.00%) use e-books 'once in a month' and 25.00% use 'everyday' followed by 24.00% 'once in a week' and 9.00% of the students 'never' use e-books in the library.

Table 5
Frequency of use of digital resources by the students

Frequency	**CD-ROM**	**Internet**	**OPAC**	**Online data**	**E-Journals**	**E- Books**
Everyday	30 (30.00)	43(43.00)	20(20.00)	23(23.00)	20(20.00)	25(25.00)
2-3 times a day	19(19.00)	25(25.00)	15(15.00)	12(12.00)	12(12.00)	15(15.00)
Once in week	23(23.00)	22(22.00)	27(27.00)	18(18.00)	18(18.00)	24(24.00)
Once in a month	15(15.00)	5(5.00)	3030.00)	30(30.00)	21(21.00)	30(30.00)
Occasionally	9(9.00)	3(3.00)	00(0.00)	10(10.00)	9(9.00)	15(15.00)
Never	4(4.00)	2(2.00)	8(8.00)	7(7.00)	20(20.00)	9(9.00)
Total	100	100	100	100	100	100

5.2.2 Student's opinion regarding the use of Internet service in the library.

Majority of the respondents (86.00%) use Internet service in the library followed by 14.00% of the respondents use Internet out side the library.

5.2.3 Students opinion regarding Preference of search engines

The Table 6 exhibits that, of the total 50.00% of respondents prefer to use 'Google' followed by 25.00% 'Yahoo' and 10.00% each of the respondents prefer to use 'AltaVista and MSN' search engines respectively. Only 5.00% of respondents indicate they prefer 'Askeeves'.

Table 6
Preference of using search engines

Search Engines	**No. of Responses**	**Percentage**
Google	50	50.00
Yahoo	25	25.00
AltaVista	10	10.00
MSN	10	10.00
Askeeves	5	5.00
Total	100	100

5.2.4 Purpose of using digital Resources by the students

Table 6 reveals that, Majority 60(34.09%) of responses indicate that the digital resources are used 'for finding relevant information in their subject' followed by 21.59% responses indicate for there 'research purpose'. However, 34(19.31%)of respondents using 'for their carrier development' and only 15(8.52%) of respondents are using digital resources for their communication purpose.

Table 6
Purpose of using Electronic Resources

Purposes	**No. of responses**	**% of responses**
For research purpose	38	21.59
For communication	15	8.52
To collect relevant information in my subject	60	34.09
Update the subject knowledge &General knowledge	29	16.47
For career development	34	19.31
Total	176	100.00

* It is a multiple-choice question. *

5.2.5 Learned to use electronic resources

Table 7 shows that, the most popular method of acquiring the necessary skill to use digital resources. Majority of 38(31.93%) respondents learned through 'Trial and error' method, followed by 27(22.68%) respondents are taking 'Guidance from Library staff' and 20(16.80%) respondents learned with the help of 'Computer department staff'. Less number of 15(12.60%) respondents learned from 'External courses'.

Table 7
Learned to use electronic resources

Learned to useElectronic resources	**No. of responses**	**Percentage**
Trial and error	38	31.93
Guidance from the Library staff	27	22.68
Course offered by the Institutions	16	13.44
Guidance from departmental staff of computer science	20	16.80
External courses	15	12.60
Any other (please specify)	3	2.52
Total	119	100.00

Multiple-choice question

5.2.6 Students opinion towards adequacy of information in digital resources

Table 8 shows that, Majority 60(60.00%) of respondents indicate the information available in the digital resources is 'Always' adequate followed by 35(35.00%) indicate 'some time' and 5(5.00%) of respondents felt that, the information available in the electronic resources is 'never adequate'.

Table 8
Adequacy of information in digital resources

Opinion	No of respondents	Percentage
Always	60	60.00
Some time	35	35.00
Never	5	5.00
Total	100	100.00

5.2.7 Hindrances in accessing the digital resources

Table 9 shows the opinion of the respondents regarding hindrances in accessing the digital resources. Majority 44(36.36%) of the respondents stated that, 'too much information is retrieved' is the main barrier to use digital resources, followed by 30(24.79%) opined that, 'time consuming' and 19(15.70%) felt that 'lack of IT knowledge to effectively utilize the services'. Only 15(12.39%) respondents stated that, 'limited access to computers' is the main barrier to use digital resources.

Table 9
Hindrances in accessing the digital resources

Hindrance	No. of responses	Percentage
Too much information is retrieved	44	36.36
Time consuming	30	24.79
Lack of IT knowledge to effectively utilize the services	19	15.70
Using digital resources often distracts from doing work	9	7.43
Limited access to Computers	15	12.39
Any other (please specify)	4	3.30
Total	121	100.00

It is a multiple-choice question

5.2.8 Students opinion regarding Impact of digital resources on their academic career

Table 10 shows that, Majority 42(30.65%) of respondents stated 'access to a current up to date information' is a benefit of using digital resources. Similarly 34(24.81%) expressed 'faster access to information' is the advantage and 32(23.35%) indicate 'easier access to information' is the benefit to develop the academic career of the students.

Table 10
Impact of electronic resources on academic career

Category	**No of response**	**Percentage**
Access to a current up-to-date information	42	30.65
Easier access to information	32	23.35
Faster access to information	34	24.81
Access to a wider range of information	29	21.16
Any other (Please Specify)	00	00
Total	137	100.00

* It is a multiple-choice question*

5.2.9 Problems faced while using digital resources

Table 11 Shows that, 30(22.72%) of respondents have faced problem of 'lack of timing', followed by 18(18.93%) of students indicate 'Lack of ' Hardware' is the main problem while using electronic resources. 10(11.11%) and 16(17.78%) of respondents are facing 'lack of Hardware and Software' problems respectively.

Table 11
Problems faced while using digital resources.

Problems	**No of response**	**Percentage**
Lack of Hardware	25	18.93
Lack of Software	20	15.15
Lack of training	22	16.66
Lack of information on digital resources	20	15.15
Lack of operating funds	15	11.36
Lack of timing	30	22.72
Total	132	100.00

* It is a multiple-choice question. *

5.2.10 Standard of academic work without digital resources.

The respondents were requested to indicate the standard of academic work without digital resources for finding required information. Table 12 shows that, majority 57(57.00%) of respondents 'Agreed' that, their academic work would suffer, without digital resources followed by 25.00% of the respondents have 'not agreed' and less percentage (18.00%) of respondents have not given their opinion.

Table 12
Standard of academic work without digital resources.

Opinion	No. of respondents	Percentage
Agree	57	57.00
Not agree	25	25.00
Don't know	18	18.00
Total	100	100.00

6. Findings of the study.

- Of the total (100) respondents 30 (30.00%) of the students visit their library 'everyday' and 55 (39.28%) of the respondents indicate their library has Internet facility and 25.00% indicate CD-ROMs are available in their library.
- Of the total 30.00, 43.00% and 23.00% of the respondents use CD-ROMs Internet and online database 'Everyday' respectively. However 27.00% of the students are using OPAC' once in a week'. 20.00% and 25.00% respondents use e-journals and e-books 'everyday' respectively.
- Majority of the respondents (86.00%) is using Internet service in their library. at the same time 50.00% of respondents prefer to use 'Google' search engine.
- Majority 60(34.09%) of the respondents are using digital resources 'for finding relevant information in their subject' field.
- 38(31.93%) of the respondents learned through 'Trial and error' method of acquiring the necessary skill to use digital resources.
- 58(58.00%) of the respondents indicate the information available in the digital resources is 'Always' adequate at the same time 44(36.36%)of the respondents stated that, 'too much information is retrieved' is the main barrier to use digital resources.
- Majority 42(30.65%) of respondents stated 'access to a current up to date information' is a benefit of using digital resources. Similarly 34(24.81%) expressed 'faster access to information' is the advantage.
- Majority 30(22.72%) of respondents have expressed 'lack of timing', is the main problem while using electronic resources.
- 57(57.00%) of respondents 'Agreed' and 25.00% of the respondents did 'not agree' that, academic work would suffer, without digital resources for finding their required information.

7. Suggestions

Based on the findings of the study the following suggestions are recommended to improve the use of digital resources among the students.

1. The authority must conduct training programmes for students regarding how to use the digital resources effectively.
2. Awareness should be created to use e-journals and e-books to obtain current information.
3. More computer/terminals should be installed in the library for the benefit of the students.
4. Authority should provide more funds to acquire the digital resources for the benefit of the users.

8. Conclusion

The digital resources play a vital role in all the field of human life. These have rapidly changed the way of seeking and disseminating information. It is clear from the study that, the students of BIET College have developed their academic career. The speed of availability and the ease of accessibility of information make the students to use digital resources more frequently. This study helps the librarian to know the importance of digital resources in academic environment.

9. References

1. Ray (Kathryn) and Day (Joan): Student attitudes towards electronic information resources. Information research. Vol (2), 1998.
2. Kumbar (Mallinath) and Praveena (J.K.) and others: Use of Electronic resources by Research Scholars in CFTRI Mysore: A Study, Paper presented at 4th ASSIST National Conference held at Dept. of Library and information science Kuvempu University B.R Project.
3. Kumbar (Mallinath) and Lohar (M.S.) (2002): Use of library facilities and information resources in Sahyadri colleges, Shimoga (Karnataka): A Study. Annals of Library and Information Studies Vol 49, (3); p73-83.
4. M.S.Lohar & Roopashree. T.N. Uuse of Electronic Resources by faculty members in B.I.E.T, Davanagere: A Survey, Paper presented at NACLIN Conference held at Pune University 2004 on 23-26 Nov.

About Authors

Mr. M S Lohar working as Librarian at University B.D.T.College of Engineering, Davangere and persuing his research from Mysore University. He has presented number of papers in seminar, conferences and journals. He is also a member of many professional bodies.
Email : manjunath_lohar6@hotmail.com

Dr. Mallinatha Kumbar woking as Reader in Department of Library and Information Science, University of Mysore, Manasa Gangothri, Mysore. Prior to this he was working as Assistant Librarian in Kuvempu University.. He has presented number of papers in seminar, conferences and journals. He is also a member of many professional bodies.

Digital Libraries and Services

K Paulraj P Balasubramanian
S Kanthimathi

Abstract

The digital library presents an opportunity for the traditional library and the librarian. Libraries and librarians have significant expertise to offer the creators of digital libraries, for example, skills in organising information.Traditional libraries will not be replaced by digital libraries, certainly not in the near future, and probably never. The reality will be hybrid libraries, with librarians needing to be able to operate in two environments: the first, physical libraries and physical collections; and the second, digital information resources. The goal will be to satisfy the information needs of the user community, regardless of whether the information needed is available in digital form or physical formats.

Keywords : Digital Library, Information Services

0. Introduction

Digital libraries aim to provide access to information 'an demand/ regardless of location of the computer in, which it is stored. In this article we examine key features of digital libraries and services.

The term library usually invokes in our mind a storehouse of information in the form of print on paper publications like books, Journals, Magazines, Newsletters and reports etc and newer media such as films, filmstrips, video and audio cassettes. Most of us view the library as place where information are acquired, organised, shelved and retrieved. Move recently, many libraries have again to use online data base- search-systems like DIALOG and connections to the online public access catalogues (OPACs) of other libraries, over telecommunication. Networks, many libraries have also taken advantage of CDROM technology to provide users networks access to large information bases. With such remote information access, the walls of the library began to be less solid

The internet and the world wide web (WWW) technologies are providing the technological environment and intellectual impetus for the development of 'digial libraries' libraries without walls, with data and ideas. The internet has enabled global connectivity of computers and the development of various tools and techniques for networked information provision and access. Starting with basic tools like e-mail (Meeaging) FTP (File Transfer Protocol) and telnet, the internet has progressed to provide user friendly tools like, Gopher, WAIS and the WWW for information publishing and access. The world wide web, which is integrating all other access tools, also provides a very convenient means for publishing and access, The world wide web, which is integrating all other access tools, also provides a very convenient means for publishing And according to multi media hypertext linked documents stored in computers spread across the world.

1. What are Dirigal Libraries ?

Digital libraries are an evolving area of research, development and application and multiple definitions have been offered by workers in this area based on common aspects among these definitions. digital libraries may be defined an electronic information collections containing large diverse repositories of digital objects, which can be accessed by a large number of geographically distributed users, Such repositories would exist in locations physically near or remote from the users. Digital objects include

text, images, maps, videos, catalogues, indices, scientific, business and government data sets as well as hyper textual multimedia compositions of such elements.

There is also general acceptances of fact that digital libraries would need to span both print and digital materials for the foreseeable future print on paper publications are expected to be around and thus digital libraries are expected to provide integrated, a coherent access to both types of maters.

Key components of digital libraries are therefore.

- Geographically distributed digital information collections.
- Geographically distributed users.
- Information represented by a variety of digital objects.
- Large and diverse collections
- Seamless access

2. Digital Electronic and Virtual Libraries

A source of confusion in this area has been the use terminologies like 'virtual' 'digital' and ' electronic libraries. One person's digital library is often another's virtual library' and some useful distinctions have recently been made.

- Electronic library : A library that provides collections and or services in electronic form for example optical videodisk, CD-Rom, online etc.
- Digital Library : A library that does not physically exist, most often used to denote a library with distributed collections or services that appear and act as one typical example is a web site with pointers to other sites.

From the above it may be seen that electronic library is more inclusive than digital library. However, digital library has come to be the preferred term perhaps keeping in line with terms like digital audio and digital video. The current usage of the term 'digital library' appears encompass both electronic and virtual libraries.

3. Networked computer Science Technical Library

Let us take a look at one of the operational digital libraries. This is the networked computer science Technical Report library (NCSTRL). NCSTRL provides unified access to catalogue records and complete documents of computer technical reports stored indistributed around the world, through the world wide web. NCSTRL put into operation. The projects participation in NCSTRL is open to all academic departments awarding Ph.D. in various faculty and to research facilities of industry and government. Currently over 200 departments around the world are participating. NCSTRL can be accessed free of cost on internet using any world wide web browser.

NCSTRL provides a very simple search interface and allows search by author title or abstracts. A retrieved document may be viewed as an HTML document or Post Script.

The technology underlying NCSTRL is a network of inter opening digital library servers. The servers provide three services.

1. Repository services that store and provide access to documents.
2. Indexing services that allow searches over bibliographic records.
3. User interface services that provide the human front –end for the other services.

These services inter-operate using a protocol, enabling development of new kinds of services.

3.1 Digital Library Projects

Since the past couple of years the idea of 'digital library' has moved to the forefront of discussion and research., several digital library projects are currently underway in the U.S.A. Europe, Australia, New Zealand and Singapore.

4. Advantage of digital libraries.

Some of the key advantages digital libraries provide include the following

- ✍ Ability to search : The ability to search provides an enormous advantage of electronic materials, When an ASCII some years been replacing printed abstract Journals. Since most of the modern material is now produced via computers. It can generally be provided in ASCII and be searched. For those documents which are searched rather than read are giving way to CD-ROMs which are small, cheaper and for more effective.
- ✍ Ubiquity : Another key advantage is ubiquity – many simultaneous users can have access to a single electronic copy from great many locations copies can be delivered with electronic speed and it would be possible to reformat the material as per the reader preference.
- ✍ Support wider range of material : Digital storage also permits libraries to expand the range of material they can provide to their users. Since audio cassette tapes and records cannot stand a large number of playings without deterioration, their digital representation can produce a format that is much super and of better quality. Digital material can also permit access to video tapes and new kinds of multi-media materials that are created only on computers and have no equivalent in any traditional format.
- ✍ Access current information - For researchers, digital libraries provide access to up to date current literature and there by help them to be aware of current trends.

5. Some of the major areas of focus are

1. Multi media object storage, retrieval and transmission.
2. Data compression
3. Digitisation
4. Hypermedia Navigation
5. Authoring tools for creating electronic documents
6. Multimedia object representation
7. Meta data bases
8. Display technologies
9. User inter faces
10. Search, retrieval and routing software.

6. Issues in Digital Libraries

Work in the area of digital libraries has thrown up several issues and challenges key issues include the following

1. Copyright: It is very easy to copy, replicate message and distribute digital information, enforcing copy right in digital environment is a major issue.
2. Technological obsolescence

   - ✍ Hardware : The major risk to digital objects is not physical deterioration, but technological obsolescence of the devices to read them while the life time of optical and magneto-optical cartridges is expected to be in decades, those of reading devices are only about one decade.

   - ✍ Software : A more serious problem is software obsolescence. It has been pointed out that the variety of software formats far exceeds the number of hardware devices manufactured and that these programs come and go more quickly than the hardware does.

7. Who will Run the Digital Libraries

All sorts of organisations are claiming the right to be the provider of information to be desk-top online services, libraries, bookstore, publishing telephone companies, telecommunication officers university centers, and new startup companies in the digital networked environment, it appears that libraries have no unique claims providers of scholarly information. In practical terms, it is likely that provision of current

Material will move back to the publishers, who can have sufficient control of who reads what. It is possible that the publishers too will be bypassed, as the authors self publish on the net or through some new venture. Witness for instance the variety of courseware available mostly free over the internet.

8. Conclusion

Digital libraries are expected to bring about significant improvements over current modes of information publishing and access methods. Educators researchers and students across the world will be among the first to benefit from digital libraries particularly those in the developing countries.

9. References

1. DESIDOC Bulletin of Information Technology (DBIT) Special on digital Libraries Vol.17 No.6 Year 1997
2. SRELS (Sarada Ranganathan Endowment for Library Science) Journals of Information Management VOL. 40, No. 3 September 2003
3. Recent Trends in Library & Information Science Technology – Seminar 2002 – Proceedings of PG Department of Library and Information Science, Bishop Hebar College, Tiruchirappalli-620017
4. PGDLAN (Post Graduate Diploma in Library Automation & Networking) Course – Materials – 2004
5. 48th Indian Library Association- Electronic Information Environment – Seminar Proceedings – January 2003.

About Authors



Mr. K Paulraj presently working as Librarian in P M T College, Melaneelithanallur, Tamil Nadu. He has done MCom, MLIS, PGDLAN and doing PhD from Manaonmanium Sundaranar University, Tirunelveli. He is a life member of various professional associations. He has contributed 11 papers in seminars, conferences & journals.
E-Mail : paulsuresh2005@yahoo.co.in



Mr. P Balasubramanian presently working as Librarian in SCAD College of Engineering & Technology, Cheranmahadevi, Tamil Nadu. He has done MA, MLIS, PGDCA, PGDPR and doing PhD from Manaonmanium Sundaranar University, Tirunelveli. He is a life member of various professional associations. He has contributed 15 papers in seminars, conferences & journals.
E-mail: bala_phd2000@yahoo.co.in



Dr. S Kanthi Mathi is working as Librarian (Senior Grade) at Rani Anna Constituent College for Women, Tirunelveli, Tamil Nadu. She has done MLIS and PhD. She is a Research Guide for MPhil and PhD scholars at Manaonmanium Sundaranar University, Tirunelveli. She is a life member of various professional associations. She has contributed 15 papers in seminars, conferences & journals.

Consortia Developments in Library and Information Centres : Some Issues

Jayaprakash H M M Bachalapur

Abstract

Due to increase in the cost of journals, dwindling library budgets and proliferation of electronic information resources, libraries have been involved in cooperation, coordination and collaboration in resource sharing. The emergence of information technology made library professionals to change their role as navigator of information and come closer willingly to share available information with other libraries. Libraries and Information Services (LIS) are being transformed by technology; consequently, LIS have to adopt to meet their user's changing needs and growing expectations. Included among the resource sharing initiatives conceived by libraries in India is the creation of a computerized network or consortium of all LIS to achieve optimum use. This paper presents the consortia developments in the library and information centers in India.

Keywords : Consortia, Electronic Resources

0. Introduction

Consortia means alliance of institutions having common interests. The application Information Technology and Information Retrieval systems for on-line catalogues, bibliographic databases and full text electronic documents have facilitated a quick information exchange among the institutions.

Cooperation amongst the institutions for sharing their library resources is being practiced for decades. Traditionally, primary purpose of establishing a library consortium is to share physical resources including books and periodicals, amongst its members. However the mode of cooperation has gone under a transformation with infusion of new information technology from print – based environment to digital environment.

The academic libraries, being the nerve centers of higher education, teaching and learning; play an important role in support of all the activities of the concerned university. The increasing growth in the enrolment of students and researcher, lack of proper and adequate infrastructure further aggravates the overall problems, challenges for the academic libraries. This situation arised the need of consortia of digital libraries. University Grants Commission(UGC), All India Council for Technical Education (AICTE), and other government bodies of education are helping academic libraries to make automation and build their own consortia of libraries in their area.

1. E-Books

An e-book or electronic book is essentially a book that has been written in electronic format so that it can be read direct from the computer screen. An e-book is a text presented in a format which allows it to be read on a computer or handheld device. Many titles which are available in printed versions can be read as e-books, including best selling fiction, classics and reference texts. E-books are also used to make out – of – print work available, or to bypass print altogether, as with new works by aspiring authors. E-books can consist simply of the electronic text or may also contain such as audio, video or hyperlinks.

2. E-Journals

Electronic Journal may be defined as any journal, periodical, e-zine, web-zine, newsletter or type of electronic serial publication, which is available over the Internet and can be accessed using different technologies such as World Wide Web, gopher, telnet, etc. E-journals are periodical, regular or irregular moderated unit made available in an e-format either on a static medium or via computer networks.

3. Library Consortia

Library purchasing methods are currently undergoing considerable changes with the growth of library consortia around the world. The consortia are increasingly negotiating with publishers for access by their members to electronic journals and databases. Under the terms of the licences involved, members may often have access to all titles taken by each of the libraries or to all titles from participating publishers. The libraries within such consortia are often turning to their subscription agents for help in the administration and management of the licence arrangements with publishers, including the maintenance of up-to-date databases of publisher options and prices; the availability of individual publishers' licence agreements and forms; liaising with publishers regarding changes to the agreements, for example in relation to new title requirements or additional libraries joining the consortium; handling renewals of licences; acting as a distribution point for passwords and IP numbers and assisting with access, etc.

3.1 Aim of the Consortia

The primary objective of the Library Consortium is to encourage and facilitate interlibrary communication, education and resource sharing within its diverse multi-type library membership.

Today Consortium purpose is shifted from mere sharing of resources to sharing of expertise between libraries and also explores the need for libraries to make the most effective use of their funds collectively.

3.2 Need for Consortia

Academic (University & College) Libraries & Research Center Libraries with the impact of Information Technology are compelled to provide relevant information essential to its end users within a short time either from its in-house holdings or through Consortia. Inflation & Budgetary reductions are the primary force that brings the idea of consortia development.

3.3 Salient features of Library Consortia

1. To eliminate the different problems faced by the libraries to provide various services to the users.
2. To meet the thrust of information of the vast people due to rapid growth of population all over the world.
3. To cope up with the newly generated knowledge published in different forms, such as, printed and non – printed documents, electronic media on various disciplines, multi-disciplinary & new generated subject areas.
4. To collect all the documents published at the national and international level, because of the library financial crunch; &
5. To overcome the language barriers i.e.:- primary documents are being published by the developed countries like USA, UK, France, Japan etc, and among them the non-English speaking countries produce majority of scientific literatures in their mother languages.

6. Single payment by one of the participants or through an agent and license has to be signed by all;
7. The members are expected to maintain same level of subscription;
8. Initial minimum subscription was for 5 titles with 10% e-access charge but was reduced to two titles thereby increasing e-access charge to 12%;
9. Publishers found it convenient to negotiate with members through an agent and agent raising individual invoices to all members and single payment to publishers;
10. Institution-wise usage statistics to be provided to ascertain as to how often user's access to all titles subscribed.

3.4 Principles to Govern the Consortia

1. Flexibility to choose your own library management solutions vendor and select the member libraries with which you will share resources.
2. Flexibility to own, manage, and control your library's records and enforces its policies.
3. Flexibility to extend access to even more information with an information portal that shows your library's face.
4. Flexibility to share physical and digital resources.
5. Flexibility to enable your library users to search and place holds on the resources of your own and other member libraries and to enable users of other member libraries to search and place holds on your library's resources.

3.5 Functions of the Consortia

1. Collection Sharing.
2. Electronic Content Licensing.
3. Electronic Content Loading/Presentation.
4. Inter – Library Loan / Document Delivery.
5. Preservation.
6. Training.
7. Union Lists / Shared Online Catalogues.
8. Other; New forms of scholarly and scientific communication.

3.6 Benefits of Consortia:

1. Consortia-based subscription to electronic resources provides access to wider number of electronic resources at substantially lower cost;
2. The Consortium, with its collective strength of participating institutions, has attracted highly discounted rates of subscription with most favorable terms of agreement.
3. The Consortium is proposed to be an open – ended proposition wherein other institutions can join and get the benefit of not only highly discounted subscription rates by also the favorable terms of licenses.
4. The Consortium have been offered better terms of license for use, archival access and preservation of subscribed electronic resources, which would not have been possible for any single institution;

5. Since the subscribed resources would be accessible online in electronic form, the beneficiary institutions would have less pressure on space requirement for storing & managing print – based library resources.

6. The Consortium is expected to trigger remarkable increase in sharing of both the print and electronic resources amongst participating library.

3.7 International Coalition of Library Consortia (ICOLC)

The International Coalition of Library Consortia first met informally as the Consortium of Consortium in 1997. The coalition continues to be an informal, self-organized group comprising nearly 150 library consortia from all over the world. The coalition serves primarily higher education institutions by facilitating discussion among consortia on issues of common interest. And dedicated to keeping participating consortia informed about new electronic information resources, pricing practices of electronic provider community, providing a forum for them to discuss their offerings and to engage in dialog with consortia leaders about the issues of mutual concern.

4. Development of consortia

Library & Information Center's networking in India have come to the existence almost two decade ago. In India we used the term "Inter-library Loan" in the parlance of library management. This scenario has extended from its limited basis to become fully grown Consortia of various kinds. Since India is the developing country, we constantly work towards improving our infrastructure and technology to meet the demands of our scientists and researchers by taking the lead from the developed countries.

Now a day, in India there are few important Consortia's are established. Those are :

4.1 INFLIBNET- UGC

The INFLIBNET was established in 1991 by UGC, which aims to link 294 traditional universities, 300 research institutes and 14,000 colleges in the country.

It avoids duplication of journals and enhances active resource sharing through consortia with various publishers. As many as 50 universities are already connected and another 70 would be shortly cleared for e-journals subscriptions.

Negotiations are being under active consideration for consortia subscriptions of e – journals with various publishers/vendors/intermediaries.

4.2 Health Inter - Network (HIN) India: The Health Inter - network is a United Nations initiative created to respond to the challenges posed by the digital divide:

The Health Inter - network aims to support & strengthen public health services & to provide access to high quality, relevant & timely health information. It further aims to improve communication & networking among public health care workers, researchers & policy makers.

4.3 HELINET Consortia

The Rajive Gandhi University of Health Sciences has been working on collaboration & virtual resource sharing & in the past one year made a significant investment in hardware and content procurement for making International Journals and Databases available to its member libraries all over the Karnataka.

5. Indian Scenario in Consortia Activities

We have looked into necessity for consortia formation, cost factors of e – journals, pricing and licensing models. In Indian context, consortia formation started much later compared to many developed countries. To begin with, a small group of libraries started coming together and made headways for negotiating consortia terms and conditions so as to have access to large amount of information. As of now, we have a few consortia formed and each one is a model of its own as far as funding is concerned, viz:

- Same funding Agency, CSIR;
- Institute's Headquarters funding main/branch libraries- TIFR;
- Central Agency funding directly- MHRD:INDEST;
- Homogeneous group under central funding – IIM Libraries
- Different departments/homogeneous group initiated consortium – FORSA.

5.1 Model 1: Same Funding Agency – CSIR

CSIR and its 40 laboratories have successfully negotiated with a major publisher for consortium licensing covering their entire database, making payment from headquarters for access to e-journals. This is a large consortium formed during 2000 after prolonged discussions keeping in view various parameters from the point of view of pricing, viz. number of subscriptions, number of subscribing laboratories, number laboratories not subscribing, print based price, add-on to e-access, access fee for non-subscribers and host of other parameters, which will not qualify for model keeping in view large number of parameters vis-à-vis spread of laboratories across the country, absence of facilities in some laboratories, heterogeneous groups, etc. With cooperation and willing support of some dedicated library professionals, the consortium went through successfully (Goudar,I.R.N., 2003). It is to be seen now, while renewing license, how potentially all 40 laboratories are making use of the database subscribed with cross e-access to large number of journals. For renewal, internal review has to be carried out to assess usage pattern from all laboratories and justify the amount spent.

5.2 Model 2: Headquarters funding main/branch libraries-TIFR

TIFR and its five branch libraries located at different locations in the country have gone into formation of consortium with major publishers. Headquarters arranged payments after negotiations and after successful completion of one year and its smooth running renewed for the second year. All the members and users are happy with arrangement for accessing important journals among themselves. The publishers should also be happy for the reason that with little administrative efforts, negotiations/payments were settled.

It was a win-win situation for both the parties, since the negotiation was based on print subscription by member libraries and one of the branch libraries has the advantage of accessing entire offer of journals and not subscribing a single title from the concerned publisher. This could be an ideal multi-site model, where administrative /payment aspects are handled by parent organization.

5.3 Model 3: Central Agency funding directly - MHRD-INDEST

An Indian National Digital Library in Engineering, Science and Technology (INDEST) was set under MHRD, Government of India, initially covering 38 major technological institute including IISc, all IITs, NITs, RECs and IIITs. It is an open ended system with provision for adding new members by shared subscription through a consortium of libraries. It hopes to increase access to e – journals and important databases on negotiation with major publishers.

Keeping in view, what science and technology libraries used to spend annually, INDEST investment is much less and could provide comparable or even better facilities of information sharing with six full text electronic resources and five on-line databases. The institutions availing INDEST have been divided into three broad categories, viz. Category I: IISc and all IIITs, Category II: RECs, NITs and all rest in category III (Sen, N. 2003; Arora, J., 2003).

There are other 'nets' getting ready or already swung into action in forming consortia, viz. UGC, ICAR, etc and small scale level- FORSA and DAE libraries. In due course, INDEST should be able rope all parallel consortia into its fold and make a truly national level consortium to negotiate national site license for all multidisciplinary areas.

5.4 Model 4: IIM Digital Library Consortium

All IIM libraries have been striving for resource sharing in areas such as cooperative acquisitions, processing and decentralized utilization. IIM libraries consortium has been in existence since last three years and negotiated for acquisition of databases and electronic journals.

INDEST Steering Committee meeting held on 18 April 2003 considered favorably to constitute Special Interest Group for IIM Library Consortium and thereby taking care of consortium based resource requirements of all management schools of INDEST members. The Group has recommended that the SIG: Management Schools to be known as "Electronic Resources for Indian Management Schools – ERIMS" and looking forward to subscribe to electronic resources(Jambekar,2003).

5.5 Model 5: Different departments/Homogeneous Groups-FORSA

This is yet another model, wherein Institutes are affiliated to different departments of central government. The model envisaged from this group is briefed in nutshell reflecting how library professionals come together willingly and support for consortia formation.

Unlike others, this group has an informally established forum, which needs a briefing to reflect how homogeneous, like minded professionals come together for cooperation, coordination and collaboration in resources sharing and initiating need based consortia formation in the changed environment.

6. Forum for Resource Sharing in Astronomy and Astrophysics (FORSA)

In early 1980s, due to proliferation of information, library professionals working in institutes where astronomy and astrophysics was one of the main thrust areas of research, felt the need to come together and to establish a forum, which can act as a spring board for sharing and exchange of information. As result FORSA was informally launched on July 29, 1981 with a mission to: compile union catalogue of scientific serials, annual and other irregular publications, reference tools, recent research in astronomy, books on order and thesis holding, holding duplicate issues of journals, directory of libraries and facilities available in each library.

At present there are eleven members in FORSA, viz. ARIES, Bose Inst., CASA-OU, HCRI, IIAP, IUCAA, NCRA (TIFR), PRL, RRI, SINP, and TIFR. For details of FORSA, one could look into- http://www.iiap.res.in/library.forsa.html.

FORSA has gone into formation of two consortia, viz. Indian Astrophysics Consortium – IAC (Kluwer Journals 2002+) and FORSA Consortium for Nature On-Line (Nature Publishing, 2002+) keeping in view the following points:

- It is a voluntary consortium with shared goals- being one of several types of consortia;
- Governed by discussion/consensus among participant library professionals;
- We started with nothing but good will and shared goals, without staff support, no office; and one of the participant has become the Coordinator for dealing with all FORSA matters;
- We have 'sunset' clause, i.e. review every three years for IAC and on-line single title every year for renewal keeping in view everybody's concern and experiences of the past years.(Patil,Y.M 2004)

6.1 Problems encountered

- It was a maiden venture for the group. We did not have guidelines or models to follow;
- We have to believe the middleman, who is expected to act on our behalf;
- The agent added one more member to our group, whose titles were outside the group thus diluting the objectives of the consortium. As a result, the consortium was burdened with other titles with marginal interest to one or two members;
- This kind of incidence could have been avoided if members were aware of consortium guidelines and a formal committee to decide the membership of the consortium or FORSA;
- The problems faced should signal FORSA members to formalize the formation and functioning of the consortium negotiations in the future;
- Usage statistics was provided on institutional basis rather than title-wise and now with the introduction of COUNTER, one can

7. Conclusion

Purpose of Library Consortia is to control and reduce information costs, improve resource sharing develop a networking information environment & share licensing issues with each other. But now a days number of consortium's are coming up with multidisciplinary subjects in India.

Libraries have an ongoing responsibility for collection development, preservation, retrieval of information from paper – based resources and now it has become more complex with the introduction of digital resources & with the help of Information Technology. Though it is late, there is a time to reinitiate consortium movement especially by the automated and semi-developed libraries attached to big libraries to acquire maximum resource & service involving minimum time, money, space etc and serve the users community at an optimum level. It is very much difficult for a single library to satisfy the needs of its users in digital environment. Keeping in view the old traditions and applying them to the new environment will make institutions grow & provide useful service to the user community. In the near future, all the libraries should be a partner of the one or the other Consortium.

8. References

1. Arora, J; National Digital Library in Engineering, Science and Technology (INDEST): A proposal for strategic cooperation for consortia based on access to electronic resources. Intl. Inform. & Libr.Rev.,35,p.1-17,2003
2. Goudar, I.R.N; Consortia models for e-journals and Indian initiatives: 9 th Prof. M.R.Kumbhar Memorial Lecture, SRELS, Bangalore, 16th April,2003.

3. Jambekar, A., Pandian, P. and Gupta, D. K; Partnership for success: a case of IIMs libraries' consortia model. http://www.ala.org/work/international/jambekar.html

4. Patil, Y. M; Managing change: Consortia efforts in IT Environment in Library and Information Profession in India: reflections and redemptions. New Delhi: B.R.Publications, 2004, p.465-486.

About Authors



Mr. Jayaprakash H. has completed his M.A., M.L.I.Sc., M.Phil., PGDLAN degrees from different universities. Having 10 years experience as a Librarian at Medical, Engineering, Nursing Libraries. Presently working as a Librarian at National College of Pharmacy, Shimoga, Karnataka State. Published / Presented 8 research articles in different National Conferences.
E-mail: jaypee_prakash@yahoo.co.in



M M Bachalapur presently working as Sernior Librarian in Kalpataru Institute of Technology , Tiptur-572 202, Karnataka. He has completed B.Sc., M.L.I.Sc and he is doing Ph.D. from Karnataka University. He has over 13 years of professional experience in Library and Information Science. Published 14 technical papers in National Conference/Seminars. Member of ILA, IASLIC and KALA. His areas of specialization include; networking and library automation.
E-mail: bachalapur@yahoo.com

Enhancing Network Applications in a University Library : A Case Study

Suresh Jange
Amruth Sherikar
R B Gaddagimath
S B Policegoudar

Absract

Information and communication technology is transforming the society, education, business and the economy. The Library managers must understand these changes in order to position their organizations to flourish in the networked environment to provide effective information services to the users. In the Local Area Network(LAN) world, an attempt has been made to describe the network establishment in Gulbarga University Library and their experiences to enhance the techno-based services to the users. Besides, network architecture with optical fiber, hardware and software requirements for the effective implementation of LAN in the library are explained in detail. Further explores steps taken to establish CD-ROM databases; Internet Lab in accessing electronic resources under UGC Infonet, status of DBMS, library web page and Institutional productivity. Future plan for establishment of Learning Resource Centre, Provision of video camera, retrospective bar coding of collections and touch screen in the library are highlighted.

Keywords: Library Automation, Digital Library, Networking, IT based Services

0. Introduction

'Without vision, the people perish' we read in the Bible, and in turbulent times, vision becomes ever more important. The enormous changes being forced upon librarians as they seek to harness the possibilities of electronic information without being overwhelmed by its quantity, variety and transience make it imperative that we have a clear vision of the future we are trying to create. As librarians we are forced to ask ourselves whether we are to pursue the technological imperative to its logical conclusion, so aiming to become the masters of cyberspace (Brophy, 1997). The outlook and taste of users have changed enormously and they become more information conscious than ever before, demanding the need-based information timely to meet their nascent needs. This could be achieved by adopting Information and Communication Technology in library by making best use of it in extending timely needs of users.

In the contemporary environment, characterized by rapid technological developments and the proliferation of information resources, information literacy empowers users to access and use resources more effectively and efficiently. It is the first and foremost duty and obligation on the part of Information Managers to make provision for Information access by optimizing the information technology through enhancing network applications especially in a university set up. In this new networked information age 'we are seeing the emergence of a web of inter-organizational trust relationships in support of … information access, implemented and expedited through new authentication and access management systems (Lynch, 1999).

In India, academic and research Libraries have made a significant investment in the information technology gadgets and associated technologies for network establishment in library to provide technology based services to the users and automate the in-house activities. A matter of concern has been how best the Library and Information Service Managers can exert significant influence as experts in locating, identifying, organizing, maintaining, compiling, and providing access and evaluating information resources using network skills to enhance efficiency and build the image of librarianship for survival and visibility.

An attempt has been made in this paper to notify the significance of network applications in a university library and examines strategy adopted for establishment of Local Area Network in the Gulbarga University Library to improve the information access to the user community by incorporating the strong technological applications.

1. Information Technology in Indian Universities : A Scenario

Advancement of technologies such as dramatic increase of digital storage media, the convergence of telecommunication and broadcasting, the availability of wealth of resources accessible through internet and reducing cost of computers with multimedia and Internet capabilities will certainly have major influence to support information needs of library users. There are also an exponential growth rate of electronic publications either online or in CD-ROMs. The emergence of Internet and its worldwide Webs expose libraries to the wealth of global information resources available in major libraries and institutions worldwide.

Thus, it is an established fact that, Information technology promotes equity, quality based education and information content at their fingertips. The extent of Information technology being used in Indian universities is depicted in below table (Shafi, Z. 2002*)

Table - Availability of IT in Indian Universities

Category	**Universities**					**Total**
	Conventional	**Deemed**	**Non-Agricultural**	**Agricultural**	**Open**	
Computers						
>1000	1	-	1	-	1	3
500-999	1	1	1	1	-	4
250-499	11	2	-	1	-	14
50-249	13	3	5	8	-	29
<50	6	2	5	5	2	20
LAN	18	6	11	12	1	48
Internet	24	6	11	12	1	54

Information about the availability of IT infrastructure in Indian universities is shown up to 2001, but there has been tremendous spurt in adoption of IT in Indian academic and research library from 2002-2004, which is a good sign for the university libraries to strengthen their infrastructure to meet the growing information needs of user community.

2. Gulbarga University Library : A Pioneering Center of Excellence

Gulbarga University Library is the heart of the university catering to the academic and research information needs of the user community has now established Local Area Network (LAN) comprising of twenty computer systems using SQL Server with dedicated library software SOUL (Software for University Libraries). This has facilitated the staff to execute their routine library activities using computers, avoiding manual processing mainly viz. Purchase of Books/ Journals/ electronic materials, placing orders, renewal, issue/ return of documents and rendering e-services i.e. searching library catalogues, bibliographical services, Current Awareness services, Content Alert services and monitoring of the work in progress of the library in-house activities and services.

2.1 Client Server Network

A well-planned Client/server technology creates a powerful environment that offered for storing and managing data has been established in the library. The server receives the structured requests from the clients, processes them, and sends the requested information back over the network to the client. This network is the most efficient way to provide:

- Database access and management for applications such as spreadsheets, accounting, communications, and document management.
- Network management.
- Centralized file storage

Because the file services and the data are on the back-end server, the servers are easier to secure and maintain in one location. Data is more secure in a client/server environment because it can be placed in a secure area away from users. The data is also more as the multi user operating system used is Windows 2000 Advance Server to prevent unauthorized access to files.

The SOUL (Software for University Libraries), a dedicated library software is being used in the library for executing library activities and services with SQL 7.0 version, as a database server referred as the server back end.

The database query is sent from the client, but processed on the server and the results are sent across the network back to the client, which involves:

1. The request is translated into SQL.
2. The SQL request is sent over the network to the server.
3. The database server carries out a search on the computer where the data exists.
4. The requested records are returned to the client.
5. The data is presented to the user.

2.2 Ensuring Hardware and Software Compatibility

In today's computer industry, hundreds of manufacturers develop hardware and software and evaluating and selecting hardware and software is a major part of planning for network implementation. Keeping in view of future up-gradation and compatible, the following hardware and software's have been procured for the library.

IT Infrastructure in the Library	
Hardware's	**Quantity**
HP TC 2120 Main Server Hyper Threading Support @ 3 Ghz	1
Computers Systems	41
Compaq P-IV Hyper Threading Support - 20	
PCS P-IV - 10	
HCL P-III - 10	
P-I - 1	
Scanners	2
Switches (24, 16, 8 Port)	4
Printers	10

Software's	
SOUL (Library Automation Software)	1
SQL 7.0 version	1
Windows Advance Server 2000	1
MS-Office 2003	1
Data Protection and recovery	4
McAfee Firewall	1

2.3 Network Architecture Layout

The establishment of LAN architecture layout in the library (Figure 1) depicts that, E-Cat 6 cable has been used with D-Link switches to connect various nodes in the library. The load time for SOUL operation is about 4 seconds, which has lead to enhance the speed of access in the first floor by drawing Optical Fiber Cable instead of E-Cat6 cable, is the significance of this LAN network established in the library. The LAN Server is housed in CD-ROM Lab and necessary provision for OPAC Nodes at the entrance (3 Nos.), Reading Hall (2 Nos.) and Serials Control (2 Nos.) have been made in the library, with only searching options opened for the users. While Internet server is connected through V-SAT drawn from the Computer Centre of the university through Optical Fiber Cable for accessing e-resources under UGC Infonet.

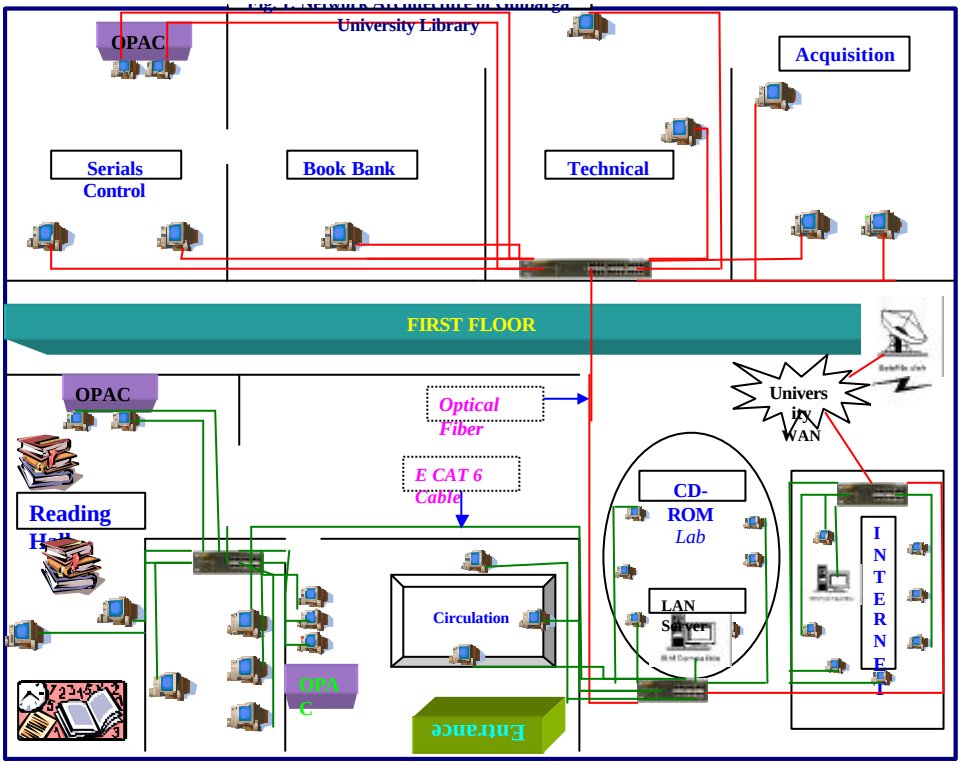


Figure 1. Network Architecture of Gulbarga University Library

After implementing security for the network's physical components, the administrator ensured that the network resources would be safe from both unauthorized access and accidental or deliberate damage. Policies for assigning permissions and rights to network resources and access library modules of SOUL are at the heart of securing the network (Figure 2). The two security models have evolved for keeping data and hardware resources safe in the library are by using password-protected shares and access permissions depending upon the nature of their work in the library.

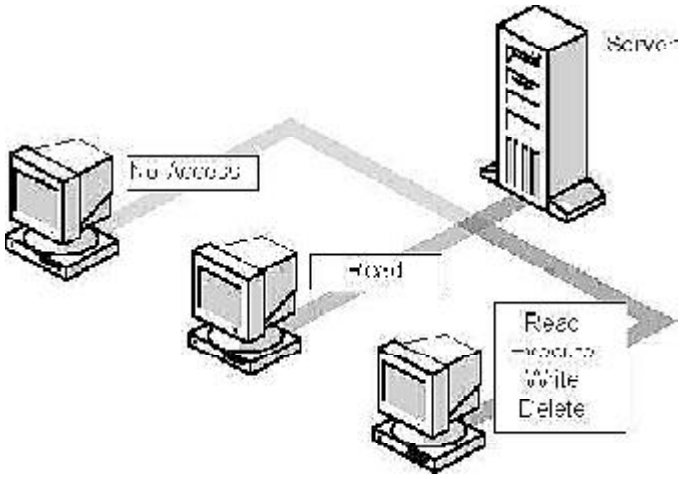


Figure – 2: Access Rights of Server

In this networking environment there must be assurance that sensitive data will remain private to secure sensitive information and to protect network operations from deliberate or unintentional damage. The major threats to the security of data on a network are unauthorized access and electronic tampering. The options available to protect the software at Windows Advance Server 2000 and specialized Data Protection and recovery software and McAfee firewall have been used for the safety of data. This is in addition to the regular backup of the data in CD-ROM. However, for the OPAC nodes kept for the users are prone to corrupt frequently has been monitored by security options and system restore utility tools available from Windows operating system.

2.4 Database Management System

Although, the library has started its automation a decade ago using CDS/ISIS, the retrospective conversion is now being switched over to SOUL. As on date, the Books collection in e-format is about 1,02,000 records with up-to-date Theses (5600) and Serials holdings (425) available in SOUL is open for Public Access Catalogue in the library for the user community.

The library has explored all the modules of SOUL including Circulation section for issue and return of books. This e-service has been started this year, by taking initially a few departments i.e. Library and Information Science, Mathematics, Statistics, Physics, Applied Electronics, Chemistry, Biotechnology, Computer Science and Management Studies comprising of faculty and research scholars. A well-documented plan has been worked out before implementing automation in Circulation unit, working out for privileges of issue/return and the type of users.

2.5 Access to CD-ROM Databases

CD-ROM technology has been established under networked environment to promote the concept of Digital Library. It is one of the pioneer libraries in the country to adopt this latest technology using CD NET Tower for multi-user access. At present, following sixteen national and international databases have been subscribed in various disciplines. The CD-ROM service is not only extended to in-house users of the University but also to the distant users from all over the country at a large scale. Even library undertakes literature search on behalf of users and send the same through e-mail or on floppy.

Sr. No.	Databases	Sr. No.	Databases
1	MathSci 1940+	9	Biological Abstracts 1992+
2	Sociofile 1974+	10	Biotech. Abstracts 1982+
3	Econlit 1969+	11	LISA Plus 1969+
4	ERIC 1966+	12	Psych-Info 1872+
5	Cross Culture 1989+	13	Georef 1785+
6	Dissertation Abstracts International 1861	14	INSPEC 1989+
7	ABI/INFORM 1989+	15	CABSAC 1973+
8	IBID 1993+	16	Supreme Court Case Finder Online 1950+

2.6 Internet Computing Center: E-Journals under UGC Infonet Program

The Internet users Lab has been established under networked environ with 16 computers using V-SAT under UGC Infonet program. Gulbarga University is now provided the facility to access 1200 scholarly journals on the Internet from Springer and Kluwer Publishers 1200 journals in full-text covering all areas of learning, which can be accessed at http://www.springerlink.com and http://www.kluweronline.com and other scientific journals and databases. Thus, the Internet has been the most important source of information for the teaching and research community as an instrumental tool; research tool and communication tool that extend the content of the curriculum enrich classroom discourse, remote communication and enhance learning opportunity. To achieve this, VSAT connectivity of UGC Infonet has been commissioned in Central Computer Center of the university, which in turn, the library has made sincere efforts to lay 6 Core Out door Optical Fiber Cable to the library premises.

2.7 Library Home Page

The library web site has been designed providing comprehensive information about the library activities and services, which is accessible at http://www.gulbargauniversity.kar.nic.in/library.htm. Efforts were made to customize the home page to provide need-based information pertaining to local and immediate needs, even for general users. It provides complete history of the region and address including telephone numbers of educational institutions, Government offices, NET Papers, Syllabus, R & D projects, research publications of the university, alumni details etc.

2.8 Institutional Repository Research Productivity:

Establishing a Central Research Database (Full Text) of the University shall provide a complete and crystal clear of the academic and research progress of the scientific community in the form of publications and help in monitoring the ongoing and completed research projects undertaken by the faculty members

and research scholars. This Institutional Repository Database includes the university productivity i.e. research publications (Full text), educational programs, Instructional materials of the faculty. This service has just begun by using D-Space under Linux environ. The teaching instructional material of Law department is now made available along with few research publications of Science and Social Science departments. The research publications have been scanned and provided a link to the Meta elements of D-Space software. Emphasis on scientometrics and bibliometric studies will be established which would help in measuring the productivity and impact of research and teaching. Efforts are on to complete the task revealing the mirror of research trend of Gulbarga University.

3. Issues and Prospects

The job of library does not end itself by establishing techno-environment, but at the same time, the outlook of user community's use, education, training and evaluation is of utmost importance to serve the purpose.

3.1 Establishing LAN in the Library

LAN establishment has been the effort for the last five years that has yielded fruits now and as such efforts should be made unless and until our demand is met. As management (administrative) support is very much required to execute the network application for building up strong IT infrastructure in building their image by carrying out innovative task to the satisfaction of user needs and aspirations.

Technical infrastructure including hardware and software selection is of utmost importance, which needs to be selected based on the compatibility of library software used and making provision for future flexibility. This has to be carried out in consultation with National documentation centers and software engineers.

3.2 Maintenance of Network

One of the major challenges faced today by the library managers in university library environ, is to ensure proper maintenance of network, so that the regular in-house and services are not affected. This calls for expertise knowledge and support through regular Annual Maintenance Contract (AMC) to the genuine System engineers to attend the complaint timely. As users have a tendency to experiment all sorts of their work on the computers and so also by the staff, monitoring and assigning user and staff rights has to be streamlined.

3.3 Limitations of Library Software

No software is perfect, but needs to be exploited for in-house activities and services of the university library. The SOUL is being used quite effectively in the library with the active support of INFLIBNET, staff in implementing various modules including Circulation and Web OPAC. The customization is required especially for the

- Retrieval of documents based on 'Type of Material',
- Editing of work based on time period
- Under the Main Menu of OPAC module, Separate sub menu is required for the collections i.e. Theses, Reports, Book-Bank, Reserve Collections, Special Collections, CD-ROM Databases
- Customized Bar Code Label comprising of university emblem
- Security options and load time of access.

The scientists from INFLIBNET, Ahmedabad is in the process to attend these requirements to overcome the limitations.

3.4 Mass Literacy to overcome inhibitions

Users information bent of mind has been the driving force for the automation of library activities and services. The library has really contributed and supported the user community in building confidence to use computers and thereby overcome inhibitions of technology to create mass literacy of technology. The OPAC (Online Public Access Catalogue) nodes meant for users has stimulated them to use computers for searching catalogue and thereby the concept of literacy has generated in the campus, besides extending Internet and CD-ROM and IT based services to the clienteles.

In this direction, Gulbarga University has actively coordinated in organizing a five-day Information Technology (IT) exhibition, with use of Internet as its main theme, to coincide with the Rajyotsava Day celebrations on 1st November 2004, with the support from Karnataka State Government. This is to provide an opportunity for students of high schools and higher secondary schools in the rural areas to have their first glimpse of computers and get basic training on the Internet facility and improve their knowledge by downloading the latest information available on the net. A notable feature of the training programme was that the manual in the computers screen was in Kannada to help the students studying in the Kannada medium.



Figure –3: IT Exhibition at Gulbarga University

3.5 Education and Training

Besides, on site education and training to use OPAC, CD-ROM Search and Internet access, regular user orientation programme is being organized to the freshers of the university, which includes online demonstration and practical visit to the labs of the library. Also need based programmes on Information search skills/ strategies and Communication skills (written/oral) are organized.

4. Future Plan

4.1 Learning Resource Center

The long-term goal of the university library has been the realization of the concept of digital library to accomplish the objective of universal access unrestricted by time and place in order to harness the ICT for enhancing the quality of teaching and research. In this direction, the University Grants Commission, New Delhi, India has already sanctioned Rs. 1 Crore for extension of library building towards establishment of Learning Resource Center in our university library. At the heart of Learning Resource Center's philosophy is the management and provision of information to meet the needs of users – students, research scholars, faculty members and external clients in the IT based environment. Each student of Gulbarga University at the center of the learning world with every resource they need would be made available by the just push of a button.

4.2 Retrospective Bar-coding of Books

Bar coding of books is on the cards has to be immediately taken up by generating subject-wise bar code labels generated from the SOUL software. This is planned to go to racks physically and paste the bar code labels and also monitor in the circulation section, in addition to the fixing of labels for the new books.

4.3 Provisions of Video camera in the Library

It is a privilege and a history in the country that, Gulbarga University Library is one of the few university libraries that has carried out stock verification in toto after 24 years of its establishment. A sense of monitoring over the user activities in the reading section of the library and proper maintenance of documents and to avoid theft and cutting of pages of books in the library by making provisions of Video Camera in the library. This also will help in proper monitoring of library staff on duty in the library.

4.4 Touch Screen Technology

Although, OPAC is open for the users for searching information in the library using mouse, but, the library plans to implement touch screen technology for browsing collections of the library avoiding mouse with fingertips on the screen.

5. Conclusion

As the change is the very essence of life, in the fast changing, confusing, information over-loaded world, the Libraries as intermediaries will have to operate in complex technology rich world for extending quality services to the user community. The Library managers have to play prime role in life long learning and community informatics in timely providing the society with state of art features to access the benefits of today's technology for survival and to build the image of librarianship in the best interest of the user community.

6. References

1. Brophy, Peter (1997). Libraries without walls: From vision to reality. London: Library Association, P177
2. Chuene, M.M (2001). The effect of information technology on library acquisitions: experiences at the University of the North, South Africa. African Journal of Library Archives and Information Science 11 (1): 25-38

3. Lynch, C. (1999). Authentication and trust in a networked world. Educom Review 34(6): 60-68
4. Martey, A.K (2002). Management issues in library networking: focus on a pilot library-networking project in Ghana. Library Management 23 (4/5): 239-51
5. Nicholson, H (1999). Planning for communications and information technology: specific aspects of design for the multi-functional library. Liber-Quarterly-the Journal of European Research Libraries, 9 (1): 81-9.
6. Rowley, J and Slack, F (1999). New approaches in library networking: reflections in South Africa. Journal-of-Librarianship-and-Information-Science 31 (1): 33-8
7. Shafi, Zeenat.S (2002*). Access to use information technology in Indian universities. New Delhi: Association of Indian University
8. Sherikar, Amruth (1996). Gulbarga University Library's Computer-aided Information retrieval and INFLIBNET related activities: A study. 17th IASLIC National Seminar at University of Calcutta, 10-13th Dec 1996, pp.117-120
9. Shipp, J (2002). Development of library networking in national infrastructure: a case study. Herald of Library Science 41 (1-2): 26-34

About Authors



Dr. Suresh Jange is working as Assistant Librarian at Gulbarga University Library, Gulbarga, Karnataka. He has presented number of papers in seminar, conferences and journals. He is also a member of many professional bodies.



Dr. R B Gaddagimath working as Librarian at Gulbarga University Library, Gulbarga Karnataka. He has presented number of papers in seminar, conferences and journals. He is also a member of many professional bodies.



Dr. Amruth Sherikar working as Deputy Librarian at Gulbarga University Library, Gulbarga, KarnatakaHe has presented number of papers in seminar, conferences and journals. He is also a member of many professional bodies.

S B Policegoudar is working as Assistant Librarian at Gulbarga University Library, Gulbarga, Karnataka. He has presented number of papers in seminar, conferences and journals. He is also a member of many professional bodies.

Importance of Digital Library for E-Learning in India

H S Chopra

Abstract

The potentialities of information technology, together with economic concerns, have been forcing various organizations to go for digitization. This has also happened to libraries, whose primary value lies not only in their collections but also in their contribution to education through providing information service, facilities for e-learning and management of collected information, which they make easily usable and accessible to users. The design and development of web-based educational systems for people is happening in India also.

Keywords: Digital Library, E-learning

0. Digital Library and E-learning

Digital library in India is under the supervision of Indian Institute of Science which is being helped by the Carnegie Mellon University. For the first time in history, the Digital Library of India is digitizing all the significant literary, artistic and scientific works of mankind and making them freely available, in every corner of the world, for our education, study and appreciation and for all our future. Learning process through digital library will be available free to any one round the world and the needy will get benefit.

Digital library at National Level is must and should have legislative responsibility to provide access to the nation's documentary heritage and to preserve the heritage of the country for the posterity. National Digital Library(NDL) has also the responsibility of providing resources to support all facets of the Indian School Curriculum. These e-sources are being used and will be extensively used by the next generation for which information technology and telecommunication will be must. As we explore next generation e-learner system needs to be introduced from the school level. Moreover, this technique needs national approach. If we are to maximise all of the advantages of the next generation e-learning and digital library environments then we must achieve this level of interoperability, interaction and seamlessness. In future education will be fully connected and supported by smart use of information communication technology.

There will be need of connectivity at every level and content creation. It is generally agreed that there will be content from a variety of sources and repositories which will have many purposes and users to support teaching learning and research. The digital libarary in all of its manifestations (e-publishing, e-print, digitization etc.) will provide content for next generation e-learning environment.

E-learning will decrease the digital divide of the country. Within India some states are still backward and some too much advanced in information technology. Similarly some states have made provision for computer education from school level whereas some states are unable to afford such facility to all the schools.

Digital library has its own implications in India as we have lakhs of books in various languages. Thus language problem is one such impediment in digitization which is likely to be solved in near future as promised by Microsoft Co. However, after the middle of twentieth century number of books in English increased at a rapid speed because of the spread of education , development in science, technology and telecommunication. Most of the books in these subjects are by foreign authors which cannot be digitized by Indian Digital Libraries as the copyright law does not permit them to do so. So far as Indian books are

concerned these relate particularly to religion, humanities, social sciences and very few to other subjects like sciences, technology, telecommunication and information technology.

Digitisation of old Indian books and manuscripts may be useful only to few scholars whereas the present need of the students and scholars is latest publications in science & technology. Old Indian literature can be digitized without much difficulty as there will not be any major copyright problem.

Information for all can be applicable only if the advanced countries provide free latest information to the developing nations. E-learning with latest knowledge will be available only if the digital library has all the latest books digitized. Changes are very fast even in education technology. The days are not afar when e-learning and digital library will be popular and much sought after by the students in India. However, certain institutions and universities have started work in this direction and are now leading the others to follow the same.

With the advent of fast development in telecommunication and information technology the traditional system of education, popular for centuries, is going under rapid change. Computer and internet technology has absolutely revolutionized the system of education. India has lakhs of schools, about 14,000 graduate and post graduate colleges and three hundred universities. At present more than 85 lakh students are going for higher education and millions of students are studying at school level. Inspite of it our 35% population is still uneducated because of several reasons.[1]

The world today engrossed by the web revolution is strongly influencing all aspects of life including education. India is no exception. Learning will shift its focus from 'Teacher centric' to 'Learner centric' education system.[2] E-learning grew out of the distance learning programmes offered by open universities and correspondence courses. Traditional libraries are being threatened by web technology.

ODES (Open and Distance Education System) is now gaining roots in India. India has ten open universities and one National University. About 235 different academic programmes and more than two thousands different courses (subjects) are offered through these universities. These universities have on rolls about 14 lac students. So far only seven languages have been included to provide education to the students.[3]

Number of students in these open universities is increasing very rapidly. However, the students of urban areas are privileged ones because of various educational facilities. Students of rural, tribal, hill areas, part-time servants, handicapped persons and military personnel are in need of education which most of them cannot avail of because of various reasons. Correspondence education is losing its popularity because of the universalization of internet.

In India print media is still popular and role of other medias like audio, video, radio, television, multimedia etc. is still only supportive. Learner has to match place and time for using the other media for learning.

Availability of information on internet has proved a boon for the developing countries. Micro or Macro Information of almost every subject is available on the internet. With the coming of internet Indian educational institutions are under great pressure to use internet for teaching.[4] Because of it E-learning has become possible even in third world countries.

Knowledge is expanding at lightning speed in the world. Indian students also need to learn more, better and faster. E-learning system can empower both students and teachers, for quality education. It will create a knowledge resource for the nation and any module can be easily shared by anyone, any where. E-learning aims to provide excellent learning support to the students, which is as good as face to face teaching. Such an effective learning, improved quality, reduced duration, cost effectiveness and flexibility

can be considered as objectives of E-learning. The mode is not truly, 'online education' but may be called as a 'web enabled education'. In India initially the prices were very high, but with the passage of time the prices of computers, spare parts, browsing charges have come drastically down. It has enabled the poor and developing countries to avail of the facility. Web based E-learning is already viewed as an important tool to improve academic quality. Quality education 'anywhere anytime' is the need of the hour in India. E-learning is a great step forward towards ensuring quality education for all', with cost effectiveness at door steps of learners.[5]

Thus for the ODES (Open and Distance Education Centres) of India there is now a strong need to consider how the Internet, with its present technology, limitations with costs, can affect the teaching, learning, expectations and employability of students.

Internet is a fast, easy and reliable communication media with a global presence. Latest 'Active Server Page' (ASP) technology offers excellent secured opportunities for interactive intelligent communication and on demand feedback about learning effectiveness, on the internet. Video interaction, virtual classroom and (virtual classroom modules) are useful in best time utilization. Other features will be information in CDs, no ambiguities but clear knowledge, best development of understanding, discussion forum on internet, online counsellors etc.[7]

For E-learning in India standardization of internet software in universities and other educational institutions is to be standardized. In order to provide education on a large scale there will be need of study centers and counsellors. Internet will affect our mode of learning and communication methodology. ODES cannot afford to ignore the benefits of internet. Quality and efficiency of academic and administrative services will improve a lot. Drastic changes in the method of learning will appear in the new millinium in India. Some universities are providing online education, online admissions as well as online examination and online results on the internet.

E-learning is an ideal mode of imparting education in an open learning system, where learners are located at different places. E-learning will be provided via all electronic media including the internet, web based digital library, intranets, extranets, satellite broadcast, audio-video tape, interactive TV and CD ROM. Content creation, content management and content distribution are other components of E-learning.

India is a developing country with 26% population still living below poverty line. Poor students unless helped financially cannot avail of facility of E-learning. E-learning allows an end user to learn at his own place, own pace and the time at which he likes. E-learning tool is a complete, composite and customized solution. It offers an engaging online environment that delivers the knowledge we need, when we need it, and where we need it. Participants irrespective of their locations can interact with other participants, facility or management through various collaborations.

There are a number of contentious issues like instructional design, quality, cost, student's outcomes, access, equality, accreditation, personal and institutional impact. E-learning is poised to catalyze both-competitive and collaborative relationships among the profit firms and non-profit colleges and universities. Unlike India, almost all the colleges and universities in U.S.A. are using e-mail, materials on the WWW (World Wide Web), or other internet applications. It will reduce average instructional cost per student and the distinction between 'distance' and campus based students will also disappear slowly and slowly. E-learning cuts the costs of travel, other expenditures, administrative over head, duplication of effort etc.

Because of digital library on internet the students can have twenty four hours accessibility, consistent quality, high retention of power and low cost. E-learning that is using the internet for instruction in post secondary education and training – has been joyously welcomed by some and bitterly described by others[12].

1. Conclusion

The challenges of providing education to a huge population in India with diverse needs and learning styles will require a new approach in our delivery strategies. It is essential that UGC should create a learner-centered, cost effective system to support the use of E-learning by the institutions. The target group of the system will be the teachers as well as learners, at all levels (formal, non-formal, informal and continuing), multimedia content developers and learner support staff.

Future of E-learning through digital library is very hopeful in a country like India. Those who are yet deprived of educational facility because of various reasons will be able to avail of E-learning facility very soon.

2. References

1. Chopra, H.S., Library Conservation, New Delhi; Common Wealth Publishers, 1995, p. 21.
2. India 2004 : Refence Annual, New Delhi; Ministry of Information and Broadcasting Government of India, 2004. p. 207.
3. Killedar, Manoj, "Web based Engineering Education in India". http://killedar. Tripod.com accessed on 10.03.04 at 7.15 pm.
4. Ibid., p. 3.
5. N. Venkateshwarlu and Subhasis Maji, "E-learning in Distance Education", Journal of Distance Education. 2004, p. 162.
6. Killedar, Manoj, p. 5.
7. Loc. Cit.
8. Ibid., p. 6.

About Author



Dr. H S Chopra presently he is working as Librarian in Guru Nanak Dev University Library, Amritsar (Punjab). He has done M.A., M.L.Sc., Ph.D. (Lib & Inf. Sc.) Ph.D. (History). He has also contributed a number of articles to National and International journals including Encyclopedia of Library and Information Sciences (U.S.A.). He has already seven books to his credit. He is also a member of many professional bodies.
Email : chopra_gndu@yahoo.com

Transformation of Library Services : With Special Emphasis on Digital Reference Service

Padmini Kaza

Abstract

Easily accessible digital information has rapidly become one of the hallmarks of the Internet. Online resources have surged in popularity as more individuals and organizations have connected to the global network. Thousands of organizations have turned to Internet-based information delivery as an effective and cost-efficient alternative to traditional communication methods, and many have expanded their services further by interacting with their users and responding to inquiries via the Internet. Digital reference services provide subject expertise and information referral over the Internet to their users. This paper provides an overview on transformation of traditional information services to digital reference service.

Keywords: Digital Reference Service, Reference Service, Digital Library

0. Introduction

The quest for information has been expanded significantly and it is , now, coupled with a desire for swift and better services. In the mist of fast emerging information explosion, finding a particular relevant information has been a very cumbersome job. Hence there is a dire need to create and manage digital resources of information and make them easily accessible to the user.

Today's researchers need to find quickly information that is usable, relevant, authoritative and verifiable. To meet that need, libraries must adapt the traditional strengths of acquiring, describing, and serving information to an environment that is not bound by time or physical place, the virtual library without walls.

1. Origin of Digital Reference Service (DRS)

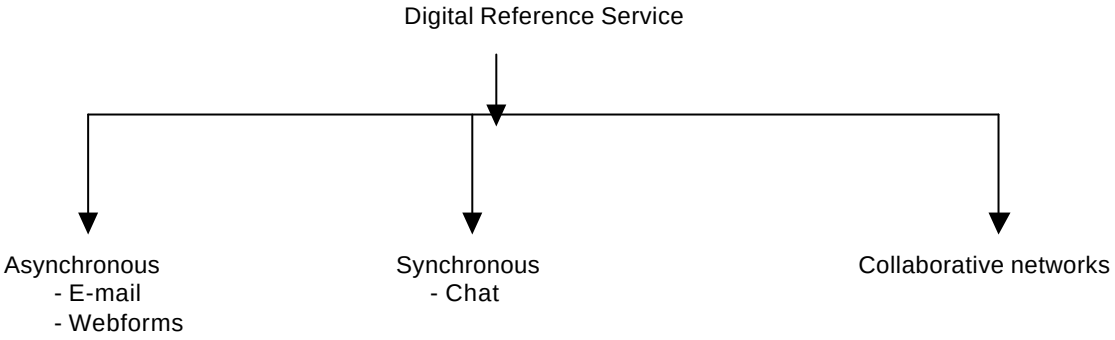
The origins of digital reference can be traced to the library field, where libraries sought to augment traditional services by providing reference assistance in an electronic environment. One of the first services to go online was the Electronic Access to Reference Service (EARS), launched by the university of Maryland Health Services Library in Baltimore in 1984. Although initial e-mail-based digital reference efforts received little attention from patrons, digital reference services proliferated over time and became increasingly popular, eventually spawning such internationally-known services as AskERIC in 1992 and the Internet Public Library in 1995.[1]

1.1 What is DRS ?

Unlike traditional reference, digital reference services allow patrons to submit questions and receive answers via the Internet and other electronic means. The users get connected with librarians or information professionals and receive direct assistance wherever and whenever they need it. In addition to answering questions, these information experts may also provide users with referrals to other online and print sources of information and support the development of skills such as information literacy. The terms "virtual reference", "digital reference", "e-reference", "Internet information service" and "AskA service" are used interchangeably to describe reference services that utilize computer technology in some way. [2]

1.2 Tyres

The following are the various types of providing digital reference service:[3]



2. Need for Standards

The purpose of standards is to promote digital reference best practices on an international basis. The online environment is uniquely suited to consortia models of work and to the development of shared resources. Libraries in different countries may have different traditions of public service, which both affect their current reference practices and their patrons' expectations. But it is also important to recognize that new technologies will enable librarians to redefine the scope of their public services.

Butler examines the need for a common, standard data format for the management of reference transactions. This need is discussed from two perspectives: that of major research libraries and that of AnswerBase Corporation (ABC), a Web-based digital publisher.[4]

3. Process

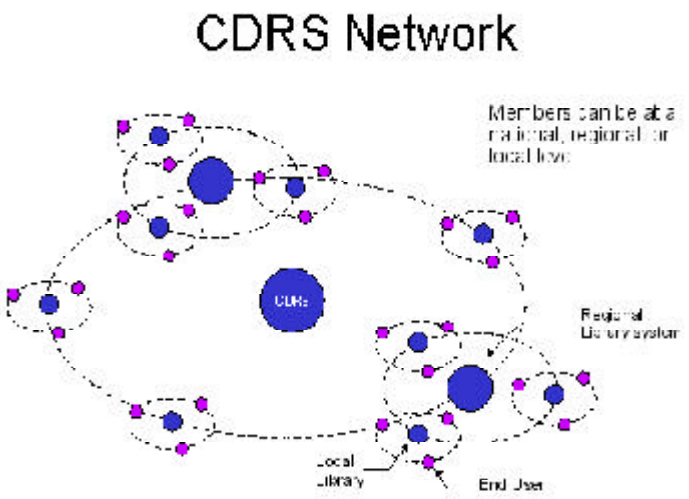
A six-step process was developed to aid organizations in the creation and operation of digital reference service, which can be applied to a wide variety of organizations and audiences including the K-12 education community, government agencies, libraries, and industry. The steps are: Informing, Planning, Training, Prototyping, Contributing, and Evaluating.[5]

A model was proposed that is particularly relevant to academic libraries in large research institutions, but extendable to other types of organizations with similar characteristics. The five critical issues examined include: integrating virtual reference service with existing services; allocating fixed resources; acting as an effective advocate to secure organizational support; developing a distributed service model integrating specialized, subject-domain expertise; and targeting and serving disparate segments of the user community.[6]

4. International Scenario

Libraries, including the Library of Congress, have a rich tradition of collaborating to get work done. Institutions have collaborated to preserve collections, to catalog materials and make them accessible, and to create virtual libraries. They have borrowed collection items from one another and used one another as service models. By linking libraries for reference services, the Collaborative Digital Reference Service(CDRS) would combine the power of local collections and staff strengths with the diversity and

availability of libraries and librarians everywhere, 24 hours a day, 7 days a week. The graphic below provides an idea of what this network system could look like:[7]



4.1 Internet Public Library (IPL)

The Internet Public Library (http://www.ipl.org), which exists only in a networked environment, started in 1995 as a student project at the University of Michigan School of Information. It was the first virtual Public Library on the Internet. As a public service organization, the IPL also serves as a learning and teaching laboratory. The IPL offers an annotated collection of high quality Internet resources and a reference service available via a Web form. This service is offered by a network of reference librarians (professional librarians as volunteers, and library students in training). The service is free of charge and open to the Internet community.

- Answer Zone. Six public libraries in Texas. http://www.answerzone.org
- Ask-A-Librarian. Eleven members of the Association of Southeastern Research Libraries. http://www.ask-a-librarian.org/
- AskNow. An Australia-wide service offered by the National Library of Australia and the eight state/territorial libraries. http://www.asknow.gov.au
- QuestionPoint Collaborative Reference Service. Library of Congress and OCLC. http://www.questionpoint.org
- Washington Research Library Consortium. Ask A Librarian. Seven academic libraries in the DC area. http://www.wrlc.org/virtualref/
- Washington State Library. Statewide Virtual Reference Project. http://wlo.statelib.wa.gov/services/vrs/
- Refdesk.com - Since 1995, free and family friendly. Refdesk indexes and reviews quality, credible, and current information-based sites and assists readers in navigating these sites

5. National Scene

Many libraries of higher learning and research institutes are stepping towards digitizing their resources. To meet the information demands of their user communities, besides rendering traditional reference and information services, they have also introduced OPAC, E-mail and Bulletin Board, Document scanning, CD-ROM networking, Internet access, Electronic referencing, Indexing and Abstracting, CAS&SDI, other bibliographical, electronic document delivery, and reprographic services. Although automated libraries are not yet sufficiently advanced to offer interactive reference services, electronically mediated reference services are increasingly available through libraries and information centers. Their next focus will be on Digital Reference Service. The following are some of such institutions:

- Indian Institute of Science, Bangalore. webman@library.iisc.ernet.in
- Indian Institute of Technology,Delhi. http://www.iitd.ac.in/accd/library/index.html
- Sir Jehangir Ghandy Library- XLRI, Jamshedpur. http://www.xlri.ac.in/library
- Indira Gandhi Memorial Library, University of Hyderabad. http://www.uohyd.ernet.in
- Jawaharlal Nehru University library, Delhi. http://www.jnu.ac.in

6. Benefits

The benefits of Digital Reference Service are:

- Regional library consortia offer member libraries the opportunity to share reference questions with each other using the Internet and other technologies.
- Collaborative Digital Reference Service allows individual institutions to share expertise and resources.
- Expanding hours of service.
- Providing access to a large collection of knowledge. Serve the public good by providing valuable information in a timely fashion, and have the potential to gain international visibility.
- Enhanced public relations benefits by having satisfied users and by providing high-quality of information.
- Accessible 24 hours a day and unrestricted by geography. [9]

7. Challenges

Some of the important struggling issues in the provision of digital reference service are listed below:

- Librarians often juggle real-time patron requests with those of walk-in or phone patrons.
- Staff must be trained to use selected real time tools.
- On going technical support must be available to maintain the system.
- Ensuring the quality and consistency of responses.
- Reaching Consensus in developing procedures and policies.
- Configuring technology that can be best accessed and used by each participating group.
- Premature launched services may not have potential impact on global audience.

- ✍ Which user population to serve?
- ✍ How to respond to question overload?
- ✍ Secure funding for continued operation.
- ✍ Task of creating and managing Internet – based question-and-answer services is complicated by the ever changing nature of the Internet

8. Possible Means

Besides acquiring adequate financial support and infrastructure, following are some of the possible means to ensured qualitative and effective Digital Reference Services, in institutions of higher learning:

- ✍ Librarians as information engineers have to learn more about the Digital Information Technology and they must of necessity, change their way of thinking if they are to meet the challenges of the present century.
- ✍ Information professional need to think creatively and adopt new technology.
- ✍ Interact with users to learn about their requirements and expectations.
- ✍ Librarians should reorient themselves.
- ✍ Web is in need of librarians who are trained in the structuring and organizing information.
- ✍ Participating staff should have the ability to locate and evaluate information resources, and have in-depth subject expertise.
- ✍ Traffic from non-primary clients needs to be controlled.
- ✍ Services are to be pilot-tested in a controlled environment.
- ✍ Regular evaluation ensures the quality of services.
- ✍ Defining realistic service goals, accompanied by workable policies and procedures, with participating staff fully cognizant ensure consistently excellent Digital Reference Service.

9. Conclusion

Digital reference services are a powerful means for the free exchange of information and the promotion of interactive learning. Thousands of organizations, at International level, have turned to Internet-based information delivery as an effective and cost-efficient alternative to traditional communication methods. As the provision of such service is very close before the Indian higher learning institutions, this is the right time to probe into the realities and current practices and step carefully forward in building and maintaining such service. Organizations interested in offering Internet-based information services must understand not only the fundamental tenets of the question-and-answer process, but also how this information is processed and translated into actual service. By proper planning and understanding of digital reference practices, libraries in higher learning institutions can ensure the effective creation and maintenance of exemplary digital reference service and high quality of the same.

10 . References

1. Wasik, Joann M. (2003) Building and Maintaining digital reference services. http:// www.michaellorenzen.com/eric/
2. http://en.wikipedia.org/wiki/Digital_reference_services

3. ibid.

4. Butler, Brett/ Answer Base Corporation. KnowledgeBit: A Database Format for Reference Version 2.0. http://vrd.org/Dig_Ref/dig_ref.shtml (14.11.2004)

5. Lankes,R.D.&Kusowitz,A.S. (1998) The AskA starter Kit : How to build and maintain digital reference services. In Wasik, Joann M. Building and Maintaining digital reference services. http://www.michaellorenzen.com/eric/

6. MacAdam, Barbara and Gray, Suzanne. A management model for digital reference services in large institutions. http://vrd.org/Dig_Ref/dig_ref.shtml (14.11.2004)

7. Kresh, Diane Nester/ Library of Congress.(2000) Offering High Quality Reference Service on the WebThe Collaborative Digital Reference Service (CDRS)) http://www.dlib.org/dlib.html (17.11.2004)

8. Kasowitz, Abby S.(2003) Trends and Issues in digital reference services. http://www.michaellorenzen.com/eric/ (14.11.2004)

9. Wasik, Joann M. Op cit.

About Author



Mrs. Kaza Padmini presently working as Associate Professor(1998) in Department Library and Information Science Institution, Sri Venkateswara University, Tirupati, Andhra Pradesh. Prior to this responsibility she has worked as a Lecturer in Dept of Library Science, D.R.W.College, Gudur, Nellore Dt. (A.P.). She has PhD in Library & Information Science, MA(Population Science); MLIS; B.Sc.(Home Science). Guided number of MLIS Dissertations and guiding research scholars. She is a life member of various professional associations – ILA, IASLIC, IATLIS, APLA, ALSD, & SIS. She has contributed 10 research articles in seminars, conferences & journals.
Email : padmisvu@yahoo.co.in

Role of Digital Libraries in E-Learning

Prachi Singh

Abstract

Today the concept of e-learning or online learning is getting emphasis in many organizations like academic, research, corporate or government. But method of implementation of e-learning is still a question. Many organizations have developed separate IT based systems for e-learning which are independent and costly as well. While others have preferred to develop e-learning modules integrated to their other IT technology based systems like KM System (Knowledge Management), Content Management System (CMS), or digital library. Digital Library provides a very well framed and compatible base for e-learning module. This article deals with various aspects of e-learning, strategies for e-learning implementation, role of digital libraries in developing e-learning, and few cases where digital libraries are successfully used in e-learning.

Keywords: Digital Library, E-learning

0. Introduction

Let's consider the term *library* the way Dr. S. R. Ranganathan defined & called it a "growing organism". In past decade we have observed a dynamic change in anatomy and physiology of library. What was considered to be a storehouse for books some few years back has now become an information provider and valuable capital of an organization.

As the type of format of information changed from written to e-format, libraries assumed the role of digital libraries. The concept of digital libraries was a major success as it came out of closed walls of library and reached users in their home, workplace and even while traveling, with the help of laptops. This approach reduced major barriers in the path of information dissemination like distance and time.

Today library is a hybrid of print and digital resources. Many developed and automated libraries who have large amount of digital content are now trying to provide e-learning through their digital libraries' web interface, thereby developing a fully dependable knowledge system. In this article we will look upon various aspects of e-learning, how digital libraries can contribute to e-learning, and practical examples of some universities and institutes which are in the process of developing e-learning mechanism through their digital libraries.

1. E-Learning

As the world is contracting and time is reducing, human aspire for more information in less time. Internet has come as boon at such a time. As far as teaching and learning is concerned, educators and students, both find it difficult to take out time from their busy schedules to meet and spend time at a place like classroom. Here Internet brings in picture "Online Learning" or "E-learning". Online education is becoming increasingly common in schools, colleges, and the training realm.

Online education in particularly best mode for distance education, i.e. settings in which learners and teachers are located in different places and all or most interaction takes place via the network. In E-learning learners are responsible for their own education & study strategies to accomplish their academic goals.

1.1 Brief History of E-Learning

Use of technology in learning, in true sense, started around the beginning of this century with the invention of films and motion pictures. During World War II world witnessed real use of technology in learning when U.S. Army used training films to educate their army-men and to maintain consistency in U.S. based training. Since then large scale developments have taken place at different fronts like academic, government, and corporate. In the sixties, early "teaching machines" and "programmed texts" paved the way for embryonic computer-based training. Instructional films became more creative and educational film business catered to both public and private sectors [1].

Next era of e-learning started when television was invented. But television did not become everyone's teacher due to the lack of interactivity with learner. The necessity of interactivity renewed efforts in the area of computer-based training (CBT).In the seventies and eighties a tremendous efforts were put in this field.

The real revolution in the field of e-learning started with the coming up of Internet. As quoted by Nobel laureate Gary S. Becker in 1992- "The Internet has begun to radically change the teaching of adults in the U.S. who want to improve their skills or further their general education."

Today's Internet (Web) is like a universal library which is anyway easy to manage, update with worldwide accessibility. Actually Internet is the perfect E-learning tool available to all and sundry.

1.2 Few Definitions of E-Learning from different sources

- ✍ Online Education or E-Learning refers to any form of learning/teaching that takes place via a computer network. Network can be of any form, viz. LAN, WAN, MAN or WWW. Various computer functions that are most commonly used for e-learning are: Electronic mail (e-mail), e-conferencing, groupware programs which include electronic whiteboard (shared writing space) kind of technologies.
- ✍ E-Learning refers to the use of Internet technologies to deliver a broad array of solutions that enhance knowledge and performance [1].

2. Digital Libraries in E-Learning

According to Gary Marchionini (1995) [2], libraries serve three roles in the learning process-

1. Sharing valuable resources- Digital library plays vital role here by allowing several information seekers to access materials simultaneously regardless of their physical location.
2. Preserving and organizing artifacts and ideas- million books project is a live example of a digital library serving for preserving old artifacts of intellectual importance by archiving their soft copies. The material is digitized with the help of scanning devices.
3. Bringing people and ideas together- Digital libraries offer diverse information resources shared by groups of learners irrespective of physical space and time. Digital libraries bring people together with different learning missions.

Various aspects that make e-learning through digital libraries an advantage are:

- ✍ Full text digital resources contained in a digital library database can be directly linked to its e-learning module.

- ✍ Efficient Digital Library software provides ability to identify, access, evaluate, organize and communicate information and knowledge. This happens because most of the Digital Library softwares use maximum librarian's tools including metadata for effective retrieval, standards like Z39.50 for effective information interchange between two different libraries, etc.
- ✍ Digital Library of an organization or institute is Intranet based, thus providing a readymade framework for e-learning.
- ✍ Digital Library provides a platform for all information and knowledge communication, user to user interaction, and provides facility for dual role of user as both teacher and learner.

3. Creating E-Learning Strategy

Implementation of E-Learning in an organization needs a well framed strategy. Normally we can consider that if an organization has a digital library with even most primary modules, it can provide a good foundation for developing e-learning system. But to develop an e-learning module we need to plan on various levels like technology level, policy level and organizational level. Marc J. Rosenberg, in his book "e-Learning: Strategies for Delivering Knowledge in the Digital Age" [1], describes in detail the various viewpoints required while planning e-learning for an organization. Following are few of the necessary steps for developing a E-learning strategy:

3.1 Who will be participants?

First and foremost we should identify who all should participate in strategy development. People who should be involved at initial level should include training managers, developers, instructors and administrators. In second stage all prospective clients should be involved like students, organizations (if the project is big one), sponsors and other stakeholders like senior managers. On the basis of the compiled result of these two discussed stages, IT department should be included to discuss technical aspects. Once whole strategy is clear a small task force should be used further

3.2 Analyzing Current Situation

The first step after identifying all participants is to fully analyze the current situation as it pertains to the ability to launch and sustain e-learning. It includes identifying main objectives of your organization, current state of your overall learning and development efforts, and analyzing how much support you will get from the administration. You also need to analyze needs of the clients, the current state of the technology infrastructure in your organization and current level of funding for e-learning.

3.3 Describing The Desired Situation

Once we know everything about our organization's needs and current situation, we need to design the future picture of e-learning in our organization as desired by us. So we create a detailed description of where you want your learning and development efforts (including e-learning) to be. For this identifying following points can be helpful-

- ✍ Identify what are the best practices in learning and development, and e-learning.
- ✍ What should be your e-learning value proposition?
- ✍ Building a vision and mission for learning and development.
- ✍ What principles are most important to you in guiding how you will implement your mission and realize your vision?

- How learning and e-learning are defined by you and your organization?
- The support you receive from top management

3.4 State your vision and mission

A vision statement describes a future state as if it were the present. It's more about how you will be recognized and valued internally and through the eyes of your users. Once you agreed-upon vision, develop a mission statement, i.e., what will you do to achieve your vision.

3.5 Gap Analysis

Create detailed specifications of the key disparities between the currect and desired situation, along with associated descriptions of root causes.

3.6 Conduct a SWOT analysis

A SWOT analysis looks at the entire organization to determine its strengths, weaknesses, opportunities, and threats, either at the moment or at some future point in time.

3.7 Strategy Recommendation

Based on the work done till now make specific strategic recommendations to close the gaps, implement the mission, and achieve the vision.

3.8 Build an Action Plan

Prepare an action plan to implement e-learning on the basis of the findings of the till now done research and studies. This is where the specific tactics are described in enough detail so everyone knows what needs to be done. In this stage,

- Identify critical success factors.
- Set and stick to timeliness and milestones.
- Provide adequate funding for implementation.
- Define and implement a change management plan.
- Define and implement a communications plan

4. Practical Examples of Digital Libraries in E-Learning

4.1 University of Strathclyde Library Services

University of Strathclyde Library Services [7] has been providing information resources for the teaching and learning of the University since many years. Its role in supporting virtual learning is no different. The Library continues to provide assistance to teaching staff engaged in e-learning. They concentrate on making the learning environment Information-rich.

It provides advantage over traditional learning in following senses:

- Rather than pointing to digital information resources such as e-books and e-journals, you may wish to import the full text of an e-resource in its entirety into your VLE space or class web pages.

- ✍ The Library can offer advice on the legal sourcing/acquisition of files and digital information resources on digital rights clearance.
- ✍ If, for example, some user want to mount a specific subject list of e-journals subscribed to at Strathclyde within his/her learning environment (not as external Library page), Library can create a stable, static url that user can place as an anchor in the code of one of his/her pages.

4.2 Cognitive Arts and Columbia University

Cognitive Arts and Columbia University [3] have come together to build high quality e-learning courses as a way of offering the educational advantages of Columbia to a wider audience. Columbia Continuing Education Online is a very good example of Columbia's initiations in the field of online education. The relationship of Columbia University with Cognitive Arts allows Columbia to examine and deploy new media as tools to enhance their educational resources and to expand content delivery opportunities, in keeping with the overall mission of the University [4].

4.3 RGUHS E-Learning Program

RGUHS (Rajiv Gandhi University of Health Sciences) situated in Bangalore, India, has developed an e-learning module called MedInfo (made from Medical Informatics) which provides information relevant to medical sciences. This E-Learning platform [5] is developed with the support and guidance of Health InterNetwork(HIN-India Project) and the World Health Organization. MedInfo aims to provide for both self education and trainer assisted education, in accessing and searching Biomedical literature on the web including:

- ✍ international and national sites in the public domain available in a variety of content categories spread throughout the world.
- ✍ using the National Health Information Collaboration

4.4 Needs Digital Library

NEEDS (National Engineering Education Delivery System) is a digital library for Engineering education. NEEDS [6] provides web-based access to a database of learning resources where the user (whether they be learners or instructors) can search for, locate, download, and comment on resources to aid their learning or teaching process over the world wide web. In addition, NEEDS supports a multi-tier courseware evaluation system including a national award competition, the Premier Award for Excellence in Engineering Education Courseware.

NEEDS' vision of what a digital library for undergraduate engineering education should be is more than just a traditional academic library in digital form. The digital library of the future will be a community of learners — encompassing faculty, students, and life-long learners.

4.5 iLumina Digital Library

iLumina is a Digital Library [8] of sharable undergraduate teaching materials for chemistry, biology, physics, mathematics, and computer science. iLumina was funded by a DLI-Phase 2 grant from the National Science Foundation. It was developed by the University of North Carolina at Wilmington, Collegis, Inc., Virginia Tech, Georgia State University, Grand Valley State University and The College of New Jersey. It is designed to quickly and accurately connect users with the educational resources they need.

5. Conclusion

Digital Libraries have proved themselves not only as effective repositories of knowledge and information, but also as effective communication medium between peers and for scholarly discourse. Jumping a step ahead, now Digital libraries are capable enough to provide an information rich platform to both instructors and students to teach, learn and share knowledge. Thus in future we will witness major role of Digital Libraries in online learning or as we call it "e-learning".

6. References

1. Rosenberg, Marc J. e-Learning: Strategies for Delivering Knowledge in the Digital Age. New York: McGraw-Hill. *pp (20-30, 291-303).
2. Jayawardana Champa, Hewagamage K. Priyantha, and Hirakawa Masahito. (2001). Personalization Tools for Active Learning in Digital Libraries. The Journal of Academic Media Librarianship Volume no (Issue no.): 8-1.
3. Schank, Roger C. Designing World-Class E-Learning: How IBM, GE, Harvard Business School, and Columbia University are Succeeding at e-Learning. New York: McGraw-Hill. *pp (187-188)
4. Columbia University. (2000). Columbia University and Cognitive Arts
5. Announce Agreement To Develop Online Courses. http://www.columbia.edu/cu/pr/00/05/cognitiveArts.html (Accessed on 29/11/2004)
6. RGUHS (Rajiv Gandhi University of Health Sciences), Bangalore. (2004). MedInfo: E-Learning Module. http://www.rguhs.ac.in/ELearning/medinfo/index.html (Accessed on 29/11/2004)
7. NEEDS, California (2004). NEEDS: A Digital Library for Engineering Education. http://www.needs.org/needs/ (Accessed on 29/11/2004).
8. University of Strathclyde (2004) University of Strathclyde Library Services. http://www.lib.strath.ac.uk/vle.htm (Accessed on 29/11/2004)
9. University of North Carolina (2004). iLumina: Educational Resources for Science and Mathematics. http://turing.bear.uncw.edu/iLumina/index.asp (Accessed on 29/11/2004)

About Author



Ms. Prachi Singh presently working in Indian School of Business as consultant of Learning Resource Centre. She did her ADIS (Associateship in Documentation and Information Science) course from DRTC, ISI, Bangalore (Documentation Research and Training Centre) in 2004.

Subject Gateways : A Case Study of the Science Campus Library, University of Madras

R Samyuktha

Abstract

Guindy Campus Library is the Science Campus Library of the University of Madras facilitating information access to the Science community belonging to the Schools of Physical, Chemical, Life, Earth, Energy and Environmental Sciences. Among the digital library services provided, the paper focuses on the Subject Gateways made accessible to the Faculty, Research Scholars and Students presently available only in the intranet as a case study.

Keywords : Subject Gateways, Portal, Digital Library

0. Introduction

Libraries have always been hybrid and complex information spaces and librarians are trained and experienced navigators of those spaces. With the advent of the web, libraries and librarians organize a large magnitude of information using the long established principles arising from traditional librarianship. There are new tools, standards and techniques emerging for the design, description, discovery and presentation of digital information, many of which are being developed in library environments.

1. Scope

This paper attempts to scan some of the digital library services provided by the Science Campus Library of the University of Madras with special reference to the Subject Gateways created specifically for their clientele, presently accessible on the intranet.

1.1 Overview of Digital Library Architectures

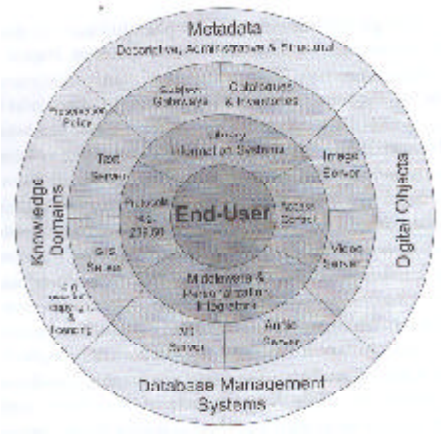


Fig. 1 OAIS model showing digital library functions

The above OAIS model (2001) shows many of the actual components that might be found in real working systems, with the user firmly located at the center, accessing resources within the controlled environment of the surrounding system components. In this context, particularly, we are concerned with the Information Systems for delivering documents and directing users to hybrid resources. In other words, the ring next to the Library and Information systems handles all requests for data objects through the data servers for the various media and information types, with supporting systems delivering Subject Gateways, Portals and Online Catalogues.

In summary, the outer ring provides structure and controlling mechanisms, the second ring serves content within a context defining system behaviors, while the inner ring provides access and interface mechanisms to present content to the user.

1.2 Definition – Portal vs Subject Gateway

In understanding the overview of the digital architecture and to lay focus on Subject Gateways it is required to note the difference between a Portal and a Subject Gateway.

A Subject Gateway facilitates access to networked resources for anyone who is looking for resources on that subject. A Portal, whilst also normally facilitating access to networked resources for anyone looking for resources on that subject, in addition offers a variety of additional services aimed largely, though not exclusively, at the relevant subject community. Furthermore, whilst gateways mostly "shallow mine" the resources in their subject areas by pointing towards them through hypertext links, portals tend to "deep mine" selected, resources in their subject areas by providing searching and sometimes other discovery services to all, or most, levels of these resources (Mac Leod, 2000).

2. Portal of Guindy Campus Library

Guindy Campus is the Science Campus of the University of Madras consisting of the Schools of Physical, Chemical, Life, Earth, Energy and Environmental Sciences. The Guindy Campus Library catering primarily to this science community has designed its library portal based on these user needs and their proactive suggestions in content development.

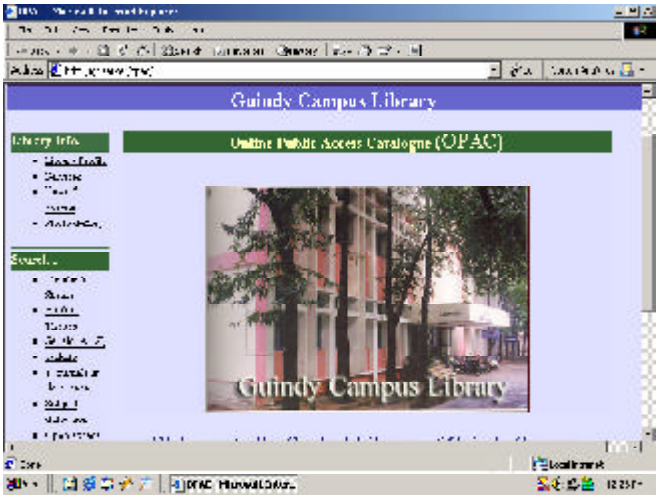


Fig. 2 Portal of Guindy Campus Library

Fig.2 shows links to the various e-services that are provided to the users besides the access to the library holdings.

The following figures restrict focus on the “Subject Gateways” developed on specific themes relevant to the Schools of Sciences in the campus.

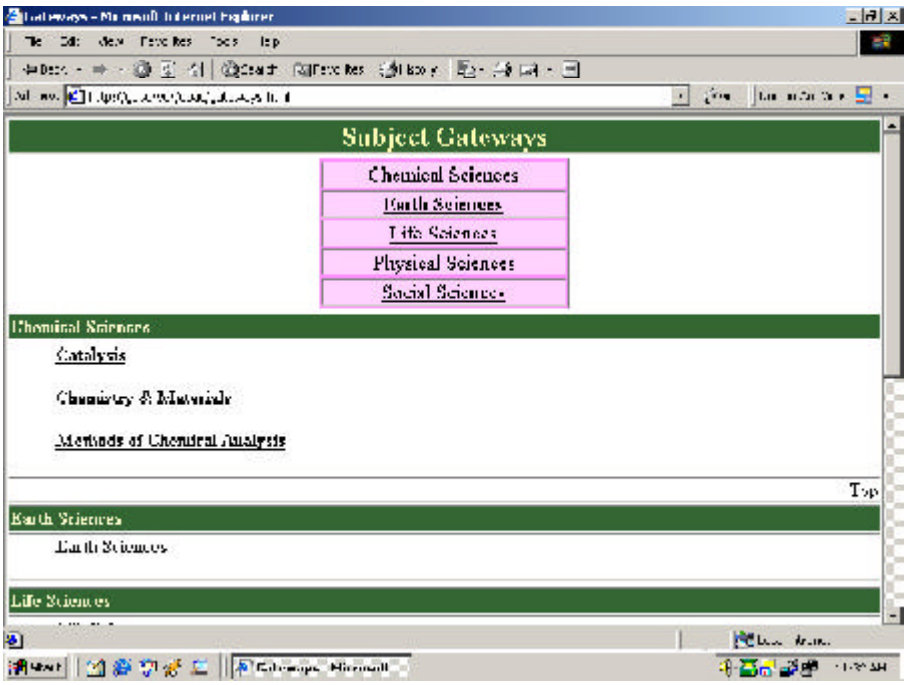


Fig. 3 Subject Gateways

Fig.3 indicates the five subject areas of Sciences based on which Gateways were compiled. In the category Chemical Sciences, Catalysis, Chemistry and Materials and Methods of Chemical Analysis are the themes.

Following Earth Sciences is Life Sciences which has themes of Microbial Biotechnology & Bioinformatics, Molecular Cell Biology and Techniques in Plant Biotechnology.

In the category of Physical Sciences, the focus area is Physics and in the Social Sciences category, Library and Information Science and Research Methods in Social Sciences are the themes.

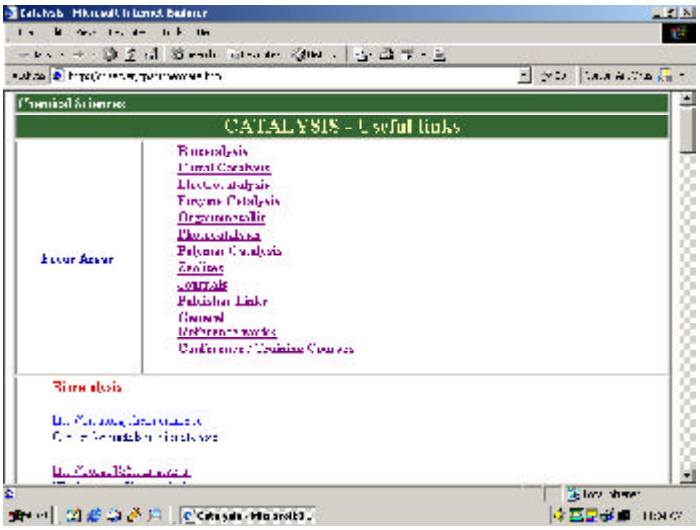


Fig. 4 Catalysis

Fig.4 illustrates compilation of useful links on the types of Catalysis such as Biocatalysis, Chiral Catalysis, Electro Catalysis, Enzyme Catalysis, Organometallic, Photocatalysis, Polymer Catalysis, and Zeolites. Besides these, there are links to centers for biocatalysis, forums, databases, journals, publishers reference works, conference and training courses.

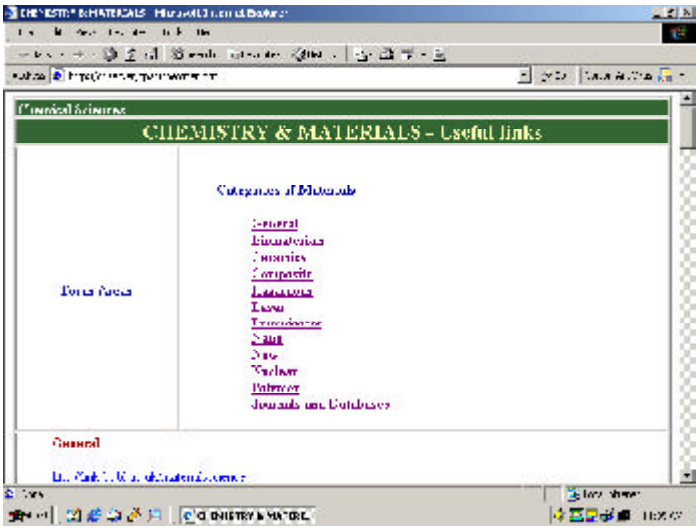


Fig.5 Chemistry and Materials

Fig.5-Besides links to different categories of materials such as Biomaterials, Ceramics, Composite, Hazardous, Laser, Luminescent, Nano, New, Nuclear and Polymer, electronic resource gateways for material science ejournals, catalogue guide for materials, databases and networks are provided.

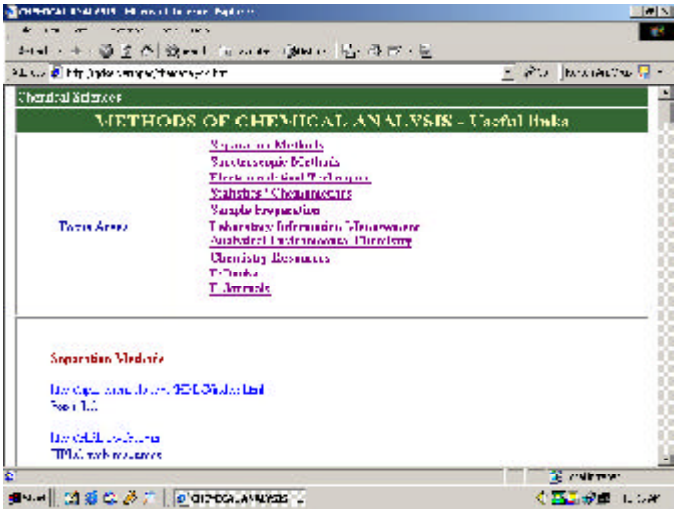


Fig.6 Methods of Chemical Analysis

Fig.6 - Different methods and techniques of chemical analysis, analytical chemistry resources, interactive labs, laboratory guides, catalogue of manufacturers and suppliers of scientific and lab equipment, societies, software, databases, e-books and e-journal links are furnished.

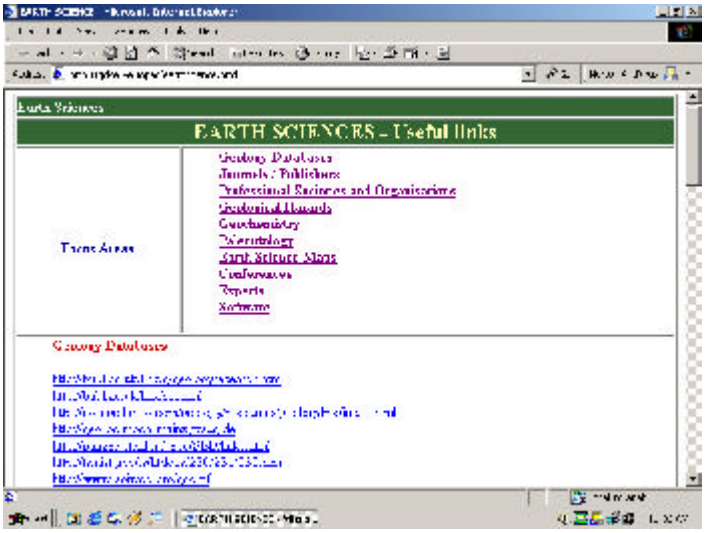


Fig.7 Earth Sciences

Fig.7 - Geology databases, journals and professional societies, geology labs online, geological hazards, paleontology, earth science maps, software, experts, research news and conference links are provided.

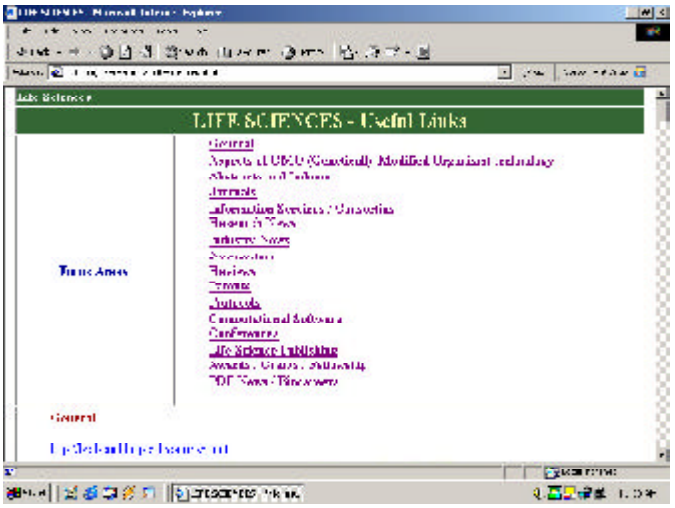


Fig. 8 Life Sciences

Fig.8 - Gateway to research in biological and biomedical research, endangered species, pesticide links, medicinal plants and properties, crustaceans, research and industry news, newsletters, reviews, patents, protocols, software, conferences and life science publishing are some of the areas to which links are provided. In addition to this, links to assays and compounds, journals, databases, indexes and abstracts, awards / grants / scholarship are also listed.

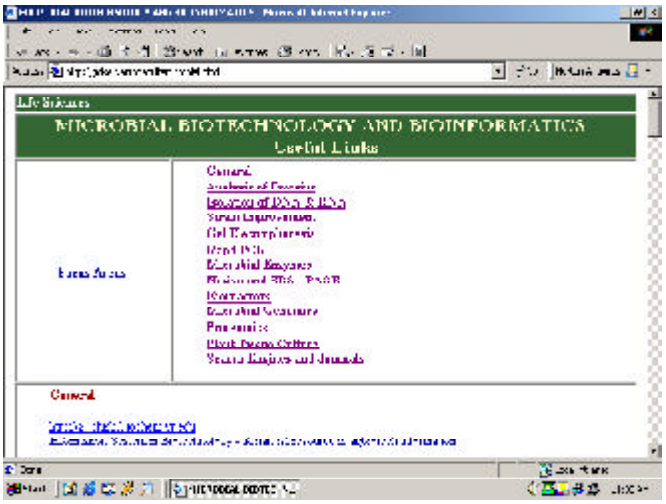


Fig.9 Microbial Biotechnology and Bioinformatics

Fig.9 - Useful links to focused areas such as analysis of proteins, isolation of DNA & RNA, strain improvement, gel electrophoresis, RAPD PCR, microbial genomics, proteomics, plant tissue culture are compiled, besides links to directories, databases, tools, tutorials, techniques, industrial products, specialized servers, power point presentations, scientific search engines and journals.

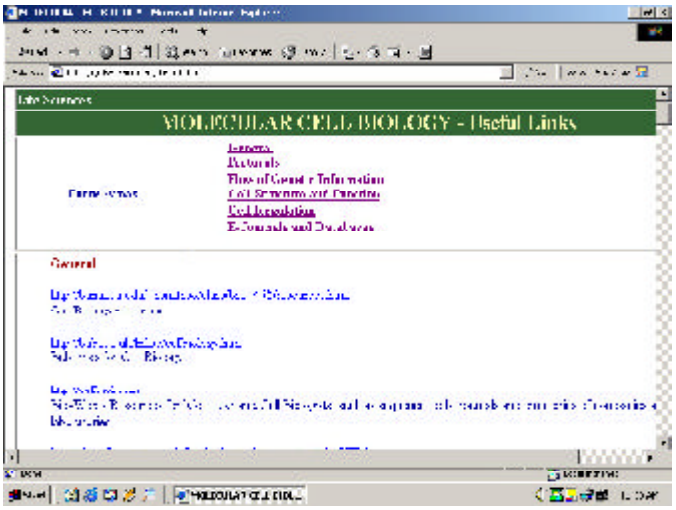


Fig.10 Molecular Cell Biology

Fig.10 - Links to useful resources, virtual libraries, gateways, tutorials, image gallery, software tools, RNA and protein synthesis, cell structure, cell signaling pathway - Free downloadable slides, hyper notes, e-journals and databases are furnished.

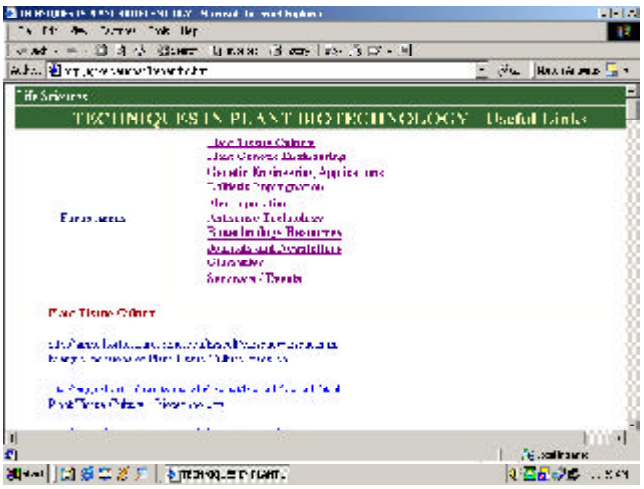


Fig.11 Techniques in Plant Biotechnology

Fig.11 - Plant biotechnology techniques such as plant tissue culture, plant genetic engineering, ballistic impregnation, electroporation, and antisense technology, biotechnology resources such as student guides, links to gateways to over 100 sites, databases, graphic gallery of process of biotechnology, biotech timeline, free resources, journals and newsletters, glossaries and links to seminars and events are the main areas for which useful links are listed.

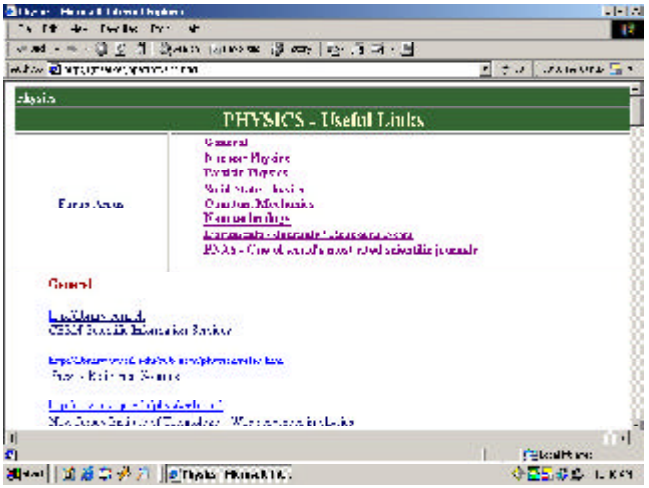


Fig.12 Physics

Fig.12 – In the above discipline Physics, links to physics resources, news, glossaries, virtual lab, virtual library, reference desk for physics, noble laureates - 1901 to present, software, directory, journals, databases, research news, eprint archives and free journal links are compiled.

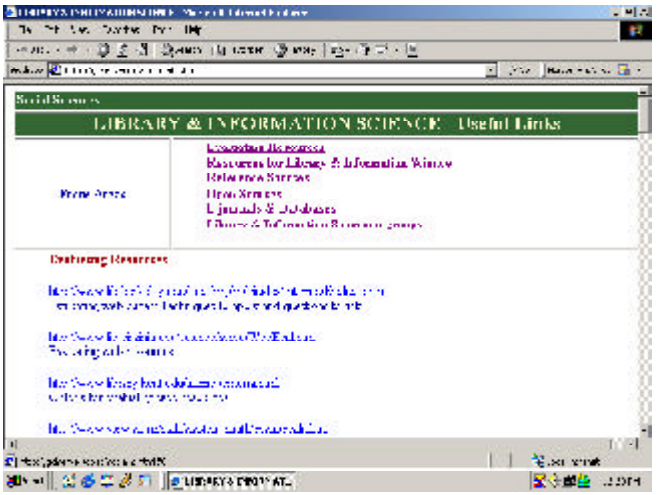


Fig.13 Library and Information Science

Fig.13 - In the Social Science category, Library and Information Science section provides links to resources for evaluating information sources, reference sources, resources for Library and Information Science, national online public access catalogue, coalition of library consortia, web tools for librarians, library services via World Wide Web, tutorial to find information on net, information skills, open sources, e-journals and databases and Library & Information Science e-groups.

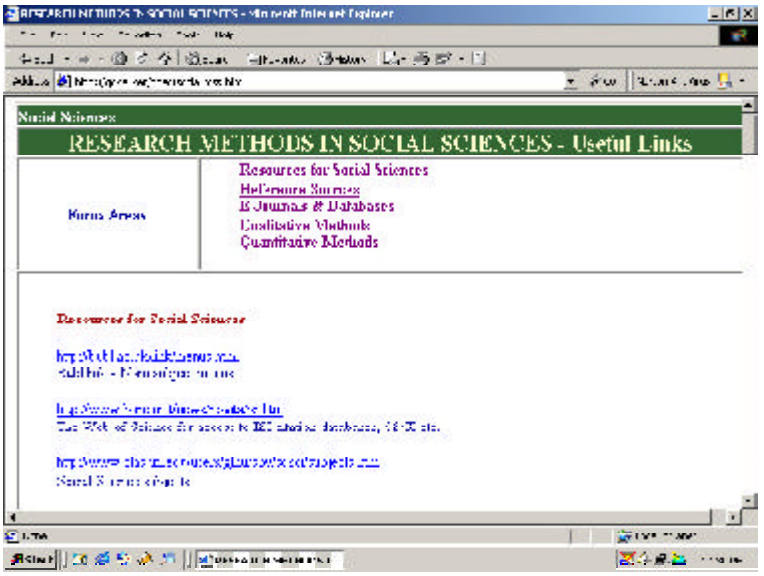


Fig.14 Research Methods in Social Sciences

Fig.14 - Research tools, Social Science Information Gateway, reference sources, ejournals and databases links are some of the areas covered, besides the links to qualitative methods such as survey resources, guide to questionnaires and survey guides, research methodology, tutorials, writing guides and qualitative data analysis software. Quantitative methods cover statistical resources, stat online textbook, free statistical software, training in statistics, statistical procedures and evaluation of information.

3. Quality of Content

When selecting resources for gateways, it should be remembered that the content must make it easy for the gateway visitor to make decisions. The content should guide the thought process or else the user is likely to abandon the site.

Secondly the user must be confident that the resources connected from the gateway are reliable and have been verifiably assured. Trust in both the resources and the delivery mechanism is an intangible but vital benefit to be conferred on the user.

According to Fisher L.M. “On the Web, most information does not have an institutional warranty behind it, which really means you have to exercise much more judgment…. If you find something in a library, you do not have to think very hard about its believability. If you find it on the web, you have to think pretty hard”.

4. Conclusion

In this above context it is essential to state here that the subject gateways compiled by Guindy Campus Library were the result of understanding the users specific needs from profiling their use of resources. Regular inputs of information requests, information searches, discussions with the science community of the campus led to development of these gateways which in turn are illustrated periodically through orientation and training schedules for Faculty, Research Scholars and Students. Many more gateways on specific themes will be periodically uploaded for the benefit of the users.

The Library has come to realize that the means of providing access to resources will be critical to the success of any digital library implementation and this will continue to be a vibrant and fast moving area of technology advancement.

5. References

1. Deegan, M and Tanner, S. (2002). Digital futures: Strategies for the information age. London: Library Association Publishing. 20,148, 161-173 pp.
2. Dussart, G. (2002). Biosciences on the Internet: a Student's guide. England: John Wiley and Sons Ltd. 173-178 pp.
3. Mac Leod, R. (2000). Comment: What's the difference between a gateway and a portal? Internet Resources Newsletter (70). http://www.hw.ac.uk/libWWW/irn/irn70/irn70.html (25.11.2004)
4. Consultative Committee for Space Data Systems. (2001) Reference model for an Open Archival Information System (OAIS). http://ssdoo.gsfc.nasa.gov/nost/isoas/us/overview.html (29.11.2004).
5. Guindy Campus Library Online Public Access Catalogue(University of Madras, Chennai) (available on intranet only at present). http://gclserver/opac.unom.ac.in

About Author



Dr. R Samyuktha designated as Deputy Librarian, University of Madras, presently heading the Science Campus. Has work 20 years of experience with opportunities of heading the Medical Science Campus of Madras University and a Post Graduate College Library earlier. She is PhD(1995) in Lib.& Inf. Sc. from University of Madras. She has contributed about fifteen publications. Specializes in "Online Information Resources in Life Sciences, Medical Sciences, Physical, Chemical, Earth and Social Sciences". Organised several Workshops, National Seminars, Conferences, Lectures, Exhibitions, and Demonstrations of access to e-content. Participated in Workshops, National and International Seminars. Successfully completed the Library Automation Project of Science Campus Library of University of Madras.
Email: samyuktharavi@yahoo.co.in

Role of Information Technology in Ayurveda in the Digital Age

G Hemachandran Nair

Abstract

The present practice of Ayurveda emphasizes on traditional way. The globalisation, patent, intellectual property rights issues and biopiracy are becoming major challenges in the indigenous traditional medical system like Ayurveda. So there is going to be crises and challenges in the Ayurveda system. In order to promote as a global medicine and equip Ayurveda to meet the global healthcare needs of the 21st century, there is an urgent need to modernise the ancient system in pace with the development of science and technology. So considering all these facts, Ayurveda is needed to restructure in the global context to meet the rising demands of a cyber society with the application of information and communication technology.

Keywords : Digital Library, Medical Information System, Ayurveda System

0. Ayurveda

Ayurveda, the ancient science of life and health, is a unique heritage of India. Ayurveda is made up of two Sanskrit words: “Ayu” which means life and “Veda” means the knowledge. Thus “Ayurveda’ in totality means ‘Science of life’. It incorporates all aspects of life whether physical, psychological, spiritual or social. What is beneficial and what is harmful to life, what is happy life and what is sorrowful life; all these four questions and life span allied issues are elaborately and emphatically discussed in Ayurveda (Gupta,1919). According to the ancient Ayurvedic scholar Charaka, “Ayu” is comprised of four essential parts. These are the combination of the mind, body, senses and the soul.

1. Basic Philosophy of Health, Disease and Treatment in Ayurveda

As per Ayurveda, ‘Health’ is a state of equilibrium of normal functions of doshas, dhatus, malas and agni with delighted body, mind and soul. It means that when Doshas-Dhatus-Malas and Agni are constantly in a state of functional equilibrium, then the health is maintained. Otherwise distortion of the equilibrium results into diseases (Dash, 1980). Erratic lifestyle is believed to be one of the basic causes behind the failure of mechanism of maintaining equilibrium. Treatment either with or without drugs and application of specific rules of diet, activity and mental status as described, disease wise, brings back the state of equilibrium i.e. health.

Fifty sixth world health assembly of WHO held in March 2003 at Geneva has mentioned that in India 65% of the population in rural areas use Ayurveda and medicinal plants to help to meet their primary healthcare needs (Sharma, 2003). In spite of its glorious past of over 5000 years as a global Medicare system, the influence of Ayurveda among the foreign public began after the Alma Ata declaration of WHO in 1980 recognised Ayurveda as an alternative system of medicine (Patel, 2000). This is because of its holistic approach and as the most user and environment - friendly system of medicine. With the changing concepts of health and disease and shifting scenario of health needs of the present times, there has been an amazing arousal of worldwide interest in Ayurveda which is likely to be accelerated with the growing trends of information technology, economic globalization and industrial activism.

2. Present Problems

After the Industrial Revolution, the rate of growth in science and technology was very fast, resulting in the inventions of computers, which are having the capacity to memorise and analyze millions of data in a nanosecond. But in practice unfortunately even 50 per cent of the available data doesn't appear to be utilized by the present day Ayurveda practitioners, perhaps with very few exceptions. Hence there is an urgent need to link computer technology and Ayurveda so that it could be utilized for present practical applications of diagnosis and treatment (Shajahan, 1998). In the present era of competition and globalisation every branch of science is trying to retain its identity in the globe by reorienting and developing itself according to the need by conducting various kinds of research. Ayurveda is also trying to prove its identity by searching newer remedies to overcome the disease for which there is no answer in modern medical science. With the growing institutionalisation of education in Ayurveda in the present century, need has been felt to launch research and development in order to update it in terms of its understanding and application to the present day need of people. The globalisation, patent, intellectual property rights issues and biopiracy are becoming a major problem in the indigenous traditional medical system like Ayurveda (Ramachandran, 2002). So there is going to be crises and challenges in the Ayurveda system. At present Ayurveda medicines are not cost-effective. Ayurveda treatment is individual based and drug production is a time consuming process. The Ayurveda medicines cannot be prepared in bulk quantity in a very short time and supplied immediately as in the case of Allopathy medicine. Another important point in Ayurveda medicine is its way of treatment. Many practitioners still resort to the traditional ways to diagnose the disease (Mathew, 1998). The rules and regulations of this sector are quite old and totally incapable to support the industry in modern developments. This system, which has proven in India for hundreds of years, has kindled the interest of the entire world and they look at it as an alternative holistic global health care system. But Ayurveda is not yet equipped to meet the challenges of the cyber society. So considering all these facts, Ayurveda needs a restructuring in the global context to meet the rising demands of a cyber society.

3. Computer based Ayurveda Practice

The potential for Information technology to help medical practitioners to perform the complex information management tasks of patient care has long been recognised. Many promising systems that incorporate advanced information technology have been developed for clinical use, with regular improvements in availability, speed, and ease to use (Gorman, 1995). The computerized Ayurveda studies have identified several important factors that affect the current and future role of computers and information technology in Ayurveda treatment. These factors include advances in information science, biotechnology and computer hardware and software, changes in the background of Ayurveda professionals, changes in the medicolegal climate and changing strategies for healthcare. At present there are few interactive Ayurveda softwares available for the diagnosis and treatment by the Ayurveda practitioners. The major computer based Ayurveda packages are:

4. Body Tune (Computerized Ayurvedic Medicare.CAM)

Body Tune, developed in 1983 is an interactive Computerized Ayurvedic Medicare software concept contributing to Ayurveda in three basic interrelated ways. It detects and communicates data about the physical conditions. It interprets that data, and actively assists in assessment and accurate diagnosis. It helps to organize the diagnostic method in a classical way envisaged by Indian Sages of Ayurveda. CAM was clinically tested by Gujarat Ayurveda University in 1993 developed by Dr. M.A. Shajahan. Its efficiency has been tested in patients and found correct. This software was particularly meant for determination of Tridosha (Vata, Pitta, Kapha) aspects only, not for any specific disease. This was the first attempt ever made in bringing computers in the field of Ayurveda (Shajahan, 1993). Itssecond and third versions came in 1988 and 1990 respectively.

Salient Features of CAM are :

- ✍ Dosha assessment : Give the signs and symptoms , after examining the persons to body tune. It will give all Doshas of body.
- ✍ Formulary: To view and search medicinal plants with its Ayurvedic properties, family, Latin Names etc. with photographs
- ✍ Rasa Guna relation: provide information about intensity and variations of rasa a and gunas
- ✍ Climatology : To know the relation of Tridoshas with climates
- ✍ Sodhana Schedule: Give awareness about various sodhana in different seasons to tune the body towards the Universe
- ✍ Weights & Measures: provide metric equivalents of classical units
- ✍ Calendar: Facility for marking schedules or appointments
- ✍ Calculator: Utility for mathematical operations
- ✍ Print:: To print case sheet or result request sheets

5. PRAKES

Prakes is an expert system for the estimation of Prakrti (body constitution) developed by CIRA (Center for Informatics Research Advancement, Kerala) in 1987. It was aimed at building a system to estimate the Prakrti of a person.

6. PRAKRTI Determination and Health Guidance by Computer

This is an expert system designed and developed by Chaitanya Consultancy, Pune in 1989. It gives users Prakrti, health advice regarding diet, instructions about daily activities, likely illness and measures for its prevention.

7. PILEX

This software is intended to diagnose the piles, its prognosis, complications and treatments. It was developed in Basic language in Gujarat Ayurveda University, Jamnagar.Gujarat in 1990.

8. MADHAVA: Ayurvedic Diagnostic System

Centre for Development of Advanced Computing, Pune has developed this diagnostics expert system based on Ayurvedic System of Medicine to diagnose a wide variety of disease in 1991. This system is developed to aid physicians in cases when the necessary information for a precise diagnosis is unavailable.

The system is capable of on – line learning as well as updating, thereby providing a scope for upgrading the system. In this system, the physician would conduct an interactive dialogue about the patient by proving information and responding to the questions generated by the system. The output of the system is a list of possible diagnosis with a certainty greater than a predefined level. The system acts as an advisor, and the physicians have the final responsibility about diagnosis of the disease as well as administration of the medicine and treatment.

9. RASEX

This package was developed by Government Ayurvda College, Trivandrum, CIRA, and ER & DC, Trivandrum in 1992. In this package an attempt has been made to correlate the pharmacological properties with that of therapeutic properties with the help of computer. A database was created after collecting, organizing and storing all the pharmacological and therapeutic properties of single rasa drug using DBase III plus. A list of drugs, which conforms to the physician's specifications is collected and displayed.

10. Role of Information Technology in Ayurveda

The process of restructuring of Ayurveda for the modernization and globalisation, the application of Information and Communication Technology (ICT) is very much required. This complex process of application of IT in the treatment and production of Ayurveda medicine needs to be studied in detail with sound theoretical and methodological foundations. However this question of developing theories and methodology poses a great challenge to Ayurveda practitioners and information technologists at international level (Nair, 2003). In this age to meet the healthcare demands of the world community, the interactions between Ayurvedic medicine and Allopathic medicine is essential. For the smooth interactions between them, application of ICT in Ayurveda is quite essential. The latest technological aids used for diagnosis and treatment in modern medicine should be used in the Ayurveda medicine also. The standardisation and production of quality drugs are important in view of export market also. There is a quantum jump of Indian Ayurvedic medicines, plants and products in international market during the couple of years which shows a tremendous growth rate. The U. S spends around $30 billion on alternate medicines every year. Consumers in Europe, too spends 13 billion Euros annually on herbal medicines. The global Ayurveda industry is estimated to be worth $62 billion annually. These mind-boggling figures justify the economics of applying the ICT on Ayurveda in its treatment, drug production and online product sales. The recent developments in the practice of modern medicine give more importance to ICT (Thomas, 2003). Therefore Ayurveda should adopt their way for its growth. The ICT revolutionises the healthcare system through new thresholds of information connectivity and higher bandwidth. These technologies have enormous capability to enable reliable storage, retrieval , transfer of the communication elements viz. test, images, audio and visual data.

The goal of Ayurveda is to improves the quality of life by preventing, and treating disease and chronic illness. The current system of treatment concentrates on three- dimensional approach, which consists of three basic elements of healthcare. They are patients, doctors, and drug vendors and suppliers(Srinivasan, 2004). The present practice involves the interaction between the patient and the doctor. There are few new major components which can be adopted in the healthcare systems of Ayurveda are Consultant Information System, Drug Information System, Patients Information System, Knowledge base in digital format and Information and Communication Technology. The above systems help to connect distant resources to work as a part of the system (Ram Mohan 1998). Therefore the revalidation and modernisation of Ayurveda can be possible through the application of ICT and research in both fundamental and applied aspects of Ayurveda

11. Conclusion

Ayurveda is the most suitable system of medicine in which Information Technology can be applied, provided both the IT experts and Ayurveda experts have very clear idea about the potentiality of both systems. The fear of Ayurveda practitioners is that if we alter the traditionality, the system will perish. So they are reluctant to apply new technologies in the Ayurvedic system. To change their mindset, they must realise that Ayurveda has global chance in this century as the most useful alternate system of medicine with vast opportunities. The adoption of ICT in Ayurveda will enhance the interactions between Ayurveda and modern medicine. So the need for modernization of Ayurveda with the application of ICT is essential to meet the challenges of future healthcare needs of a cyber society. It is a new area where the application of ICT is more evident with regards to the modern Allopathy medicine.

9. References

1. Dash, Bhagavan (1980) Basic principles of Ayurveda. New Delhi,Concept.
2. Gupta, Nagendra Sen (1919) The Ayurvedic System of Medicine, New Delhi, Logos Press.
3. Gorman, P.N. (1995) Information needs of physicians. Journal of the American Society for Information Science. 46: 729-736.
4. Mathew, Raju M. (1998) Role of Information Technology for the sustained development to Kerala. Strategies and policies. Kelpro Bulletin. 2 (1) : 3-8.
5. Nair, Hemachandran (2003) Application of information technology in the treatment and preparation of medicine in Ayurveda with special reference to Kerala. Ph. D Thesis. Dept. of L& IS, University of Calicut.
6. Patel, Aravind (2000) Ayurveda in foreign countries. In Proceedings of the International Seminar on Ayurveda and other traditional medicines scope and challenges in 21st century. Jamnagar, Gujarat Ayurved University: 82-83.
7. Ramachandran, K.V. (2002). Globalization: Threats faced by Ayurveda industry. Aryavaidyan. 16 (1): 9-15
8. Ram Mohan (1998) Information technology in Ayurveda. Apta. 5: 20-25.
9. Shajahan, M.A. (1998). Computer and Ayurveda. Apta. 5 (1): 5-17.
10. Shajahan, M.A. (1993) Clinical evaluation of Ayurvedic pharmacological principles based on computerized Ayurvedic Medicare. Ph.D Thesis. Jamnagar. Department of Dravyuguna, Gujarat Ayurved University.
11. Sharma, Ajay Kumar (2003) Ayurvedic research aspect. In Proceedings of the 4th International Seminar on Ayurvedic Education, Research & Drug Standardization - A Global Perspective. Jamnagar, Gujarat Ayurved University :54.
12. Srinivasan, K (2004) et al. Ayurveda and information technology : A preventive and curative approach to healthcare. Sajosps. 4 (2):141-144
13. Thomas, Hilary (2003) Clinical networks for doctors and managers. British Journal of Medicine. (326): 655

About Author



Mr. Hemachandran Nair is working as Technical Assistant Kerala University Library, Kerala University, Trivandrum. He has Graduated from Aligarh Muslim University and Ph.D from Calicut University. Before joining to the University, he worked as a Librarian at Jawarlal Nehru College, Kavaratthi, Lakhadweep. He had attended two training programs at INFLIBNET center at Ahmedabad in 1997 and 2003. Visited Claremont Graduate University, California as a Visiting Scholar in the Fall 2002. At present assisting as Technical Expert in the Fulbright Educational Partnership between the two Political Science departments of Kerala and Claremont Graduate University California

Email : ghnair@yahoo.com

Institutional E-Print Repositories for Schorarly Communication : Issues and Implications

B Maharana
B K Choudhury
D K Pradhan
S K Pathy

Abstract

Institutional e-print repositories offer a strategic response to systematic problems in the existing scholarly journal system and distribution of research output by making faster communication and transformation of scholarly information over the long run. This paper introduces e-print archives in general and institutional repositories in particular. The article also discusses the purpose, architecture, elements, and issues of institutional repositories. A guideline for the design of institutional archive has also been discussed. A detailed list of major institutional archives has been presented.

Keywords : E-print archives, Repositories, Scholarly communication, Open Archive Initiative.

0. Introduction

In the current networked information environment, individually driven innovation, institutional progress and the disciplinary scholarly practices are shifting dynamically to digital medium. It is the primary duty of the academic institutions that, they would take interest in capturing and preserving the intellectual output of their faculty, students and staffs. Traditionally, the institutional libraries have been serving for preservation of the institutions' intellectual legacy and facilitating the scholarly communication. But now in this digital age, these institutional repositories have changed their model and such repositories serve for providing scholarly communication by accessing research articles, supporting the institutions and libraries, reduce the monopoly of journals by demonstrating the scientific, societal and economic relevance with research activities.

Technological growth, usefulness and its trends developed new efforts in institutional repositories. Online storage costs have dropped significantly which can be afforded by repositories. Standards like Open Archive Initiative Metadata Harvesting Protocol (OAI-MHP), progressive metadata standard like 'Dublin Core' have been extensively used as underlying infrastructure for repositories. The development of free publicly accessible journal articles and extraordinary digital work has led to digital institutional repositories system such as, DSpace (http://dspace.org), California Digital Library (CDL) & e-Scholarship Repository (http://repositories.cdlib.org/), Academic Research in the Netherlands Online (ARNO), Scholarly Publishing and Academic Resource Coalition (SPARC) (http://www.arl.org/sparc/), Dispute (http://dispute.library.uu.nl/), E-print (http://www.eprint.org), and many others. The content of e-print repositories consist of narrowly peer-reviewed journal articles, conference papers, posters, pre-prints, multimedia, dissertations and even primary data. The open access of these materials encourages e-print archives as institutional digital repository for scholarly communication.

1. What are 'e-prints'?

'E-prints' are electronic copies of academic research papers. They may take the form of 'pre-prints' (papers before they have been refereed) or 'post-prints' (after they have been refereed). They may be journal articles, conference papers, book chapters or any other form of research output. An 'e-print

archive' is simply an online repository of these materials. Typically, an e-print archive is normally made freely available on the web with the aim of ensuring the widest possible dissemination of their contents.

Leslie Chan and Barbara Kirshop (2003), in their article, have categorically classified the open e-print archives into four broad groups: individual archives or self archives, institutional archives, discipline based archives, and other special archives. There are a number of successful discipline based open access e-print archives already in existence. The best known is arXiv (http://www.arxiv.org), a service for high energy physics, mathematics and computer sciences. Another example is CogPrints (http://www.cogprints.soton.ac.uk), which covers cognitive sciences.

These subject-based centralized archives work; but so far they have only been taken up by a limited number of subject communities (Pinfield; 2002). Because of this an alternative model is being suggested by advocates of e-prints: institutional e-print archives. Institutions, are assumed to have the resources to substantiate e-print archives, they also have the organizational and technical infrastructures to support ongoing archive provision. In addition, they have direct interest in wishing to expose their research output to others as this would promote the institution's standing in the research community.

2. Institutional E-print Repositories

Institutional archives are developed, maintained, and administrated by an organization or scholarly society, commonly by institutions, such as, Universities, R&D establishments, Libraries, Museums, etc. to offer universal e-print access facilities stored in their servers. Scholarly Publishing and Academic Resource-Coalition (SPARC) has defined 'Institutional Repositories' as 'digital collections that capture and preserve the intellectual output of single and multi university community' (Crow; 2002). Similarly, Lynch (2003) is of the opinion that, "a university based institutional repository is a set of services that a university offers to the members of its community for the management and dissemination of digital materials, created by the institution and its community members". Hence, Institutional Repositories are an effective organizational commitment for long-tern preservation as well as access or distribution for digital materials and support by a set of Information Technologies.

Institutional repositories are focused on the collection and preservation of all types of research literature, scientific data, learning objects, administrative records, multimedia and any other type of collection (Harnad; 2003). Thus, Institutional e-prints repositories are the globally searchable system of distributed interoperable repositories which will impact on the scholarly communication by facilitating dissemination of research result and the e-print institutional repositories are worked under OAI-PMH umbrella. A growing number of institutions and consortia are actively engaged in setting up and running institutional repositories. A country wise list of open institutional e-print archives could be found in the Appendix-1.

3. Purpose of Institutional Repositories

The origin of the e-print archives lies in the increasing interest in alternatives to the scholarly publishing paradigms. According to Crow, Institutional Repositories have two main rationales; such as:

- Scholarly Publishing Paradigm: Institutional Repositories centralize, preserve, and make accessible by institution's intellectual capital and they will form global system of distributed interoperable repositories that will help facilitate reform of scholarly communication system.
- Institutional Visibility and Prestige: Institutional Repositories serve as indicators of academic quality by capturing, preserving and disseminating the collective intellectual capital. The intellectual product created by the researchers, faculty, and other knowledge workers of an institution, deposited in the Institutional Repository; demonstrates its scientific, social and financial value. Thus, the Institutional Repositories measure institutional productivity and prestige and increased visibility of high quality of scholarship.

4. Elements of Institutional Repository

An institutional repository is a digital archive of the intellectual product created by the faculty, research staff and students of an institution and accessible to end users both within and outside the institution. In other words, according to Crow (2002), the content of institutional repository carry the following elements:

- Institutional Defined: Institutional repositories capture the original research and other intellectual property generated by an institution's activity in many fields. In this way, it represents the historical and tangible intellectual assets and output of and institution.
- Scholarly Content: Depending on the goals of establishment of institution, an institutional repository could contain any work product generated by the institutional faculty, student, non-faculty, researchers, and staff. This material is such as electronic portfolio, teaching materials, annual report, video recording, computer programmed, datasets, photographs and digital materials etc.
- Cumulative and Perpetual: The role of Institutional Repository for scholarly communication is that the content collected is both cumulative and maintained in perpetuity; in this regard it has two roles:

  (a) In Institutional Repository what ever is deposited is protected under legal right to avoid plagiarism, copy right infringement, etc. to sustain perpetually. Hence, the cumulative nature of institutional repository is scaleable.

  (b) Institutional repository aims to preserve and make accessible digital content on a long-term basis. Digital preservation and long-term access are inextricably linked.
- Open and Interoperable: The institutional repository must provide access to broader community, user outside the institution must be able to find and retrieve information from the repository, means institutional repository must be open access. Therefore, the institutional repository system must be able to support interoperability in order to provide access with the help of search engines and other discovery tools.

5. Architecture of E-prints Repositories

The architecture (Fig.1) of e-print service is based on harvesting metadata from OAI-PMH (Open Archive Initiative – Protocol metadata Harvesting) compliant e-print repositories from different institutions, non-institutions or by persons into a centralized database. Once gathered, both the metadata and full text of e-prints will be available or passed to external web server by web supporting protocols and that will be able to enhance metadata records by (Day;):

- Adding/validating authoritative forms of *author names;*
- Automatic assigning *subject classification terms;*
- Analyze the bibliographic reference into structured forms, using the Open URL standards.

This enhanced metadata formed the basis of e-print service. It will be made available to end users in a number of ways:

- Through general search interface which is integrated with other information gateway;
- Through the developed configured discovery service by which the academic institutions and other organization are directly embed e-print with their own service.

5.1 Open Archive Initiative

Open Archive Initiative (OAI) is supported by the Digital Library Federation, Coalition for networked information environment. Its mission is to develop and promote interoperability standards that aim to facilitate the effective dissemination of content (Simpson; 2004). The OAI has given momentum to any type of institutional archives that contain e-prints of published journal papers produced in research and education institutions to enhance scholarly communication.

The OAI archives can be disciplinary or institutional. The facilitating software is the OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting) which creates the framework for interoperability between distributed e-print archive/ repositories servers by enabling metadata format to harvest and aggregate into one searchable database/ interface. The metadata format is based on the 'Dublin core' metadata standard elements (Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, and Rights). OAI compliant archive/repositories server may be full text papers or may not and the interoperability is of open (free) accessible or not (Hichcock; 2003).

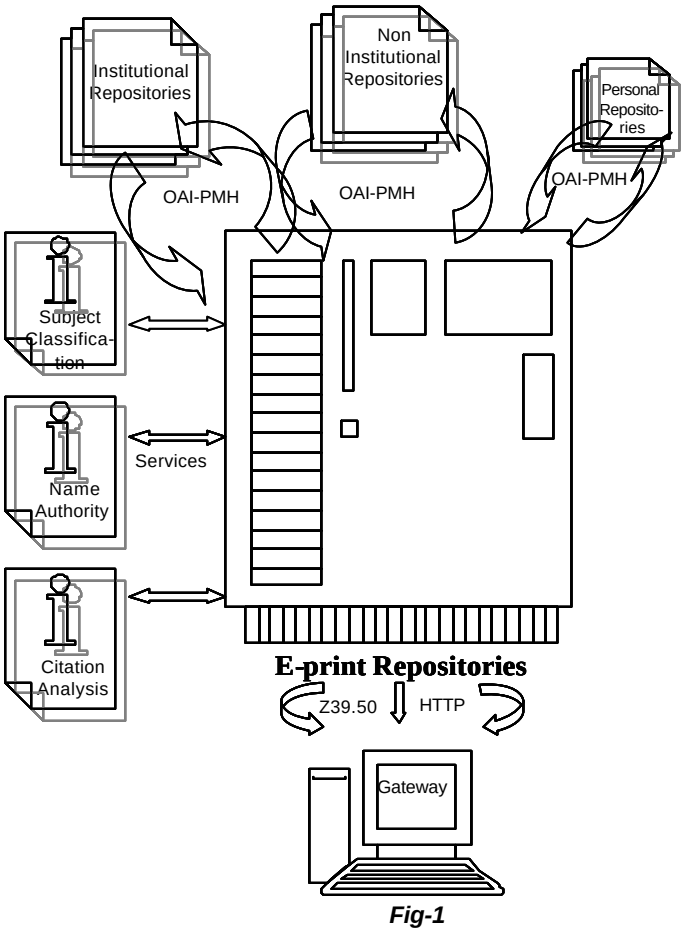


Fig-1

6. Design of Institutional e-print Repositories

Institutions can provide both the incentives and the means to spread the self- archiving e-print repositories across disciplines by facilitating the following incentives (Harnad; 2003):

- Installation of 'OAI-complaint e-print archives' using various free software that confirm to OAI – PMH. This will guarantee interoperability of all such e-print archives, as if all the papers deposited there are one seamless global archive, accessible to and navigable by any and every where.
- Adoption of a policy that all faculties (members) maintain and update standardized online curriculum vitae (CV) for annual review;
- Mandate that the full digital text of all refereed publications should be deposited in the institutional archive and linked to their entry in the authors' online CV.
- Train digital librarian/staffs are employed for assisting in self archiving (proxy self-archiving) to the authors who feels that he is personally unable to self-archive for them;
- Digital librarians, in collaboration with web system staff; provide the proper maintenance, backup, mirroring, upgrading, and preservation of e-print archives (Mirroring and migration are handled in collaboration with that counterparts/institution which supports OAI complaints e-print archives).

7. Issues and Challenges of e-print Repositories

There are number of issues related to self-archiving institutional e-print repositories, which could be related to a lack of awareness or opportunity, or a less number of records in those repositories, etc , but there are a number of practical or cultural issues which are stated as below (Pinfield:2003):

7.1 Copyright

The main obstacle in the success of institutional e-print repository is the traditional assignment of copyright to the publishers. In general cases when the paper has been accepted for publication then the author has to assign the copy right to the publisher but in case of e-print repository the exclusive license is excluded in the publication of paper.

7.2 Peer-review and quality control

Review is an essential part of the existing scientific and scholarly publishing process. But in this repository the peer-review is outside the scope which can be needed for quality control.

7.3 Long term Preservation

Long term preservation is a potential problem for e-print repositories as digital preservation has always been a challenge for all.

7.4 The popularity of traditional journal

However e-print solves the problem of serial pricing crisis and permission crisis, but the traditional published journal is more popular in scholarly communication medium because it solves the problem of copy right, quality control, long-term preservation, etc. So the popularity of traditional journal for scholarly communication medium is a partition in the way of e-print repositories.

8. Conclusion

Institutional repositories are now being recognized as a significant way of valuing and show casing as institution's intellectual assets. It is a major tool in opening access to research. OAI-compliant e-print archives are a real opportunity to improve the access to the research literature to enhance the scholarly communication process. Library and information professionals should have the vision to be leading on this important innovation taking the lead on these important developments. Although the institutional repositories have a huge potential for information communication model, they still need more testing and implementation issues, need to be explored.

Appendix-1
SPARC List of Institutional E-print Archives (Source: http://arl.org/sparc)

Sl. No.	Country	Name of the Repositories	Institution/ Organisation	Contents	System Software
1	AUSTRALIA	EPrint Repository http://www.eprints.anu.edu.au/	Australian National University	Preprint, which have been sent for publication, Post-prints, etc.	Eprint.org
2	CANADA	Papyrus http://papyrus.bib.um	Universitie de Montreal	Preprints, articles, and other research papers	Eprint.org
3	DENMARK	Electronic Library http://www.aub.auc.dk/phd/	Aalborg university	Research papers and publications of lectures and researchers (PDF)	In-house web-based
4	FRANCE	Archive Electronique http://jeannicod.ccsd.cnrs.fr/	Institute Jean Nicod	Preprints, published articles (in journals and anthologies), published correspondence	Eprint.org
5	GERMANY	Eldorado http://eldorado.uni-dortmund.de	Universitat Dortmund	Preprints, published articles(in journals and anthologies), published correspondence	Hyperwave
	GERMANY	MILESS:Die Essener Digital bibliothek http://miless.uni-essen.de/	Universitat Essen	Preprints, published articles, teaching materials, theses & dissertations, & multimedia files	MyCoRe
	GERMANY	Online publications http://elib.uni-stuttgart.de/opus/	Universitat Stuttgart (OPUS)	Preprints, journal articles, proceedings, lecture notes, these & dissertations.	OPUS System
	GERMANY	KOPS-Databank http://www.ub.uni-konstanz.de/kops/	Universitat Konstanz	Preprints, published articles, teaching materials, theses & dissertations	OPUS System
6	INDIA	http://eprints.iisc.ernet.in/	Indian Institute of Science	Preprints, post prints & others scholarly publications.	Eprints.org
7	IRELAND	Eprint archive http://eprints.may.ie/	NUI Maynoot	Preprints & post prints, research papers and other materials	Eprint.org

8	ITALY	Archive E-prints http://e-print.unifi.it/	Universita degli studi de Firenze	Didactic materials, technical reports, theses, working papers preprints as well as published articles conference papers and chapters from books.	Eprint.org
9	THE NATHER LANDS	Digital Academic Repository http://dare.uva.nl/en/	Universiteit van Amsterdam	Scientific publications, reports, preprints, articles, books, chapters, book review, inaugural lecture & dissertations.	ARNO and DLXS
	NATHER LANDS	Electronic documents http://137.120.22.236 /www-edocs/	University of Maastricht	Primary research papers	Eprint.org
	NATHER LANDS	Disputehttp://dispute. library.yy.nl/	Utrecht University	University publications (Full Text) online dissertations.	In house
10	SWEDEN	Electronic Research Archive http://www.hk-r.se/fou/	Blekinge Institute of Technology	Research papers(PDF) web based	In house &
	SWEDEN	Publications http://epubl.luth.se/	Lulea University of Technology	Abstracts describe research papers, technical reports, theses & dissertations(PDF)	In house & web based
	SWEDEN	LUFThttp://www.lub. lu.se/luft/	Lunds Universitet	Teaching materials, report series and research papers.	In house & web based
11	SWITZER LAND	CERN Document Server (CDS)		Over 630,000 bibliographic records, 250,000 full text documents, preprints, articles, books, journals, photographs and many more.	In house
12	UNITED KINGDOM	http://eprints.bath. ac.uk (UKOLN)	University of bath	Preprints and post prints research papers and others research materials.	Eprints.org
	UNITED KINGDOM	Glasgow ePrints Service http://eprints. lib.gla.ac.uk/	University of Glasgow	Full text of the research out put of university scholars, scientist and researchers.	Eprints.org
	UNITED KINGDOM	Nottingham ePrints http://eprints. nottingham.ac.uk/	University of Nottingham	Preprints, post prints and offprint of published papers.	Eprints.org
13	UNITED STATE	eSchlorship http://repositories.cd-lib.org/eschlorship/	California Digital library	Any research scholarly output by university research units, centers, or departments.	Berkeley Electronic Press

	US	CODA http://coda.caltech.edu/	Caltech	Completely scholarly research or educational materials in the final form submitted or sponsored by Caltech professional.	Eprints.org
	US	DSpace https://hpds1.mit.edu/	MIT	Digital research articles, preprints, technical reports, working papers, conference papers, images and more.	DSpace
	US	Hof Prints http://hofprints.hofstar.edu	Hofstar University	Papers written by Hofstar faculty and administrators and papers delivered by Hofstar members and conference paper sponsored by them	Eprints.org
	US	Digital Library and Archives http://scholar.lib.vt.edu/DLASPS/	Virginia Polytechnic Institute and State University	Preprint, published . articles, image, theses and dissertation	

9. References

1. Crow ,R, The Case for Institutional repositories: A SPARC Position paper.2002. <http://www.arl.org/sparc/Ir/ir.html>
2. Day, M, Prospects for institutional e-print repositories in the United Kingdom. Version 1.0, may 2003. <http://www.ukoln.ac.uk/m.day/>
3. Harnad, S, Eprints: electronic Preprint and Postprints. Encyclopedia of Library and Information science, ed. Marcel Dekker.2003. <http://www.cogsci.soton.ac.uk/~harnad/>
4. Hichcock, S, Metalist Open Access E-print Archives: The Genesis of Institutional Archives and Independent Service.2003. <http://www.arl.org/newsltr/227/ir.htm>
5. Lynch, C. A,. Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age 2003. <http://www.arl.org/newsltr/226/ir.htm>
6. Open Society Institute. A guide to Institutional repository. <http://www.soros.org)/openaccess/software/>
6. Pinfield, S, Open Archive and UK Institution: An Overview. 2003. <http://www.dlib.org/dlib/pinfild/o3pinfild.htm>
7. Pinfield, S, Gardner, M and Maccolla, J,. Setting up an institutional e-print archive.
8. Ariadne. Issue 31; 2002. <http://www.ariadne.ac.uk/issue31/eprint-archives/intro.html>
9. Simpson, P and Hey, J M N, Institutional E-Print Repository for Research Visibility. <http://www.soc.soton.ac.uk>

10. Chan, L and Kirshop, B. Open Archives Opportunities for Developing Countries: Towards a suitable distribution of global knowledge. Ariadne. 30; 2001. <http://www.ariadne.ac.uk/isssue30/oai_chan/intro.html>
11. http://dspace.org
12. http://repositories.cdlib.org
13. http://www.arl.org/sparc/
14. htp://dispute.library.uu.nl
15. http://www.eprint.org
16. http://www.arxiv.org
17. http://www.cogprints.soton.ac.uk

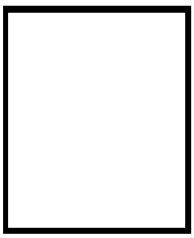
About Authors



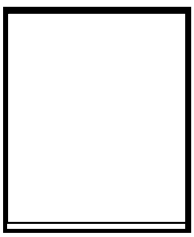
Bulu Maharana is working as a Lecturer in the Post Graduate Department of Library & Information Science, Sambalpur University, Jyoti Vihar, Orissa since 2001. He has a professional experience of working in Indian Institute of Management, Indore for more than two years. He has a good number of publications in LIS journals and presented papers in conferences.
Email : bulu_maharana@yahoo.com



Dibya Kishor Pradhan, presently working as Associate Lecturer in the Post Graduate Department of Library & Information Science, Sambalpur University, Jyoti Vihar, Orissa.
Email pradhandibya1@yahoo.co.in



Dr. B. K. Choudhury is currently Head and Coordinator, UGC-DRS-SAP Autonomous Department of Library & Information Science, Sambalpur University, Jyoti Vihar-768019. He is a product of Jadavpur, Karnatak and Utkal University. He has a professional experience of 12 years and teaching for 21 years. He has a good number of publications both in form of journal article and books.
Email : bkc_123@rediffmail.com, bkc_2008@yahoo.co.in



S. K. Pathy is working as Information Scientist in the Prof. B. Behera, Central Library, Sambalpur University, Jyoti Vihar. Earlier to this he was working as Librarian, Reliance School, Jamnagar and also worked with CEE, Ahamadabad. He has more than five years of working in the computerized library environment.
Email : skpathy@rediffmail.com

Digital Library Management in German University Libraries : The Bochum Perspective

Erda Lapp

Abstract

The presentation demonstrates cooperative approaches in the field of digital library development in Germany from the perspective of Bochum University Library. It demonstrates the integrated OPAC, the regional union catalog, a search engine for national/international catalog resources, solutions for accessing international periodicals titles in printed and in electronic form, a database of databases, a regional digital library and further relevant national initiatives. The paper argues for teaching information competence, for integrating e-learning and e-publishing into the library's information architecture and for international cooperation to further enhance information products and services.

Keywords: Digital Library, University Library, Portal

0. Introduction

In the following presentation I shall demonstrate some cooperative approaches in the field of digital library development in Germany, which show that cooperation can create much better products and services than any library could ever offer alone. I shall demonstrate these approaches from the perspective of the library I come from: Bochum University Library.

I shall also make clear that the modular structure of our digital library does not always have a seamless architecture. With e-learning and e-publishing as emerging fields in which libraries are getting involved increasingly, it becomes evident that much work remains to be done in order to offer resources and services which will fulfil our users' needs in these turbulent and rapidly changing times. It is my personal opinion that this task can only be solved cooperatively.

1. The Digital Library as a Cooperative Service

1.1 Integrated Library System / OPAC

The nucleus of our library services is the integrated electronic library system with the modules Acquisitions, Serials, Cataloging, OPAC and Circulation. Bochum University Library has created electronic cataloging records ever since its foundation in 1962. German university libraries with a longer history completed large retrospective conversion (retroconversion) projects in the 1980s and 1990s. Large and valuable resources of early and rare titles have been made accessible through these projects. Since retroconversion is expensive through the high labor cost involved, some libraries have digitized their card catalogs. We chose this technology for the card catalog with the holdings of the department libraries before 1990 on the Bochum campus. < www.ub.ruhr-uni-bochum.de >

1.2 Regional Union Catalog

In the information age libraries do not operate alone, they cooperate. So do we. Our regional union catalog is at the Interuniversity library center (HBZ) in Cologne. < www.hbz-nrw.de >

1.3 Search Engine for National / International Resources

The German National Library operates a library in Frankfurt and one in Leipzig with separate catalogs, and we have a number of regional union catalogs. A special search engine searches the national resources (national libraries and regional union catalogs) and displays the results successively after each search in a different catalog. < www.ubka.uni-karlsruhe.de/html >

1.4 International periodicals titles

German libraries have created 2 databases of international periodicals titles, one for print journals, one for e-journals.

- ZDB contains bibliographic descriptions of all print journal titles in German libraries including title changes and library holdings in each region. < www.pacifix.ddb.de:7000 >
- EZB contains all e-journal titles (commercial and open access journals) in German libraries and full text if possible. Each library offers the users a local perspective on accessible full text journals. The traffic light system indicates the local availability: green for free journals on the internet, yellow for journals for which the Bochum library has a licence and red for journals that cannot be accessed from the Bochum campus because of licence restrictions. Other libraries in the country use the same database, but their yellow and red lights will be differently placed.

Our users find that this database is extremely useful: the journal titles can be sorted alphabetically and according to subjects, a quick search entry is also available. We are not an affluent library and cannot afford many expensive commercial science journals. We think (and hope) that the future will belong to open access publications. Our library supports the SPARC (= Scholarly Publishing and Academic Resources Coalition), which was founded by research libraries in the US but has sparked to Europe.

Recently the open access initiative gained momentum on a national scale, when major science institutions signed the Berlin Declaration which calls for support and recognition of peer reviewed open access publications. < www.rzblx1.uni-regensburg.de/ezeit/fl.phtml?notation=&bibid=RUBO&colors=7 >

1.5 The Bochum Web Site / Database of databases DBIS

The Bochum library has journal contents databases and subject databases in all fields; these are accessible over our web site. < www.ub.ruhr-uni-bochum.de >

Via our web site we allow access to our catalogs including an access point for a quick search in the OPAC.

We announce current events. We list our services: the library from A-Z, reference service, e-mail reference, acquisitions requests online, information literacy classes, ILL and document delivery, Internet workstations and access for users. We share information about the library: how to find or contact us, opening hours, the library mission statement, exhibitions, projects, national and international cooperation partners, departmental libraries. The column “search” gives a floor plan of the stacks and access to universal and subject specific electronic information sources.

A cooperatively maintained database of databases with a structure similar to the database of e-journals also uses the traffic light system to indicate a campus perspective: free databases on the Internet, locally licensed databases and databases which are not accessible from the Bochum campus. The database can be sorted alphabetically or according to subjects.
< www.bibliothek.uni-regensburg.de/dbinfo/suche.phtml?bib_id=rubo&colors=15&lett=l >

1.6 Regional Digital Library

Our regional library center maintains a digital library interface which allows a metadata search in national and international library catalogs, journal contents databases, subject databases and full text databases. The metadata search does not always yield complete search results. However, the advantage of the digital library interface is, that it also offers a local view, in our case the Bochum University library view, and if the result of a search shows that a certain title is not available in Bochum, the system offers and processes an ILL request. < www.ub.ruhr-uni-bochum.de/DigiBib/digibib-nrw.htm >

1.7 More Relevant Digital Library Projects in Germany

- The national subject portal vascoda, is being built cooperatively with funds from the German Research Association. The subject portal offers valuable material and is a valuable access point. However, there is no agreement yet as to how this portal should be integrated into our existing digital library structures. < www.vascoda.de >
- Digitization projects with digitization centers in Göttingen and Munich
- Network of multimedia resource centers

Initially, in the year 2000, 15 archives, museums, documentation centers and libraries with large multimedia resources formed the nucleus of the project; meanwhile over 30 institutions from all German regions are partners in the network and cooperatively provide access to multimedia collections and materials.
< www.netzwerk-mediatheken.de >

1.8 Information Competence

Our digital library is not easy to use. But heavy use is in our interest, because the digital library costs a lot of money to maintain. We teach information competence in the Bachelor's program, and we integrate information competence in the curricula of the university departments by lecturing to freshmen.

Also, librarians take a laptop and a beamer to the cafeterias of the departments and show the students how the library can help them find information.

On an international level, we are working with the Seton Hall University Libraries in South Orange, N.J., USA, on information competence. We have been exchanging ideas and experience for 3 years now.

2. The Digital Library as a Modular Structure

It is obvious that the digital library I have presented has a modular structure, and the modules (the OPAC, regional union catalog and ILL, national resources, periodicals titles in printed and electronic form, journal contents databases, subject databases, full text databases) do not always form a seamless architecture. In spite of the digital library interface and in spite of the care we are taking to maintain a user friendly web site, there is still a classical and an electronic library with different demands and challenges to library personnel and users.

The situation is further complicated by the fact that most German university libraries are currently working on digital information issues together with the computer and /or media centers on campus. The most prominent issues are e-learning resources and e-publishing.

2.1 e-learning

The Bochum library is participating in the university's e-learning initiative together with the Computer Center, the Media Center and the Center of Continuing Education. In the framework of a project we cooperatively create, provide and maintain e-learning resources for the Archaeology Department and for an interdisciplinary neurosciences group. The university's e-learning platform is the Blackboard system. This is expensive vendor software. The library, being under constant pressure to fund expensive databases and e-journals, favours open source software out of necessity. On the other hand, Blackboard is an internationally applied platform, and since our American partners are also using it, it will facilitate international projects.

The library has not yet decided how to integrate the Blackboard system into its services. Currently, access to the e-learning resources is via the project web site, and the library has integrated its portal to course materials and e-learning resources (Virtual Book) into the Blackboard system.

2.2 e-publishing

In Germany there is a publication requirement for dissertations, and libraries have always been responsible for distributing these publications. In the last few years university departments have decided to permit electronic publication in fulfilment of the requirements, and the library accepts data files instead of print publications.

We have built up a dissertation server which is a BRS database with the functions of a data provider in the framework of the Open Archives Initiative (OAI). Currently we have 1200 dissertations online with bibliographic descriptions and full texts which can be searched and accessed from anywhere in the world. < 134.147.247.178/HSSSuchMaske/hs.cgi >

An expansion from dissertation server to publications/media server is desirable and being discussed.

On the regional level some libraries of the region and their universities are producing peer reviewed e-journals and using the publication workshow system (GAP), the document server (OPUS) and the online presentation system (Fedora) cooperatively. The systems are regionally provided and maintained by the Interuniversity Library Center in Cologne. < www.dipp.nrw.de >

3. Conclusion

I hope I have made clear that we are working on very similar problems as you and probably under similar financial constraints. It is a global village. We live in fascinating times, and libraries can be the winners of the globalization process, because they have always known the concept of sharing in order to gain. Together we shall work on solutions for even better digital libraries to serve our customers better.

4. References

1. Bochum University Library. www.ub.ruhr-uni-bochum.de

2. Interuniversity Centre, Cologne http://www.hbz-nrw.de

About Author



Dr. Erdmute Lapp, Director of Bochum University Library, Germany since 1996. She has been deputy director and head of user services at the Central Library of Juelich Research Center as well as project coordinator of several library planning projects. Her special interests are international projects on digital library development and the teaching library. She is teaching information competence in her university's Bachelor's program.

Email : erda.lapp@ruhr-uni-bochum.de

Digital Libraries and Open Source Software

Umesha Naik D. Shivalingaiah

Abstract

Open source software (OSS) is popular with technically sophisticated users, who are often also the software developers, and has not yet made a significant impact on the desktop of most users. OSS has much potential for libraries and information centres, and there are a number of projects, including Greenstone, DSpace and Ganesha, etc that demonstrates its viability in this context. OSS is becoming an increasingly popular software development method. This paper highlights what is an OSS, its features, software licensing, advantages and disadvantages. The paper also highlights the features, functions and use of three popular digital libraries software viz. Greenstone, DSpace and Ganesha.

Keywords: Digital Library, Open Source Software, Licensing, Greenstone, DSpace, Ganesha

0. Introduction

The implementation of OSS in libraries represents a method for improving library services and collections. A variety of interpretations exist with regard to the nature of OSS, sometimes confusing it with different kinds of gratis software or liberally using the term for either the development process, the software product or a particular licensing scheme. Free and OSS is also often mentioned in the same breath as open standards or interoperability, which are distinct issues in their own right.

OSS is built and enhanced through public collaboration. It is free in that it gives the user unrestricted access to the source code. The source code shows how the software works in a language that programmers can understand. In order to use OSS, users must agree to a license, which usually includes the ability to run the program, have the source code, change the source code, and distribute it. Collaboration is also how problems with the software are detected. Glitches are more easily detected when many people look at and use the software. However, some licenses restrict users from putting OSS into proprietary licensed software.

The most important aspect of the OSS is the participation of users. When a user(s) want a feature added or bug fixed for a program, they have traditionally been at the mercy of the software vendor. However, with open source they can modify the program to their own needs or fix what is broken. Many users will help develop the program for free, simply to improve the product and benefit the community.

1. What is OSS?

The term open source in common usage may also refer to any software with publicly available source code, regardless of its license, but this usage provokes strong disapproval from the OSF open source community, which may call them “disclosed source” rather than open source.

OSS means any computer software whose source code is either in the public domain or, more commonly, is copyrighted by one or more persons/entities and distributed under an open source license such as the GNU General Public License. Such a license may require that the source code be distributed along with the software, and that the source code be freely modifiable, with at most minor restrictions, such as a requirement to preserve the authors’ names and copyright statement in the code.

OSS is primarily defined as software which is freely redistributable and includes the source code. The licenses under which OSS is released vary greatly, but these two points remain consistent. This is vastly different from the mainstream software industry where source code is highly guarded and programs are only distributed in their binary, un-modifiable format.

OSS is typically created and maintained by developers crossing institutional and national boundaries, collaborating by using internet-based communications and development tools. Products are typically a certain kind of “free”, often through a license that specifies that applications and source code (the programming instructions written to create the applications) are free to use, modify, and redistribute as long as all uses, modifications, and redistributions are similarly licensed; General public licence, Berkeley Software Distribution and Mozilla Public License etc.

2. Features of OSS

The main features of OSS, and the mechanisms which drive the working of open source projects which enable these features are:

- One of the main attractive features of OSS is that its source code is available.
- It is possible to customize a particular software application according to local needs.
- Have the software at their disposal to fit it to their needs. Of course, this includes improving it, fixing its bugs, augmenting its functionality, and studying its operation.
- Redistribute the software to other users, who could themselves use it according to their own needs. This redistribution can be done for free, or at a charge, not fixed beforehand

3. OSS Licenses

With the current legal framework, the license under which a program is distributed defines exactly the rights which its users have over it. For instance, in most proprietary programs the licence withdraws the rights of copying, modification, lending, renting, use in several machines, etc. In fact, licences usually specify that the proprietor of the program is the company which publishes it, which just sells restricted rights to use it.

Authors can choose to protect their software with different licensees according to the degree with which they want to fulfill these goals, and the details which they want to ensure. In fact, authors can distribute their software with different licences through different channels. Therefore, the author of a program usually chooses very carefully the licence under which it will be distributed. And users, especially those who redistribute or modify the software, have to carefully study its licence. Under the OSS, licenses must meet ten conditions in order to be considered open source licenses:

1. Free Redistribution: the software can be freely given away or sold.
2. Source Code: the source code must either be included or freely obtainable.
3. Derived Works: redistribution of modifications must be allowed.
4. Integrity of The Author’s Source Code: licenses may require that modifications are redistributed only as patches.
5. No Discrimination Against Persons or Groups: no-one can be locked out.
6. No Discrimination Against Fields of Endeavor: commercial users cannot be excluded.

7. Distribution of License: rights must apply to everyone who receives the program.
8. License Must Not Be Specific to a Product: the program cannot be licensed only as part of a larger distribution.
9. License Must Not Restrict Other Software: the license cannot insist that any other software, it is distributed with, must also be open source.
10. License Must Be Technology-Neutral: no click-wrap licenses or other medium-specific ways of accepting the license must be required.

4. OSS Licencing bodies: Some of the common OSS licenses are as follows:

4.1 Berkeley Software Distribution (BSD)

The BSD License is similar to the GPL, but does not require derivative works to be subject to the same terms as the initial BSD License. Under the BSD Licenses, distribution of source code is permitted, but not mandated for derivative works. Programs under the BSD Licenses can be combined with proprietary software. The BSD licence is a good example of a “permissive’’ licence, which imposes almost no conditions on what a user can do with the software. The authors only want their work to be recognized. In some sense, this restriction ensures a certain amount of “free marketing’’. It is important to notice that this kind of licence does not include any restriction oriented towards guaranteeing that derived works remain open source.

4.2 General Public License (GNU) (GPL)

This is the licence under which the software of the GNU project is distributed. The GPL is based on the international legislation on copyright, which ensures its enforceability. The main characteristics of the GPL are; it allows binary redistribution, but only if source code availability is also guaranteed; it allows source redistribution (and enforces it in case of binary distribution); it allows modification without restrictions (if the derived work is also covered by GPL); and complete integration with other software is only possible if that other software is also covered by GPL.

4.3 (Mozilla Public License (MPL)

This is the licence made by Netscape to distribute the code of Mozilla, the new version of it network navigator. It is in many respects similar to the GPL, but perhaps more “enterprise oriented’’.

5. Advantages of OSS

Open source offers a radically different and exponentially better software development model. Companies can improve their products greatly and significantly increase their market share. Overall, open source is good for everyone.

- ✍ Access to source code and ability and right to modify it: The availability of the source code and the right to modify it is very important. It enables the unlimited tuning and improvement of a software product.
- ✍ Right to redistribute modifications to benefit wider community: The right to redistribute modifications and improvements to the code, and to reuse other open source code, permits all the advantages due to the modifiability of the software to be shared by large communities.

- ✍ The right to use the software in any way: This, combined with redistribution rights, ensures a large population of users, which helps in turn to build up a market for support and customization of the software, which can only attract more and more developers to work in the project.
- ✍ Cost effective: Usually, the first perceived advantage of open source models is the fact that OSS is made available gratis or at a low cost.
- ✍ Customizable: Since OSS comes with the source, one can customize existing software to suit one's needs. Open source licenses typically guarantee the right to be able to customize the software.
- ✍ Preventing re-invention of the wheel: Since we can reuse existing code, effort is not wasted re-developing software that already exists. Open source development can build on the entire body of work already released under a suitable open source license.
- ✍ Helping the progress of technology: Effort can be concentrated in making existing software better. This helps the progress of technology.
- ✍ More secure: Since the source code is open, more people scrutinize the source code and hence more flaws are found and corrected.
- ✍ Technology transfer at zero cost: Since the source code is open, anyone can learn how the software was developed, thus facilitating technology transfer at zero cost.
- ✍ Allows for easier localization: To translate a particular software package into another language using proprietary software.
- ✍ Prevent misuse of monopoly positions: The availability of the source code dictates that software vendors will always have to follow market demands and will not be able to misuse monopoly positions.
- ✍ Development advantages: With many open source projects, a virtual community of developers grows around the software. The company then incurs lower overhead because of unpaid, outsourced work and is closer to customers who use the product.
- ✍ More Programmers are Better: One would think that by having more programmers, a piece of software could be created faster and better.

6. Disadvantages of OSS

There are several disadvantages, some of which are aspects of higher life cycle costs. Because of the disadvantages listed below, open source products for the most part have become popular as black box, server-side appliances, not as interactive applications, the main disadvantages are:

- ✍ Perceived disadvantages of open source models: Of course, open source development models lead also to the perception of some disadvantages. However, some of them are only disadvantages if we stick to classical (proprietary) development models, which is of course not the case with open source.
- ✍ Limited or no accountability: Limited domain of solutions, Limited hard real-time support
- ✍ Patented Proprietary File formats: Some file formats have been patented, or for other reasons, cannot be read by Open Source products. Software patents are often given out loosely.
- ✍ Resistance to Migration: Most of the world's offices and desktops are currently using proprietary software. The migration to open source costs money and takes efforts in the short term, before long term benefits can be obtained.

- The Total Cost of Ownership Argument: For a long time, it was argued that although OSS was initially cheaper, the long term 'total cost of ownership' was higher. Increasingly, OSS is winning this argument.
- Lack of Advertising: There are only a few major proprietary software companies, and they've made a lot of money, which they can then spend on advertising.
- Fear, uncertainty and doubt: The majority of the commercial software industry finds it easier to criticize or scare people away from OSS, than embrace it, and change their business models.
- Proprietary software offering 'open' source code: Proprietary software sometimes tries to blur the line between proprietary and free or open software. This is an attempt to show that proprietary software has the same openness as OSS.
- Lack of an 'ecosystem': A problem often cited is the lack of an open source 'ecosystem', comprising lots of companies both large and small, willing to offer support etc. Major organisations need this before they are willing to use any product.
- Piracy: Piracy is common in the proprietary software world, since the legally purchased software is so expensive. Piracy makes proprietary software seem cheaper than it really is. It is sometimes alleged that proprietary software vendors 'look the other way' in developing countries when they know piracy is happening, until the country is heavily locked into the proprietary software.
- Restricted choice: In virtually every area of software there are dozens if not hundreds or even thousands of choices for different commercial packages, but rarely are there more than one or two, if any, open source options.
- Poor integration: Open source products tend to be created by people, so as a result their products are poorly integrated.
- Poor interactive capabilities: OSS with an interactive user interface as good as "average good" interactive packages in Windows.
- Difficult to use: A subset of the above that should be enumerated explicitly. OSS tend to be written by engineers for other engineers and for many of them it is accepted that ordinary function will involve creation of configuration files, writing scripts, or actually editing the source code and recompiling.
- Higher cost of installation: Commercial vendors are forced by intense competition to configure their products for easy installation. Open source tends to have much higher installation costs because a much greater degree of expertise usually is required for installation.
- Higher cost of operation: Open source products tend to require a much higher degree of technical expertise to operate and maintain, so they end up costing more.
- Higher cost of technical support: Open source costs more to support because the software is typically self-supporting.
- Lack of capabilities/features: Open software packages tend to have far fewer features and capabilities than commercial equivalents.
- Poor customer response: A well-run commercial software company will immediately turn around customer requests for enhancements. With open source, if you don't do it yourself you are at the mercy of a disjoint community of developers.
- Lack of innovation/codification of obsolete architectures: The glacially slow pace of development within open source movements and the design by committee, consensus process tends to assure that obsolete architectures get implemented within open source.
- No warranty: There is no single company backing the product.

7. Popular OSS

7.1 Greenstone software (http://www.greenstone.org/english/home.html.)

It provides a new way of organizing information and publishing it on the Internet or on CD-ROM. It is open source, multilingual software, issued under the terms of the GNU. The system operates under UNIX, Windows, and Mac OS/X, and works with standard Web servers. The Unicode character set is used throughout, so documents - and interfaces - can be in any language. It builds collections with effective full-text searching and metadata-based browsing facilities that are attractive and easy to use. Moreover, they are easily maintained and can be augmented and rebuilt entirely automatically. The system is extensible: software "plug-in" accommodate different document and metadata types.

The aim of the Greenstone software is to empower users, particularly in universities, libraries, and other public service institutions, to build their own digital libraries. The latest version of the software is 2.52 released on October 2004. Interfaces available for the Greenstone digital library software (version 2.51 only): the four "core" languages English, French, Spanish, Russian; and interfaces for Arabic, Armenian, Chinese, Croatian, Czech, Dutch, Farsi, Finnish, Galician, Georgian, German, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Kannada, Kazakh, Maori, Portuguese (Brazil), Portuguese (Portugal), Serbian, Thai, Turkish, Ukrainian.

The Greenstone "Collector" is an interactive subsystem for managing and accessing collections.

The Collector can be used to:

- create a new collection with the same structure as an existing one;
- create a new collection with a different structure;
- add new material to an existing collection;
- modify the structure of an existing collection;
- delete a collection;
- write an existing collection to a self-contained, self-installing Windows CD-ROM.

Greenstone is:

- Widely accessible: Collections are accessed through a standard web browser.
- Multi-platform: Collections can be served on Windows and UNIX, with an external Web server or (for Windows) a built-in one.
- Metadata-driven: Browsing (and, if desired, searching) indexes are built from metadata. Metadata may be associated with each document or with individual sections within documents. It must be provided explicitly (often in an accompanying XML or spreadsheet file) or derivable automatically from the source documents.
- Extensible: Plugins can be written to accommodate new document types. Classifiers can be written to create new kinds of browsing indexes based on metadata.
- Multi-language: Unicode is used throughout and is converted on-the-fly to an encoding supported by the user's Web browser. Separate indexes can be built for different languages: a plug-in allows automatic language identification for multilingual collections.

- International: The interface is available in multiple languages: new ones are easy to add.
- Large-scale: Collections containing millions of documents, and up to several gigabytes, have been built. Full-text searching is fast. Compression is used to reduce the size of the indexes and text
- Z39.50 compatible: The Z39.50 protocol is supported for accessing external servers and (under development) for presenting Greenstone collections to external clients.

Greenstone provides:

- Flexible searching: Users can search the documents' full text, choosing between indexes built from different parts. Queries can be ranked or Boolean; terms can be stemmed or unstemmed, case-folded or not.
- Flexible browsing: Users can browse lists of authors, lists of titles, lists of dates, hierarchical classification structures, and so on. Different collections offer different browsing facilities, determined at build time.
- Zero maintenance: All structures are built directly from the documents themselves. New documents in the same format can be merged into the collection automatically. No links need be inserted by hand, but existing hypertext links in the original documents, leading both within and outside the collection, are preserved.
- Phrases and key phrases: Standard classifiers create phrase and key phrase indexes of text — or indeed any metadata.
- Sustained operation: New collections can be installed without bringing the system down. Even active users rarely notice when a collection is updated.

Greenstone enables:

- Multimedia: Collections can contain pictures, music, audio and video clips. Currently, non-textual material is either linked in to documents or accompanied by written descriptions to allow access. However, the architecture allows plugins and classifiers to be written for generalized documents.
- CD-ROM option: Collections can be published on a self-installing CD-ROM. A multi-disk solution has been implemented for larger collections.
- Distributed collections: Collections served by different computers can be presented to users as though they were part of the same library, through a flexible process structure.
- What you see — you can get!: Greenstone is available from the New Zealand Digital Library (http://www.nzdl.org) under the terms of the GNU. It is easy to install on Windows and UNIX.
- Easy modify: And last but not least, because Greenstone is OSS, it is easily modified!.

7.2 DSpace: Open Source Digital Library (DL) System (http://www.dspace.org)

DSpace is a groundbreaking digital institutional repository that captures, stores, indexes, preserves, and redistributes the intellectual output of a university's research faculty in digital formats. It manages and distributes digital items, made up of digital files (or bit streams) and allows for the creation, indexing, and searching of associated metadata to locate and retrieve the items.

DSpace was designed and developed by Massachusetts Institute of Technology (MIT) Libraries and Hewlett-Packard (HP). DSpace was designed as an open source application that institutions and organizations could run with relatively few resources. It is to support the long-term preservation of the digital material stored in the repository. It is also designed to make submission easy. The latest version of the software is 1.2.1 beta2 released in November, 2004.

DSpace support the types of content

DSpace accepts all manner of digital formats. Some examples of items that DSpace can accommodate are:

- Documents, such as articles, preprints, working papers, technical reports, conference papers
- Books
- Theses
- Data sets
- Computer programs
- Visualizations, simulations, and other models
- Multimedia publications
- Administrative records
- Published books
- Overlay journals
- Bibliographic datasets
- Images
- Audio files
- Video files
- Reformatted digital library collections
- Learning objects
- Web pages

Institutional Repository

DSpace is a digital library system to capture, store, index, preserve, and redistribute the intellectual output of a university's research faculty in digital formats.

- DSpace is organized to accommodate the multidisciplinary and organizational needs of a large institution.
- DSpace provides access to the digital work of the whole institution through one interface.
- DSpace is organized into Communities and Collections, each of which retains its identity within the repository.
- Customization for DSpace communities and collections allows for flexibility in determining policies and workflow.

Digital Preservation

DSpace identifies two levels of digital preservation. First one is Bit preservation, which ensures that a file remains exactly the same over time - not a single bit is changed - while the physical media evolve around it. - Functional preservation goes further: the file does change over time so that the material continues to be immediately usable in the same way it was originally while the digital formats (and physical media) evolve over time. One of the primary goals of DSpace is to preserve digital information.

- DSpace provides long-term physical storage and management of digital items in a secure, professionally managed repository including standard operating procedures such as backup, mirroring, refreshing media, and disaster recovery.
- DSpace assigns a persistent identifier to each contributed item to ensure its irretrievability far into the future.
- DSpace provides a mechanism for advising content contributors of the preservation support levels they can expect for the files they submit.
- For all three levels, DSpace does bit-level preservation so that “digital archaeologists” of the future will have the raw material to work with if the material proves to be worth that effort.
- Access Control: DSpace allows contributors to limit access to items in DSpace, at both the collection and the individual item levels.
- Versioning: New versions of previously submitted DSpace items can be added and linked to each other, with or without withdrawal of the older item. Multiple formats of the same content item can be submitted to DSpace, for example, a TIFF file and a GIF file of the same image.
- Search and Retrieval: The DSpace submission process allows for the description of each item using a qualified version of the Dublin Core metadata schema.

Benefits of using DSpace

- Getting your research results out quickly, to a worldwide audience.
- Reaching a worldwide audience through exposure to search engines such as Google
- Storing reusable teaching materials that you can use with course management systems
- Archiving and distributing material you would currently put on your personal website
- Storing examples of students’ projects (with the students’ permission)
- Showcasing students’ theses (again with permission)
- Keeping track of your own publications/bibliography
- Having a persistent network identifier for your work, as shown in this image:
- No more page charges for images. You can point to your images’ persistent identifiers in your published articles.

7.3 Ganesha Digital Library (URL: http://gdl.itb.ac.id/)

Ganesha Digital Library enables institutions or personals to share their knowledge as well as simultaneously access and utilize knowledge. Ganesha Digital Library or GDL is a tool for managing and distributing digital collection using web-based technology. GDL enables institutions or persons to share

their knowledge as well as simultaneously access and utilize knowledge in Indonesian “giant memory” in the form of network of Indonesia DLN digital libraries. The latest version of the software is 4.0 released in June, 2004.

Features of Ganesha digital library software

- Distributed Knowledge Management: Knowledge management done through distribution, by partner in each digital library server.
- Centralized Knowledge Distribution: To make information closer to user, GDL Partner Server can benefit GDL Hub Server (Central Server) in order to disseminate metadata to all Digital Library Partner Server within IndonesiaDLN.
- Online Member Registration: User registration can be done online on the web. Validation number sent by email to make sure to make any contact to user in the future.
- Roaming Membership: Once user is registered in any GDL server, he/she can use his/her account in every online GDL server.
- Searching: GDL 3.1 supports fast information searching and detail to all managed metadata.
- Category-Based Organization: Organizing collection done with creating category and sub-category. This makes browsing easier.
- Upload Metadata and Files: Every member can publish his/her digital collections by submitting metadata form and upload the file easily.
- Personal Directory: Every member automatically possesses personal directory which he/she can freely manage.
- Review Forum: Every uploaded article can be set whether asked for review by visitor or not. Visitor can post and read review. Contributor will receive email notification if there is any review posted.
- Access Restriction: Uploaded articles can be arranged whether accessed by Intranet (any particular group) or open to the Internet.
- Image Thumbnail: Image file (jpg and PNG) can be appeared in smaller size (thumbnail) at abstract page.
- Knowledge Organization: Member, editor, and knowledge officer can organize where to put uploaded articles to appropriate categories in regard to their privileges.
- News: Editor and knowledge officer can upload fresh news to appear in GDL News GDL easily.
- Synchronization: GDL Partner Server can upload and download file and metadata to/from GDL Central Server through Synchronization facilities. Membership and publisher information can also be synchronized.
- Member and Group Administration: Administrator can manage member data, create group, and regulate editor access right.
- Statistics: Administrator can view statistics of knowledgebase content and its contributor.
- Advertisement: Administrator can show advertisement banner completed with keyword and subject matching facilities.

- Dublin Core / IndonesiaDLN Metadata: GDL utilizes IndonesiaDLN Metadata Standard that is based on Dublin Core metadata standard. It opens possibilities of information exchange with other system on the Internet that also utilizes Dublin Core.
- XML Based Transaction: Data transaction between client and server within GDL-Network uses XML format. It makes it possible for further development of GDL to become more extensive web-based networking application in the future.
- CD-ROM Enabled: GDL uses Apache, MySQL, and PHP free-software that can be run directly from CD-ROM to make easy information dissemination.
- CD-ROM Enabled

Ganesha Purposes:

- Managing scholar resources: theses, dissertations, research reports, journal, publication, etc.
- Promoting the SME's products: E-Mall (currently the e-transaction is not supported).
- Managing the art work and heritage resources: pictures, songs, videos, etc.
- Managing the expertise directory of people and organizations.
- Extend the metadata schema for other purposes easily.
- And the most important, develop distributed knowledge repository network.

8. Conclusion

It gives library staff an option to be actively involved in development projects, and this involvement can take many forms, such as reporting bugs, suggesting enhancements, and testing new versions. Organisations adopting OSS will need to provide their staff with additional development and training to enable them to take on these new roles effectively, and will need to have a long-term commitment to the projects. Currently available open source projects cover application areas ranging from the traditional library management systems to innovations like Greenstone, DSpace and Ganesha, which complement traditional systems.

OSS is well worth considering, particularly for stand-alone applications that complement traditional commercial library management systems. Systems librarians and library managers should watch this trend for future developments. The most important resource for the whole exercise is staff time and expertise. Although there is a lot of hi-tech and computers involved in creating and running a digital library, most of it is hard work. Resources for emergencies need to be considered and contingency plans (stand-by machine(s), access to temporary staff, etc.) need to be made.

OSS is any software whose code is available for users to look at and modify freely All Open Source projects have an owner All Open Source projects are governed by some type of license agreement: General public licence, Berkeley Software Distribution and Mozilla Public License etc.

9. References

1. ACM Digital Library. Retrieved October 01 2004 from http://www.acm.org/dl/
2. AsiaOSC Asian open source centre. Retrieved October 2 2004 form http://www.asiaosc.org/enwiki/page/Advantages_of_OSS.html

3. Association of Research Libraries. Retrieved September 30 2004 from http://arl.cni.org/

4. What is Open-Source Software? Retrieved November 10, 2004, from http://www.darwinmag.com/learn/curve/column.html?ArticleID=108

5. Digital library standards and practices. Retrieved September 25 2004 from http://www.diglib.org/standards.htm

6. Digital Resources from Library of Congress. Retrieved September 30 2004 from http://www.loc.gov/loc/ndlf/digital.html

7. DSpace: open source Digital Library (DL) system. Retrieved November 05 2004 from http://www.dspace.org/

8. The Economics of Open Source Software. Retrieved September 25 2004 from http://www.cs.virginia.edu/~pev5b/writing/econ_oss/index.html

9. Free Software / Open Source: Information Society Opportunities for Europe. Retrieved September 25 2004 from http://eu.conecta.it/paper/Contents.html

10. Ganesha: the first web-based digital library software in Indonesia. Retrieved November 05 2004 from http://gdl.itb.ac.id

11. Greenstone Digital Library. Retrieved November 05 2004 from http://www.greenstone.org/english/home.html

12. IEEE Computer Society Digital Library. Retrieved September 30 2004 from http://www.computer.org/publications/dlib/

13. Open Source Initiative (OSI). Retrieved September 25 2004 from http://www.opensource.org

14. Open Source Systems for Libraries Retrieved September 25 2004 http://www.oss4lib.org

15. Sun Microsystems Digital Library Toolkit. Retrieved October 30 2004 from http://www.sun.com/products-n-solutions/edu/libraries/digitaltoolkit.html

16. World Wide Web Consortium (W3C). Retrieved September 30, 2004, from http://www.w3.org/Status.

About Authors



Mr. Umesha Naik is currently working as a Lecturer in the Department Library and Information Science, Mangalore University, Mangalore. Prior to this he has worked 8 years at INFLIBNET Centre. He obtained his BL.ISc degree from Mangalore University and MLIS from IGNOU. His areas of interest are Networking, Internet, Web Design, Digital and Electronic Libraries. He published six articles in journals and seminar/conferences.

Email : umeshai@yahoo.com



Dr. D. Shivalingaiah is a Reader in Library and Information Science, Mangalore University, Mangalore. He holds M.A. in Rural Development and MLISc from Bangalore University and Ph.D. from Mangalore University. He has successfully guided a candidate for Ph.D. programme. Presently six candidates are working under him for Ph.D. programme. He has publications in Journals and Conference Proceedings and edited books. He is presently working as Deputy Registrar (Administration) on deputation.

Email : d_shivaling@yahoo.com

Building Up Digital Resources for Effective E-learning Programmes

T Rama Devi

Abstract

Today on line education using computer-mediated communication to connect the learners and the instructors via the Internet (asynchronous learning networks) is more effective. Online queries are more in use than the traditional class tests. Tutorials conucted in traditional ways are all replaced by audio-visual conferences, e-mail, chat, listservers, news groups, simulations and guest chat on the Internet. Thus new learning modes and media is being initiated for enhancement of quality. The present paper discusses about the concept of digital library and building up of digital resources collection for effective E-Learning programmes. The importance of digitized grey literature collection to enhance the quality of e-learning programmes is highlighted. The digital exposition is a remedial answer by the little screen but extensive knowledge becomes positive and possible through a resourceful digital database.

Keywords : E-Learningl, Digital Libraries

0. Introduction

The learning and teaching methods are undergoing tremendous changes with rapid development of technology. Particularly, proliferation of the Internet and availability of powerful computers have made it possible to access any information from anywhere in the world. E-learning has become pervasive worldwide through the growth and evolution of modern technologies and is simultaneously accompanied by new applications and increased adoption by end users.

e-Learning is a combination of learning services and technology to provide high value integrated learning, anytime, anyplace. It is being presented in the market place as the next evolution of the training and education industry and the next phase in the digital revolution. It is about a new blend of resources, interactivity, performance support and structured learning activities. This methodology makes use of various technologies to enhance or transform a learning process, achieving real business and educational value, and reaching a large, more diverse learner population with minimal expenditure.

CISCO defines e-learning as “the overarching umbrella that encompasses education, information, communication, training, knowledge management and performance management. It is the web-enabled system that makes information and knowledge accessible to those who need it, when they need-anytime, anywhere”.

In its short history, e-learning has come a long way, offering increasing benefits with each interaction. E-learning can be used to reduce costs, improve quality and accelerate time to market. There are other very good reasons as well, notably manageability, flexibility, speed and learning effectiveness. E-learning can be delivered anywhere, any time, and can provide flexible models, such as just-in-time learning

The effectiveness of any education or training depends on the methods and techniques used for conveying the content. Besides, the traditional method, namely, lecture, there are a number of other methods like demonstration, group discussions, panel discussions etc. which help to improve the quality of learning. With the advent of computers and Internet with Information technology, the learning programmes could

reach the remotest corner of the country. The global changes that occurred in education lead to a shift from old learning process and there is a development of online learning.

1. Need for Digital Collection

Learners access the courseware through standard Web browsers and Multimedia Players. Many institutions and agencies are engaged in training in the form of short term courses, workshops, tutorials, etc. Most of these courses produce a printed volume. The institutions organizing such courses and the funding agencies sponsoring the same should encourage preparing the course materials in electronic form to distribute it to participants and provide network access to these resources.

With the increasing demand for online programme, many universities especially in USA are getting equipped themselves with necessary technologies suitable to initiate the e-learning programmes / Online courses/web based learning etc. Content development, organization and delivery are the key components in this programmes. The user should be able to access various digital resources comfortably during the course while preparing the assignments/exams/participating in the discussion form or quiz etc. Hence, the building up of Digital Resources plays a major role for the successful completion of the online programmes.

The information revolution not only supplies the technological horse power that drives digital libraries, but fuels an unprecedented demand for storing, organizing and accessing information. If information is the currency of the knowledge economy, digital libraries will be the banks where it is invested. Digital libraries have the potential to be far more flexible than conventional ones. At present many Digital Library initiatives have been taking place at national and international level. In the developed countries digital libraries are in the progressive stage. On the other hand in the developing countries the digital library set up is yet to be realized due to the economic factors, cost of hardware and software and other infrastructure facilities. Creation of digital resources with the establishment of digital libraries is the need of the day. Digital library technology allows organized collections of information, graced with comprehensive searching and browsing capabilities , to be created rapidly (witten et al., 2001b)

2. What is a Digital Library?

Digital library, a global virtual library, is a library of thousands of networked electronics libraries. The library must be a network based distributed system with local servers responsible for maintaining individual collections of digital documents. The basic functions of the digital libraries are:

1. Provide digital content to virtual, geographically dispersed users
2. Pull in digital information electronically from outside sources irrespective of location.
3. Online access to the external digitized information with the provision of user ID and password to the users.

Digital library encompasses two possibilities

1. Library which contains material in digital form (Digitising physical counterparts e.g. paper)
2. Library contains digital material(initial content itself is created in digital form)

2.1 Definition of Digital Library

Yang et al have defined a digital library as generally including “a large collection of objects, stored and maintained by multiple sources, such as databases, image banks, file systems, E-mail servers, and web based repositories”.

2.2 Why Digitise?

The growing impact of Information and Communication Technologies (ICT), web technologies and database techniques has compelled library and information centers to use these technologies effectively to render services. With the growing number of E-resources, it has become imperative for information providers to redefine their role in disseminating information to the users.

2.3 Approaches to Digitization

1. Retrospective conversion to convert all the existing collection from A to Z
2. Digitisation of a particular special collection or a portion of one, which is highly valued for the use of the particular institution
3. Highlight a diverse collection by digitization particular good examples of collection strength.
4. High use materials making that material are in most demand more accessible.

2.4 Types of Digital Resources

Digital resources include a wide range of material such as:

1. Collection in which complete contents of documents are created or converted into machine readable for online access.
2. Scanned images, images of photographic or printed text, etc.
3. Online databases and CD-ROM information products, particularly those with multimedia and interactive video components
4. Computer storage devices such as optcal disk,jute boxes,CD_ROM/DVD-ROM
5. Databases accessible through Internet and other Networks
6. Digital audio, video clips or full length movies

2.5 Acquisition of Digital Collection

1. Digitisation of existing important and useful print material, it also helps in preserving rare and fragile objects without denying access to those who wish to study them.
2. Link and pointers, resources which are freely available on the Internet and are of significant scholarly value can be added to library catalogues and network resources.
3. Purchased or licensed material such as electronic journals or databases. In many cases this material is not “ physically owned” by the library in the same sense that printed books or journals may be owned, but instead the library has acquired specific access right to the material on behalf of library clientele.
4. Special efforts should be made to acquire Grey Literature . These can be digitized, stored and indexed for easy access

2.6 Forms of Digital Resources

- Creating Databases of Library catalogues: Libraries should create databases of holdings and collections and provide access , these should be searchable both on Intranet and Internet
- Providing Links: Organisations should create and maintain databases of their publications and collections, generally available on their web sites and may be accessed online free provided with the hyperlink to the organizations (NIRD provides this facility, go to www.nird.org.in , on the right site there is icon of CLIC)
- Procuring reference sources/E-books (Digital Form) or providing links: General and subject reference sources such as encyclopedias, dictionaries, handbooks, e-books and atlas etc. are available in CDs. Free e-books are available on web identify and links may be provided for accessing
- Subscribing/procuring bibliographical databases on CD: There are many bibliographical databases being published world wide in the field of their interest
- Accessing Bibliographic Database Online/through web: Most of the bibliographical databases are available on the web for access against payment using password., provided with web- based interface to search
- Subscribing contents pages of journals through E-mail/accessing through web
- Getting contents pages of journals through E-mail: Leading publishes viz. Elsevier, Blackwell and Wiley are providing this service free of cost
- Subscribing/Accessing Full-Text journals (E_Journals): These would be available either online or offline. Three types of online subscriptions to procure the same 1. Journals which are totally free online 2. Journals with online access free along with print subscription 3.Journals with online access whose price is marginally less than the printed version
- Providing links to important websites: Maintaining a list of URLs of various organizations/agencies provide very useful and latest information about the work being done/undertaken in their respective organizations in their web site.
- Digital Preservation of Video Programmes: Information available on Video tapes can also be digitized.
- Documents retrieved from Internet search: The material surfed regularly from the Internet on relevant topics, download documents, photographs etc and organize systematically by subject wise or any other convenient way especially to use as a reference source
- digitization of Newspaper Clippings: The news items are scanned and kept ready for further use.
- Online Paid Content Publishers Report : Subscribe the online paid content service from the publishers, OCLC is providing this service
- Digitising Resources Generated Internally: In house R&D journals, articles published in other national and international journals, newsletters of the institution, annual reports, Directories, Technical reports generated in house, technical brochures,. Pamphlets, regular course materials, lectures, video/audio clippings of demonstrations, publications/services from the library like CAS,SDI, subject bibliographies etc. Newspaper clippings, internally created files available in various file formats like ASCII, txt, pdf, xls, html etc.

Providing access to the internally created digital resources is very important. Therefore, a suitable software and format has to be chosen and adopted in order to access the resources using various data elements. It is necessary to catalogue these resources using a suitable software which supports Dublin Core (DC) standard that is being adopted internationally to create meta data of digital resources.

3. Digital collection of Grey Literature (GL)

Care to be taken for the coverage of Grey Literature in the above digital resources. The course content provider has to put lot of efforts in acquiring, digitising and managing the GL resources. Every research organisation as apart of its activity conducts specific researches focusing on its thrust areas. As a result of these researches a variety of documents are lively to emanate in the form of research reports, occasional papers, monographs, case studies, working papers, annual reports and the like. Often it is found that these documents are produced for a limited purpose, sometimes available only in draft form and not circulated extensively. This type of literature called grey literature, constitutes 60% of the total literature produced in the field of development related sciences, according to the estimates of International Development Research Council (IDRC), Canada.. Grey Literature can be obtained on a routine basis by a variety of methods like exchange agreements with other organizations, purchases by subscription or gratis. All these documents should be available in the digital format The creation and development of exhaustive digital information resource base is essential for effective e-lerning programmes which is a major task for the content provider.

Realising the importance of GL National Institute of Rural Development (NIRD) attempts to collect GL by continuously scan the newspapers, journals and annual reports to know the existence of this variety. Extensive correspondance with the organizations and browsing the Internet sites, search engines to acquire the GL on exchange/gratis basis. This collection should be digitized, organized and disseminated to the users. The GL generated internally by NIRD is being digitized with the cooperation of city central library under One Million Project of Carnie Mellon University , USA. The digitized material is receiving in the form of CDs . This can be organized and make it access through online for the development community consists of policy makers, planners, researchers, trainers, elected representatives and NGOs while attending training programmes and short term courses.

Conclusion: Setting up a Digital Library is very important for any E-learning programmes. According to Donald Waters the “Promise of digital technology is for libraries to extend the reach of research and education, improve the quality of learning, and reshape scholarly communication.” No single institution can effectively manage and provide access to more than a small portion of the information universe. There is an increasingly diverse array of networked digital library products and services. In a digitized environment it is possible for the users to have access to a library’s own digitized collections, CD_ROM databases, online database, E-book, E-zines, digitised holdings of other libraries, Internet and its myriad resources.

4. References

1. Bhavina, J.Naik (2002). Free Digital Information Resources on Environment and Environmental Engineering. In Library and Information Networking, pp.88-96, edited by H.K.Kaul and M.D.Baby. New Delhi: Developing Library Network.pp388

2. Centre for Development of advanced Computing (2002). E-Learning through Web Technologies, Manual.Hyderabad: CDAC.pp91.

3. Chitra, M (2002) et al. Developing a Digital Library in Civil and Structural Engineering R&D Institutions. In Library and Information Networking, pp.68-88, edited by H.K.Kaul and M.D.Baby. New Delhi: Developing Library Network.pp388

4. International Development Research Council IDRC (1976). Preliminary design of an International Information System for the development sciences, DEVISIS Study Team. Ottawa:IDRC, p247

5. Johnson Sophia(2002).Information Services for E-Learning. In Library and Information Networking, pp.284-289, edited by H.K.Kaul and M.D.Baby. New Delhi: Developing Library Network.pp388

6. Rama Devi, T. (2003). Bibliographical Control of GL in Social Sciences pp274-282. In National Bibliographical Control : Problems and Prospects edited by A.A.N. Raju and L.S.Ramaiah. Hyderabad: Allied Publishers (p) LTd p.430

7. Richvalsky, James and Watkins,D (2000).Design and implementing a Digital Library. ACM Crossroads student magazine, April issue 23-27

8 Waters, D.J (2003). In www.clir.org/pubs/issues/issues04.html#dbf (Accessed on 13.12.2004)

9. Witen, L.H. (2001b) et al. Power to the people, end user building of digital library collection. Proc.ACM Digital Libraries.

10. Yang, Y et al(2002).Agent based data management in digital libraries. Parallel Computing 28(5):773-792

About Author



Dr. Rama Devi Tella is working as a Doccumentation Officer as well as Associate Professor I/C. at Centre on Rural Documentation of the National Institute of Rural Development (NIRD), Hyderabad. She holds MA , MLIS and PhD in Library Science. She has an experience of 20 years under various capacity in various institutes. She was awarded Fulbright Scholarship under the subject Information Science and Technology for the year 2003-2004 and working as Mortenson Associate of the University of Illinois, Urbana-Champaign. She has attended many Seminars/ Conferences/Workshops and presented papers at national and international level.
Email : trd@nird.gov.in

Library Portal : A Knowledge Management Tool

Daulat Jotwani

Abstract

Describes the pivotal role being played by the Central Library, the Indian Institute of Technology Bombay in supporting its march towards its vision. The library has applied knowledge management practices in organizing and providing the seamless access to the knowledge resources to help users, and in doing so has acquired core competencies in several areas. Discusses the critical factors for success of knowledge management in the library, viz., knowledge resources, knowledge (dissemination) services, human resources, sustained strategic commitment, and technology. The library portal has been described as the most popular form of the technology that provides networked information about library's collections, digital resources, web sites, and services. Explains in detail the salient features of the library portal of IIT Bombay to provide single window shopping for users. Underlines the need for an aggregator to facilitate broadcast searching across databases and search engines. Concludes that the knowledge management technologies have helped the Central Library, IIT Bombay to systematically synchronize all the critical components and to serve its users more effectively and efficiently, and thus to contribute to organizational goals.

Keywords : Portal, Knowledge Management, Digital Library

0. Introduction

Indian Institute of Technology Bombay set up in 1958, is a world class institution of higher learning and research in engineering, technology and science. It has several firsts to its credit in offering programs that are flexible and innovative with a strong focus on research. In tune with its vision, "to be the fountainhead of new ideas and of innovators in technology and science", IIT Bombay recognises that knowledge is a forward as well as backward integration of ideas, experiences, institutions, systems, skills, lessons learnt and the ability to create and add value for all stakeholders. The Central Library - a proud partner in the institute's march towards its vision, plays a pivotal role in generation, assimilation, and dissemination of knowledge by promoting knowledge exchange, strengthening innovation, creating the enthusiasm and abilities for learning, and facilitating the efficient knowledge application. It also promotes relationship in and between libraries, between library and users to strengthen knowledge internetworking and to quicken knowledge flow.

1. Knowledge management in the Central Library

Knowledge management in the Central Library means organizing and providing the seamless access to the knowledge resources to help users, librarians and administrators carry out their tasks more effectively and efficiently. In doing so, the Central Library has acquired core competencies in the following areas:

- Building comprehensive collections of world-class knowledge resources with strengths in relevant subject areas
- Sharing of relevant best-practices, case-studies, lessons-learned, etc. from both internal & external sources, forging partnership within and beyond the organization
- Creation of knowledge bases and warehouses by integrating explicit and tacit knowledge sources.

- facilitating seamless, single-point access to all resources irrespective of format, language, subject and location
- Implementation of appropriate information-communication technology (ICT) tools and techniques for acquisition, processing, dissemination and sharing of knowledge
- reorientation of library personnel to acquire newer skills and develop expertise in ICT-enabled systems and services, and other critical areas.
- provide library personnel with opportunities for open communication and participation in decision-making
- provide highest level of user-focused services
- conduct information literacy programs for users' empowerment, enabling them to do their own things.
- creative in finding new solutions and better ways of operating

The Central Library with its strengths in collection building, processing, organization and dissemination accomplished by a pool of trained and experienced professionals imbued with a service-oriented value system and the expertise in knowledge sources, knowledge users and knowledge technology has achieved the knowledge management objectives of improving services and expanding its bases of resources and users.

2. Critical components for Knowledge Management

The factors critical for success of the knowledge management in the central library, IIT Bombay can be broadly categorized into : knowledge resources, knowledge (dissemination) services, human resources, technology and sustained strategic commitment.

2.1 Knowledge Resources

2.1.1 Print resources

The print resources are the biggest and most valuable assets of the library. Its collection include Text books, Reference books, Standards, Patents, Reprint and Pamphlets, Bound volumes of journals, Technical reports, Theses and other material in science, engineering, technology, humanities, social sciences and management. This well organized collection of print resources is highly valued and heavily used by not only our own users but also by the corporate and industrial houses, and educational institutes in the region. The access to this collection is provided though Online Public Access Catalogue (OPAC).

2.1.2 Digital resources

The Central Library is one of the first few libraries in the country to obtain for its users, web-based access to bibliographic databases and full text journals. Our users have access 24 x 7 to over 9000 full text journals and several important bibliographic databases on institute-wide intranet. The digital resources facilitate browsing, searching, downloading and printing of required information without limitation of time and space. In addition to above, the library has several information sources on videocassettes, CDs and DVDs which can be used and accessed in the library premises. Details of all digital sources available can be seen in Annexure- I.

2.1.3 Electronic Theses and Dissertations (http://etd.library.iitb.ac.in)

The Central Library maintains a database of all theses and dissertations tull-text submitted online by M. Tech. and Ph. D students. All Masters Dissertations from 1999 and Ph. D. Theses from 2000 onwards are available in the database which is hosted on ETD Web server on Intranet can be accessed through our web site. The database contains over 1660 records (1450 Masters and 210 Doctoral).

2.2 Knowledge (dissemination) services

An important component of knowledge management in the library is provision of the services satisfying users requirements. The use and quality of resources and the technology application will greatly improve if appropriate user focused services are offered by the competent and service oriented staff. In the process of helping users locate relevant information, staff has amassed enormous amount of tacit knowledge about print and digital resources, users' specializations and requirements, and resources most appropriate to satisfy their needs. This knowledge has been of immense help in targeting user services. Library services are also available to the IITB Alumni, Corporate houses and engineering educational institutions. The Central Library offers following services :

- Reference and consultation
- Membership and circulation
- Document delivery service
- Information Alert Service
- Resource sharing and Partnerships

2.3 Human Resources

The library has a strong team of about 60 personnel including professional librarians and support staff who are encouraged to think independently and work collectively. A key to success is an all-round improvement of library staff's quality and positioning of the human values. To ensure the participation of staff in the knowledge sharing, collaboration and re-use, they are given visibility, recognition and credit as experts in their respective areas of specialization - while leveraging their expertise for success. Library personnel are encouraged to participate in various programs for improvement of the skills and knowledge so that they are able to implement and use newer tools and techniques of ICT as well as become more productive, effective and service-oriented. An environment of openness and free communication is maintained where staff can directly meet their seniors and discuss the issues concerning them. Application of flexible management methods facilitate giving due attention to diversity and variation of library staffs' requirements, encouraging them to participate in decision-making and consultation, and undertake more jobs so as to bring their management abilities into full play and realize organizational and personal objectives.

Library users are an important component of KM who greatly influence the policies, procedures, resources and services of the library. They are the raison d'etre for the library to innovate and improve. A regular dialogue between users and library is maintained through various mailing lists, bulletin boards and personal interaction. An orientation program is organized every year for new entrants to the IIT Bombay wherein they are given in depth knowledge of library's resources, services and other facilities available for them. The library also organizes short term training sessions for users when ever a new product or service is introduced. Several vendors or producers also conduct similar training for users. Our website also functions as an important user education tool.

2.4 Sustained strategic commitment

Strategic management has a key role to play in promoting the desired behaviour both through example and by constant communication across the organization of the importance it attaches to KM. The Central Library is the nucleus of all academic activities in IIT Bombay. It is the knowledge hub around which all teaching, learning and research activities revolve. The library receives support from the management and administrators, Senate Members, faculty and the Library Committee in all its endeavors. It receives full support in policy planning, decision making, strengthening of infrastructure, modernization and introduction of new technologies and services. The Director of the institute takes keen interest in library's affairs and is always available for the help. The library also receives generous support from the Alumni Association.

2.5 Technology

The application of ICT today is indispensable as it enlarges the scope of knowledge acquisition, processing, organization and dissemination, rises speed, reduces cost and over comes space, time, language and media barriers. It links knowledge sources with knowledge workers and creates knowledge networks. It also supports knowledge sharing, collaboration, workflow, document-management, etc. across geographical boundaries.

The Central library has adequate ICT infrastructure to streamline its operations, improve efficiency, integrate its resources and provide fast access mechanisms for dissemination and sharing of knowledge. It has 8 servers, 55 PCs and other hardware to cater to the needs of library. All PCs and servers are connected to the campus-wide network that is built around a Fiber-Optic, ATM backbone comprising of an ATM switch and 5 Powerhubs. One of the powerhubs (CC Powerhub) connects the Library to the ATM switch and the backbone. The Institute's ATM backbone, in turn is connected to 2 Mbps radio link for faster access to the Internet through VSNL gateway. An additional 512K Internet link is also available from an ISP called Software Bandwidth. This network provides 10 Mbps bandwidth to the library.

The Library has computerized all its operations using a software developed in-house, uses bar coded technology for circulation of books and has installed a 3M's electromagnetic security system. It supports electronic submission of theses and dissertations, and is planning to develop an open access repository of all institutional publications.

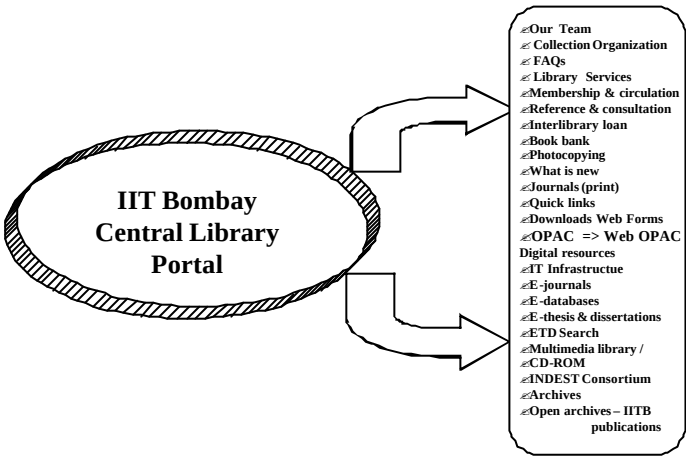
3. The Library Portal (http://www.library.iitb.ac.in)

The most popular form of KM technology that provides a secure central space where staff, users, administrators, partners and suppliers can exchange information, share knowledge and guide each other and the library to better decisions is the library portal. It is networked information space that presents the Library's collections, digital assets, Web sites, and services to its users. It allows libraries to rapidly innovate and select, organize and successfully deliver high quality Web-based content, served up through easy-to-use information discovery and management systems.

Strauss defines a portal as a special kind of gateway to Web resources—"a hub from which users can locate all the Web content they commonly need." "A portal is user-centric, while a home page is owner-centric"—in other words, the site design is built around some target community of users, rather than around the organization that hosts or "owns" the site. Elements that might appear on portals include access to various kinds of data, a search box, links, calendars or schedules, e-mail or address books, discussion groups or chat, and support for collaborative activities.

The library portal is the gateway to the Central Library providing information about its activities, functions, resources and services. The purpose of an information gateway of this type is to help our users discover high quality, relevant Web-based information quickly and effectively. The library portal has three main

components (a) it provides factual information about the staff ("Our Team"), "Collection Organization" and "Library Services", (b) it allows access to the entire collection of books, reports, theses etc available in the library through the Online Public Access Catalogue (OPAC) "Search Library Catalogue" and (c) provides direct link to full text journals, e.g. Science Direct, and bibliographic databases, e.g. COMPENDEX on publishers' site. "Multimedia library" links to CD-ROM collection available in the library. Users can download library guide ("Know Your Library"), proforma to request for book purchase, and library memberships forms from the web site. The portal provides access to list of print journals being currently subscribed to, list of all bound journals held in the library, and union catalogue of journals available in libraries of 5 IITs and BARC. M. Tech and Ph. D Students can submit their theses electronically through an intranet link provided from the website. All the theses thus received can be searched under "ETD Search" from web portal. Announcements of new activities and services are made on "What is New" which also displays recently added books and reports. Users can go to "FAQs" and find out for themselves the information about the library. "Quick links" also facilitates direct access to the desired page of the portal. User interaction is encouraged through a number of e-mail links. The schematic diagram given below illustrates all the links from the library portal:



The Library portal of IIT Bombay is one of the best examples of knowledge management which brings together all its resources and services on a single platform for convenience of its users. However, the library continues to work towards improving the portal, making it more user-oriented, interactive and customizable. It is planned to put in place an interface - an aggregator that will facilitate searching across databases and search engines – broadcast search facility, to save the users switching from one source to another. It is also being contemplated to allow users to develop their own sub portal where they receive information related only to their work.

4. Conclusion

Knowledge management has the potential to assist libraries in capturing, collecting, organizing, disseminating and sharing the knowledge and collective memory of the organization with the help of information and communication technologies. It also helps libraries streamline their day-to-day operations, improve their visibility and involvement in the organizational affairs and assume a leadership

role in helping to capture the institutional memory. The Central Library, IIT Bombay has been one of the pioneering libraries in India to adopt knowledge management technologies for serving its users more effectively and efficiently and thus contributing to institute's mission. It started adopting technology during mid 80s' and has continued to march in this direction. The current web-based technologies have been of immense help to systematically synchronize all the critical components, viz. knowledge resources, knowledge services, technology, human resources and support of the management to achieve the organizational goals. The support and inputs from our management and users have been of great value and our motivation. The success of knowledge management initiatives depends upon us if we function as learning community, have a knowledge sharing culture, versatility to accept new challenges and the ability to harness power of ICT. The Central Library, IIT Bombay has all these in abundance.

5. References

1. Calhoun K. 2002. From Information Gateway to Digital Library Management System : a case analysis. Library Collections, Acquisition, and Technical Services 26 : 141-150
2. Delu W. 1999. The Collection and Processing of Knowledge. http://www.bsti.ac.cn/bsti_kmchina/gei /048_001.htm
3. Gandhi S. 2004. Knowledge Management and Reference Services. Journal of Academic Librarianship 30 : 368-381
4. Hariharan A. 2002. Knowledge Management : a strategic tool. Journal of Knowledge Management Practice 3 : 1-8
5. Rui C. 1999. Thoughts and Technologies of Knowledge Management. Information Knowledge in Libraries 1 : 10-13
6. Shanhong T. 2000. Knowledge Management in Libraries in 21st Century. In 66th IFLA Council and General Conference, Jerusalem, Israel, 13-18 August 2000.
7. Strauss H. 2000. What is a portal, anyway? CREN (Corporation for Research, and Educational Networking) Tech Talk, January 20, 2000. http://www.cren.net/know/techtalk/events/portals.html .
8. Xiaoping S. 1999. Knowledge Management of Libraries in the 21st Century. Library Magazine 8 : 29-32
9. Yi C. 1999. The Reorientation of Libraries in the Knowledge Economy Era. Library Work and Research 3 : 4-26
10. Yunhua W. 1999. Knowledge Economy & the Development of the Library. Library Work & Research 6 : 17-19

About Author



Mr. Daulat Jotwani, presently he is working as Librarian at IIT, Mumbai, He holds B.Sc., M.Lib.Sc & Doc, M.A.,Cert. French Lang. He has professional experience of over 25 years during which he has served in National Medical Library (NML) and ICRISAT. Prior to joining the IIT Bombay in March 2004, he worked as Deputy Director & Head (1998- Feb.2004), National Medical Library, New Delhi. The main contribution of Mr Jotwani to NML include computerization of all its activities, organization of continuing education programmes for medical librarians in India & Southeast Asia. He is the recipient of the WHO Fellowship in 1997. He has visited National Library of Medicine, USA and British Medical Association Library, UK. Besides, he has also visited number of other countries Botswana, Malawi, Tanzania, Zambia and Zimbabwe. He has presented number of papers in seminar, conferences and journals. He is also a member of many professional bodies.

Email : librarian@iitb.ac.in

Annexure - I
Digital sources – full text (web based)

Source	**Publisher**	**Scope**	**No of Titles**
ABI INFORM	ProQuest Information & Learning Co (formerly UMI Company)300 North Zeeb Road Ann Arbor , MI 48103	Dateline Global Trade & Industry	171 2608 1068
PROQUEST SCIENCE	"	Journals	409
ACM DIGITAL LIBRARY	Association for Computing Machinery, 1515 Broadway New York, NY, 10036	Journals Magazines Transactions News Letters Affiliated Inst Pub SIG Proceedings	5 10 21 50 17 40 206
ACS	American Chemical Society Columbus Ohio, OH 43202 USA	Journals	18
AMS	American mathematical Society P O Box 6248 Providence RI 02940, U S A	Journals Databases	9 2
ASCE	American Society of Civil Engineers 1801, Alexander Bell Drive Reston, VA 20191	Journals	30
ASME	American Society of Mechanical Engineers International, 3 Park Avenue, New York, NY 10016, USA	Transactions + AMR	21
CRIS-INFAC	CRISIL Ltd Andheri, Mumbai 400093	Business Manage-ment Database	
EBSCO Databases	EBSCO Information Services P.O. Box 1943Birmingham, AL USA 35201	Business Source Premier Academic Search Elite	1100+ 2050+
Elsevier's Science-Direct	Elsevier Science B V Amsterdam, The Netherlands	Journals	1800 +
Emerald Full Text	Emerald Group Pub 60/62 Toller Lane Bradford England BD8 9BY	Journals	100+
Euromonitor GMID	Euromonitor Plc 60-61 Britton Street London EC1M 5UX UK	Global Market Information Digest	3500+ Companies 200+ Countries
IEL Online (IEEE + IEE)	IEEE Information Handling Services (IHS Englewood, Colorado, USA	Journals Standards Conferences /Proceedings	121 + 900 + 400 +

INSIGHT	Asian CERC Inf Tech Ltd Koramangala Layout Bangalore 560095	Company Database	8000+
IOP	Institute of Physics Publishing Dirac House, Temple Back Bristol BS1	Journals (all years) NCO Option (latest 2 years) Back Files (upto 1992)	3 all titles all titles
NATURE	Macmillan Journals UK	Journal	1
SIAM	Society for Industrial & Applied Mathematics, Univ City Sci Centre, Philadelphia PA 19104	Journals	13
Taylor & Francis	Taylor and Francis 4 Park sq., Milton Park, Abingdon Oxfordshire, OX14 4RN, UK	Journals	12
RSC	Royal Society of Chemistry Science Park, Milton Road, Cambridge, UK	Journals	18
Wiley's Polymer Science Collection	John Wiley & Sons West Sussex PO19 1UD UK	Journals	10
OUP	Oxford University Press London, UK	Journals	10
IchemE	Institute of Chemical Engineers 165-189 Railway Terrace, Rugby CV21 3HQ UK	Journals	1
ImechE	Institute of Mechanical Engineers London, UK	Journals	5
AIP	American Institute Physics New York, NY 11747	Journals	17
AIAA	American Institute of Aeronautics and Astronautics, 1801 Alexander Bell DriveSuite 500, Reston, VA 20191	Journals	5

Bibliographic databases (Web-based)

Database	**Producer**	**Coverage**	**Titles**
Compendex	Elsevier Engg Information Inc Hoboken, New Jersey, U S A	Journals, Conferences Reports	5000+
INSPEC	Elsevier Engg Information Inc Hoboken, New Jersey, U S A	Journals,Conferences Books, Reports, Dissertations	4200+
MathSciNet	American Mathematical Society P O Box 6248, Providence, RI 02940 U S A	Journals, Reviews	1800+
Scifinder Scholar	Chemical Abstracts Services American Chemical Society Columbus, Ohio, OH 43202 USA	Journals, PatentsConference Proceedings	8000+
Web of Science	Thomson Scientific Corp Philadelphia, PA, USA	Journals, Conferences	5000+
J-Gate	Informatics India, Bangalore	Journals	10000+
JCCC	Informatics India, Bangalore	Journals	4000+

Multimedia Library (intranet, CD-net)

COMPENDEX (1991-2001)	**Indian Standards**
Chemical Abstracts - 12th & 13th Collective Index	Index to Scientific & Technical Proceedings (1990-2000)
Chemical Abstracts on CD (2003-04)	INSPEC (1991-2001)
Current Contents on Disc (1999-2001)	Powder Diffraction Files
Dissertation Abstracts	Videocassettes 337

The CD-ROMs received along with books/journals/conferences are available in the Reference Section for browsing and consultation. All CDs received along with books are mirrored at CDH CD-MIRROR server and accessible through Windows machines at http://cdmirror.library.iitb.ac.in/

An Improved Hybrid Routing Protocol for Mobile Ad Hoc Networks

P Kamalakkannan A Krishnan
V Karthikeyani

Abstract

A novel routing scheme for mobile ad hoc networks (MANETs), which combines the on-demand routing capability of Ad Hoc On-Demand Distance Vector (AODV) routing protocol with a distributed topology discovery mechanism using ant-like mobile agents is proposed in this paper. The proposed hybrid protocol reduces route discovery latency and the end-to-end delay by providing high connectivity without requiring much of the scarce network capacity. On the one side the proactive routing protocols in MANETs like Destination Sequenced Distance Vector (DSDV) require to know, the topology of the entire network. Hence they are not suitable for highly dynamic networks such as MANETs, since the topology update information needs to be propagated frequently throughout the network. These frequent broadcasts limit the available network capacity for actual data communication. On the other hand, on-demand, reactive routing schemes like AODV and Dynamic Source Routing (DSR), require the actual transmission of the data to be delayed until the route is discovered. Due to this long delay a pure reactive routing protocol may not be applicable for real-time data and multimedia communication. Through extensive simulations in this paper it is proved that the proposed Ant-AODV hybrid routing technique, is able to achieve reduced end-to-end delay compared to conventional ant-based and AODV routing protocols.

Keywords : Network, Route Discovery, Data Communication, Routing Protocol, Wireless Network

0. Introduction

Current routing protocols for mobile ad hoc networks (MANETs) suffer from certain inherent shortcomings. On the one side the proactive routing schemes like Destination Sequenced Distance Vector (DSDV) [1] continuously update the routing tables of mobile nodes consuming large portion of the scarce network capacity for exchanging huge chunks of routing table data. This reduces the available capacity of the network for actual data communication. The on-demand routing protocols like Ad Hoc On-Demand Distance Vector and Dynamic Source routing [2,3] on the other hand launch route discovery, and require the actual communication to be delayed until the route is determined. This may not be suitable for real-time data and multimedia communication applications. Mobile agents similar to ants [4,5,6,7] can be used for efficient routing in a network and discover the topology, to provide high connectivity at the nodes. However the ant-based algorithms in wireless ad hoc networks have certain drawbacks. In that the

nodes depend solely on the ant agents to provide them routes to various destinations in the network. This may not perform well when the network topology is very dynamic and the route lifetime is small. In pure ant-based routing, mobile nodes have to wait to start a communication, till the ants provide them with routes. In some situations it may also happen that the nodes carrying ants suddenly get disconnected with the rest of the network. This may be due to their movement away from all other nodes in the network or they might go into sleep mode or simply turned off. In such situations the amount of ants left for routing are reduced in the network, which leads to ineffective routing.

The current paper tries to overcome these shortcomings of ant-based routing and AODV by combining them to develop a hybrid routing scheme. The Ant-AODV hybrid routing protocol is able to reduce the end-to-end delay and route discovery latency by providing high connectivity as compared to AODV and ant-based routing schemes. The hybrid scheme also does not overload the available network capacity with control messages like the proactive protocols.

1. Background Description of Aodv and Ant-Based Routing Protocols

1.1 AODV Routing Protocol

If a node using AODV [2] desires to send a message to a destination node for which it does not have a valid route to, it initiates a route discovery to locate the destination node. The source node broadcasts a route request (RREQ) packet to all its neighbors, which then forward the request to their neighbors and so on until either the destination or an intermediate node with a “fresh enough” route to the destination listed in the RREQ is located. AODV makes use of sequence numbers to ensure that the routes are loop free. Each node maintains its own sequence number, and a broadcast ID. The sequence number is incremented whenever there is a change in the neighborhood of a node and the

broadcast ID is incremented for every route discovery the node initiates. Along with its own sequence number and the broadcast ID, the source node also includes the most recent sequence number it has for the destination node. Intermediate nodes may reply to the RREQ if they have a route to the destination with a destination sequence number equal to or more than the one listed in the RREQ. If additional copies of the same RREQ are later received, these packets are simply discarded.

When the RREQ reaches the destination or an intermediate node (having fresh enough route to the destination), it responds by sending a route reply (RREP) packet to the source. Periodic HELLO broadcasts are used in AODV by the nodes in the network to inform each mobile node of other nodes in its neighborhood. These broadcasts are used to maintain local connectivity. If a node along the route moves, its upstream neighbor notices the move and propagates a link failure notification/route error message (RERR) to each of its active upstream neighbors to inform of the removal of that part of the route.

1.2 Ant-based routing

Routing algorithms for MANET which employ ants have been previously explored by [4,5,6]. Ants in network routing applications are simple agents embodying intelligence and moving around in the network from one node to the other and updating the routing tables of the nodes they visit with what they have learned in their traversal so far [4,5,6,]. Routing ants keep a history of the nodes previously visited by them. When an ant arrives at a node it uses the information in its history for updating the routing table at that node with the best routes it has for the other nodes in the network. The higher the history size the larger the overhead, hence a careful decision on the history size of the ants has to be made. All the nodes in the network rely on the ants for providing them the routing information, as they themselves do not run any program for finding routes. The ant-based routing algorithm implemented in this paper does not consider any communication among the ants. Each ant works independently. The population size of the ants is another important parameter, which affects the routing overhead. Ants that take the “no return rule” [4] while selecting the next hop at a node have been implemented in this paper. In the conventional ant algorithms the next hop is selected randomly. If the next hop selected is the same as the previous node (from where the ant came to the current node) then this route would not be optimal. Data packets sent on such routes would just be visiting a node and going back to the previous node in order to reach the destination. Every node frequently broadcasts HELLO messages to its neighbors so that every node can maintain a neighbor list, which is used for selecting the next hop by the ants.

2. Ant-Aodv Hybrid Rouring Protocol

Ant-AODV technique, forms a hybrid of both ant-based routing and AODV routing protocols to overcome some of their inherent drawbacks. The hybrid technique enhances the node connectivity and decreases the end-to-end delay and route discovery latency. Route establishment in conventional ant-based routing techniques is dependant on the ants visiting the node and providing it with routes. If a node wishes to send data packets to a destination for which it does not have a fresh enough route, it will have to keep the data packets in its send buffer till an ant arrives and provides it with a route to that destination. Also, in ant routing algorithms implemented so far there is no local connectivity maintenance as in AODV. Hence when a route breaks the source still keeps on sending data packets unaware of the link breakage. This leads to a large number of data packets being dropped. AODV on the other hand takes too much time for connection establishment due to the delay in the route discovery process whereas in ant based routing if a node has a route to a destination it just starts sending the data packets without any delay. This long delay in AODV before the actual connection is established may not be applicable in real-time communication applications.

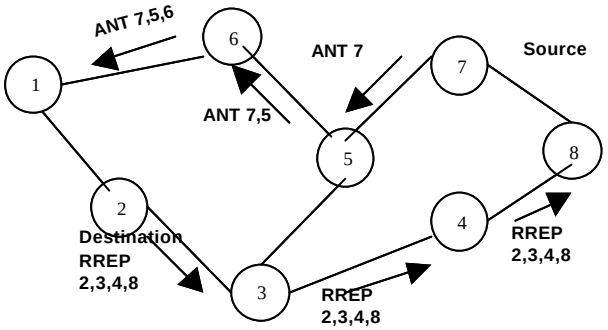


Fig. 1. Propagation of route reply and traversal of ant packet in Ant-AODV routing protocol.

In Ant-AODV ant agents work independently and provide routes to the nodes as shown in fig. 1. The nodes also have capability of launching on-demand route discovery (fig. 1) to find routes to destinations for which they do not have a fresh enough route entry. The use of ants with AODV increases the node connectivity (the number of destinations for which a node has un-expired routes), which in turn reduces the amount of route discoveries. Even if a node launches a RREQ (for a destination it does not have a fresh enough route), the probability of its receiving replies quickly (as compared to AODV) from nearby nodes is high due to the increased connectivity of all the nodes resulting in reduced route discovery latency. Lastly, as ant agents update the routes continuously, a source node can switch from a longer (and stale) route to a newer and shorter route provided by the ants. This leads to a considerable decrease in the average end-to-end delay as compared to both AODV and ant-based routing. Ant-AODV uses route error messages (RERR) to inform upstream nodes of a local link failure similar to AODV. Routing table in Ant-AODV is common to both ants and AODV. Frequent HELLO broadcasts are used to maintain a neighbor table. This table is used to select a randomly chosen next hop (avoiding the previously visited node) from the list of neighbors by the ant agents.

3. Simulation Model

Extensive simulations were carried out to compare the Ant-AODV hybrid routing protocol proposed in this paper with the conventional ant-based and AODV routing protocols. Network Simulator (NS-2) [7] is used to simulate these protocols. NS-2 is a discrete event simulator. The latest version of NS-2 (ns-2.1b8a)

which can model and simulate a multi-hop wireless ad hoc network was used for the simulations. The physical layer for the simulation uses tworay ground reflection as the radio propagation model. The link layer is implemented using IEEE 802.11 Distributed Coordination Function (DCF), Media Access Control Protocol (MAC). It uses "RTS/CTS/Data/ACK" pattern for unicast packets and "data" for broadcast packets. Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) is used to transmit these packets. All protocols simulated maintain a send buffer of 64 data packets, containing the data packets waiting for a route. Packets sent by routing layer are queued at the interface queue till MAC layer can transmit them, which has a maximum size of 50 data packets. The interface queue gives priority to routing packets in being served. The transmission range for each of the mobile nodes is set to 250m and the channel capacity is 2Mbps. Simulations were run for 600 simulated seconds. The routing table used for all the three protocols are similar. Every route entry in the routing table has a destination node address, number of hops to reach that destination, the next hop to route the packets, the sequence number of the destination and the time to live for that route.

3.1 Ant history size and ant population

Several combinations of ant population and history sizes were used in the simulations to arrive at the values that gave the best performance. These values of ant population and history size were then chosen so as to keep a balance between control overhead and efficient routing. For simulating antbased routing protocol the number of ants was kept equal to the number of nodes (which was 50) with a history size of 15. For Ant-AODV, 10 ants with a history size of 12 were used.

3.2 Mobility

A network of 50 mobile nodes migrating within an area of 1500m X 300m with a speed of 0 - 10m/s was simulated. A rectangular space was chosen in order to force the use of longer routes between nodes than would be there in a square space with the same amount of nodes [8]. The mobility model uses the random waypoint model in the rectangular field. The simulations were run multiple times for 6 different pause times: 0, 30, 60, 120, 300 and 600 seconds. After pausing for pause time seconds the mobile node again selects a new destination and proceeds at a speed distributed uniformly between 0 and a maximum speed.

3.3 Traffic

The simulations carried out consisted of 20 Continuous Bit Rate (CBR) sources. CBR traffic sources were chosen, as the aim was to test the routing protocols. Source nodes and destination nodes were chosen at random with uniform probabilities. The sending rate used was 4 packets per second with a packet size of 64 bytes. Each data point in the comparison results represents an average of multiple runs with identical traffic models but with different movement scenarios. Same movement and traffic scenarios were used for all the three protocols simulated.

4. Simulation Results

4.1 Average end-to-end delay

The average end-to-end delay includes buffering delay during route discovery, queuing delay at interface queue, retransmission delays and propagation and transfer times. The average end-to-end delay for AODV and Ant-AODV hybrid protocol (fig. 2) is very less. But in case of Ant routing technique (fig. 3) the average end-to-end delay is high. The high end-to-end delay in ant-based routing is attributed to the lack of on-demand route discovery capability of the nodes in ant routing. Due to this the packets to be sent by a node keep waiting in the send buffer till the ants visit that node and provide it with routes.

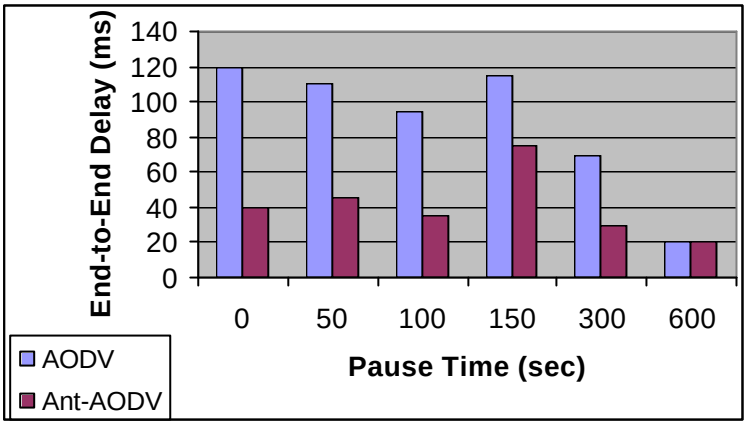


Fig. 2. Average end-to-end delay provided by AODV and Ant-AODV routing protocols.

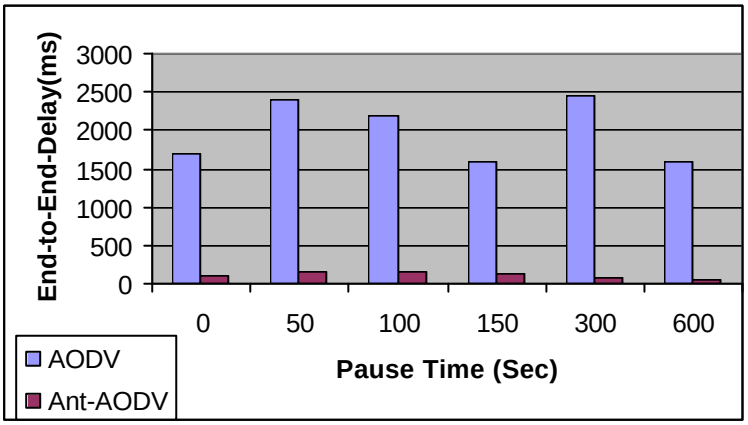


Fig. 3. Average end-to-end delay provided by Ant-based and Ant-AODV routing protocols.

Comparing Ant-AODV and AODV it can be observed that the end-to-end delay (fig. 2) is considerably reduced in Ant-AODV as compared to AODV. Ants help in maintaining high connectivity in Ant-AODV, hence the packets need not wait in the send buffer till the routes are discovered. Even if the source node does not have a ready route to the destination, due to the increased connectivity at all the nodes the probability of its receiving replies quickly from nearby nodes is high resulting in reduced route discovery latency. Lastly, the dynamic nature in which routes are kept updated by the ants leads to the source node switching from a longer (and stale) route to newer and shorter ones hence reducing end-toend delay for active routes.

4.2 Goodput and Packet delivery fraction

Goodput is the total number of useful packets received at all the destination nodes and packet delivery fraction is the ratio of number of data packets sent to the number of data packets received. Packet delivery fraction is very high for AODV and Ant-AODV (fig. 4) as compared to ant-based routing. Goodput is also higher for Ant-AODV and AODV as compared to ant-based routing (fig. 5).

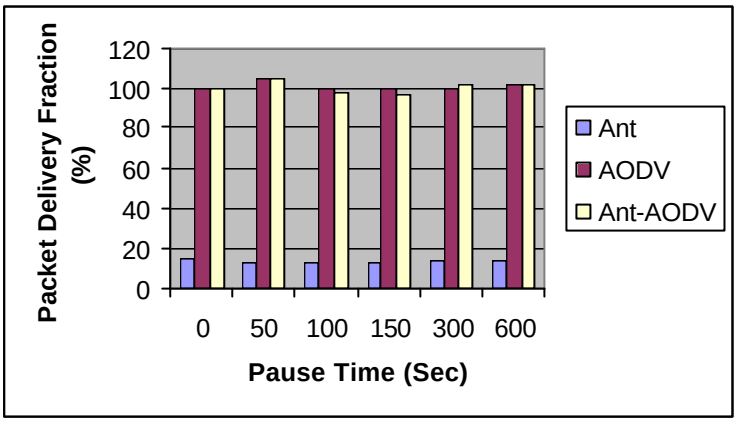


Fig. 4. Packet delivery fraction provided by Ant-based, AODV and Ant- AODV hybrid routing protocols.

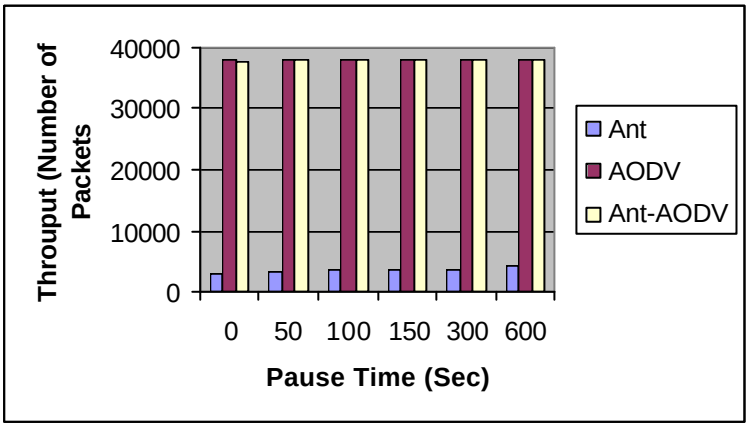


Fig. 5. Goodput of Ant-based, AODV and Ant-AODV routing protocols.

High packet delivery fraction and goodput in Ant-AODV and AODV is because they make use of link failure detection and route error messages. Whereas in case of ant-based

routing there is no such feature and so the source nodes keep on sending packets unaware of the link failures. This leads to a large amount of data packets being dropped which reduces the packet delivery fraction and the goodput. Also seen in the graphs for packet delivery fraction (fig. 4) and goodput (fig. 5) is that as the pause time increases the goodput and packet delivery ratio increase due to less link failures at low mobility rates (high pause time).

4.3 Normalized Routing overhead

Normalized routing overhead is the number of routing packets transmitted per data packet received at the destination. The routing overhead in case of ant-based routing is independent of the traffic. Even if there is no communication the ants would still be traversing the network and update the routing tables. However in case of AODV, the overhead is dependent on the traffic and if there is no communication then there will be no control messages generated in the network. In Ant-AODV the overhead has two components. It has the ants traversing in the network, and the route discovery and route reply messages being generated in case the nodes do not have routes provided to them by the ants for some destinations. From the comparison results (fig. 6) it is seen that the normalized overhead is too high in case of ant-based routing scheme. The reason for this is that the actual data packets delivered are too less and hence the ratio of control overhead to data packets delivered becomes too high. In case of AODV (fig. 6) the normalized overhead is the least. The normalized overhead (fig. 6), is slightly greater in Ant-AODV as compared to AODV because of the continuous movement of ants in the network. The continuous gradual drop in normalized routing overhead (fig. 6) for all the three protocols is attributed to the increased packet delivery fraction and goodput at higher pause times (normalized load is the ratio of total control packets generated to actual data packets received).

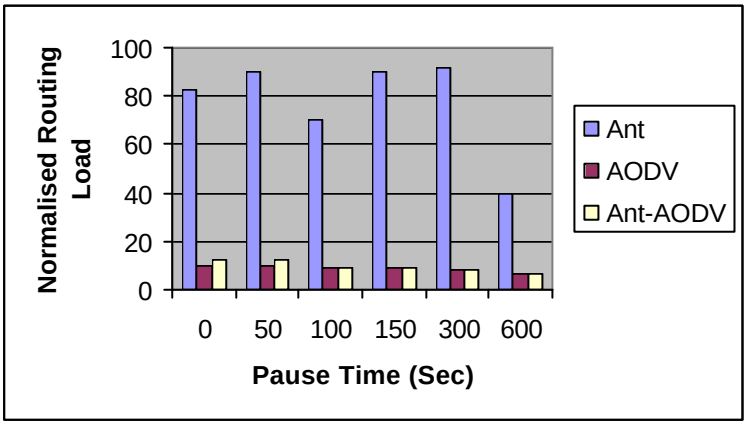


Fig. 6. Normalized routing overhead of Ant-based, AODV and Ant-AODV routing protocols.

4.4 Connectivity

Connectivity is the average number of nodes in the network for which a node has un-expired routes. In case of Ant-AODV and ant-based routing protocols (fig. 7), ant agents continuously traverse the network and update the routing table entries. Due to this, a node has fresh enough (or un-expired) routes to a large number of nodes in the network at any given point of time. Connectivity in Ant-AODV and ant-based routing schemes is more than double the connectivity in AODV (fig. 7). Higher connectivity leads to lesser route discoveries and reduced end-to-end delay.

5. Discussion

5.1 Ant Visit Period

During the simulations an important characteristic of ant agents for routing in MANETs was observed. After a certain period (nearly 100 simulation seconds), the ant activity (ant hopping from one node to the other and updating routes) would almost subside. This could be due to various reasons such as (i) the ant packets could be lost in wireless transmission, (ii) the next node which was to receive the ant packet moves out of the wireless range of the sending node, or (iii) the ant bearing node goes out of wireless range of every node in the network and there is no next hop node available for the ant. In such situations the number of ants actually available for routing purpose decreases. To overcome this decrease in number of ants available for routing, a "minimum ant visit period" was set. If no ant visited a node within this period the node would generate a new ant and transmit it to one of its neighbors selected randomly. This way the ant activity would never subside and the network would not become devoid of ants. The simulations carried out used a minimum ant visit period of 5 seconds.

5.2 Performance of Ant-AODV

It is evident from the simulation results that by combining ant-like mobile agents with the on-demand route discovery mechanism of AODV, the Ant-AODV hybrid routing protocol would give reduced end-to-end delay and route discovery latency with high connectivity. Such low end-to-end delay cannot be achieved from either of the two base protocols (antbased and AODV) because of their inherent shortcomings.

6. Conclusion

The shortcomings of on-demand routing protocols like AODV and ant-based routing have been tried to overcome in this paper by combining both of them to enhance their capabilities and alleviate their weaknesses. Ant-AODV hybrid protocol is able to provide reduced end-to-end delay and high connectivity as compared to AODV. As a result of increased connectivity the number of route discoveries is reduced and also the route discovery latency. This makes Ant-AODV hybrid routing protocol suitable for real-time data and multimedia communication. As a direct result of providing topology information to the nodes (using ants), the foundations for designing a distributed network control and management get automatically laid. Higher connectivity and reduced end-to-end delay are achieved at the cost of extra processing of the ant messages and the slightly higher overhead occupying some network capacity. However this does not adversely affect the packet delivery fraction or the goodput. The future work would be to explore the use of back up or multiple routes provided to the nodes by ants in their frequent and continuous visits to the node.

7. References

1. C. E. Perkins and P. Bhagwat, "Highly dynamic Destination Sequenced Distance-Vector routing (DSDV) for mobile computers," in Proc. of the SIGCOMM '94 Conf. on Communications Architecture, Protocols and Applications, pp. 234-244, Aug. 1994.

2. C. E. Perkins, E. M. Royer and S. R. Das, "Ad Hoc On-Demand Distance Vector (AODV) Routing," in Proc. IEEE Workshop on Mobile Computing Systems and Applications, pp. 90-100, Feb. 1999.

3. D. B. Johnson and D. A. Maltz, "Dynamic Source Routing in ad hoc wireless networks," Mobile Computing, edited by Tomasz Imielinski and Hank Korth, chapter 5, pp. 153-181. Kluwer Academic Publishers, 1996.

4. N. Minar, K. H. Kramer and P. Maes, "Cooperating Mobile Agents for Dynamic Network Routing," Software Agents for Future Communications Systems, chapter 12, Springer Verlag, 1999.

5. H. Matsuo and K. Mori, "Accelerated Ants Routing in Dynamic Networks," in Proc. Intl. Conf. On Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp.333-339, Aug. 2001.

6. R. R. Choudhary, S. Bhandhopadhyay and K. Paul, "A Distributed Mechanism for topology discovery in Ad Hoc Wireless Networks Using Mobile Agents," in Proc. of Mobicom, pp. 145-146, Aug. 2000.

7. K. Fall and K. Varadhan, editors. The ns manual. The VINT Project, UC Berkeley, LBL, USC/ISI and XEROX PARC, 2001. http://www.isi.edu/nsnam/ns/ns-documentation.html

8. J. Broch, D. Maltz, D. Johnson, Y.-C. Hu and J. Jetcheva, "A Performance Comparison of Multi-Hop Wireless Ad Hoc Network Routing Protocols," in Proc. of the Fourth Annual ACM/IEEE International Conference on Mobile Computing and Networking, pp. 85-97, Oct. 1998.

About Authors

Mr. P. Kamalakkannan is working as Assistant Professor in Department of Computer Science, K S R College of Technology, Tiruchengode, Tamil Nadu. He is also pursuing his research work.
Email : kamal_karthi96@yahoo.co.in

Dr. A Krishnan is a Professor & Head in Department of Computer Science & Engineering and Principal of R R Engineering College, Tiruchengode, Tamil Nadu. He has contributed number of papers in seminars, conferences and journals and he is also a member of professional associations.
Email : a_krishnan26@hotmail.com

Mrs. V. Karthikeyani is working as Assistant Professor in Department of Computer Science, K S R College of Technology, Tiruchengode, Tamil Nadu. She is also pursuing her research work.

Student's Perceptions Toward the Use of the Digital Library for Higher Learning

A Manoharan M Anu Vasanthi
T Deepa

Abstract

This study attempted to investigate students' perceptions toward the use of the digital library for higher learning. Attention was given to three variables namely sex, year of study, level of course. A Likert – type instrument consisting of 10 items was designed to collect information about students perceptions. The respondents of this study were 72 students enrolled in various Post-graduate courses at Bishop Heber College, Tiruchirappalli. Collected data were analysed using ANOVA (Analysis of Variance). The overall results suggest that students had positive perceptions toward the use of digital library. Sex, year and level of course were found to be significant factors. Females, first year students and social science students had significantly positive perceptions. Implications for practices are discussed and recommendations are made for future research.

Keywords : Digital Libraries

0. Introduction

The educational field has been attracted by the promise and potential of technology from the advent of films in the 1920s to television in the late 1950s, computers in the 1980s and information technology in the 1990s. In the 1980s, during the micro computer revolution in higher education, the computer emerged as a personal tool: students, faculty and institutions purchased desk top systems by the track load; emerging applications, falling prices and increased power and convenience brought the desktop and note book computers to thousands of students who never previously though of themselves as "Computer users". Most people would agree that modest productivity benefits emerged as growing number of faculty transferred much of their work from secretaries, mainframes and mini computer to desktop systems and word processors.

Midway through the 1990s however, colleges and universities benefited by the second major phase of this revolution – a shift in emphasis from computer as a desktop tool to computer as the communications gateway to contents (data bases, images, text libraries, video and more) increasingly accessible via computer networks to both the faculty and students. Information Technology supporters are fond of describing a future information rich environment that will support learning and scholarly activities in new and exciting ways.

1. The Digital Library

Drabenstott describes the concept digital library as the

- The digital library is not a single entity;
- The digital library required technology to link the resources of many;
- The linkages between the many digital libraries; and information services are transparent to the end users;

- Universal access to digital libraries is a goal; and
- Digital library collections are not limited to document surrogates; they extend to digital artifacts that can not be represented or distributed in printed formats.
- Leiner describes the digital library as the collection of services and the collection of information objects and their organisation, structure and presentation that support users in dealing with information objects available directly or indirectly via electronic / digital mean.

The term virtual libraries and electronic libraries are often used simultaneously and / or interchanged with the term digital libraries.

2. Purpose of the study

The study focuses on the use of the digital library in higher learning. The purpose of this study was to investigate students' perceptions toward the use of the digital library that used as the one of the major resources for higher learning. The perceptions of the students may be an important factor in influencing a positive learning outcome. Differences in students' perceptions toward the use of digital library were examined using three variables – gender, year of study and course.

3. Methodology

A likert-type instrument consisting of ten items was designed to measure the students' perceptions toward the use of the digital library for higher learning. The instrument represented positively worded statements that collect information about students perceptions.

- The digital library is a valuable tool
- By using digital library, I often find materials relevant to what I want.
- My time is well spent using digital library for higher learning.
- Using digital library enhances my learning.
- Use of digital library increases my ability to do a better job.
- Using the digital library increases my ability to do a better job in learning assignments.
- Using the digital library is an important part of the learning process.
- It is worthwhile using the digital library.
- I feel that I gain a lot by using the digital library.
- It is a rewarding experience to use the digital library.

The responses to the items were recorded as strongly agree=5, agree=4, neither agree nor disagree = 3 disagree = 2 and strongly disagree = 1.

4. Respondents

The respondents of this were 72 students who were enrolled in post graduate programmes in Bishop Heber College, Tiruchirappalli. The respondents were from the courses of Library and Information Science, Maths, Physics, Chemistry, Environmental Science, Biotechnology Information Technology, Computer Science, Computer applications and management. Out of 72 students, 41.66 percent were males and 58.34 were females the purpose of the study and the scoring strategy was carefully explained to the respondents. They were assured that their responses would be anonymous and confidential.

5. Research questions

This study was designed to answer to following questions.

- Is there a difference in perceptions between males and females using the digital library?
- Is there a difference in perceptions between first year and second year post graduate students?
- Is there a difference in perceptions among various course levels (Science, Computer science and Technology and Social science) of respondents using the digital library?

6. Data Analysis

Analyses of variance (three separate ANOVAs) were conducted to answer the research questions. The F statistics generated from the analysis indicate significant differences between the selected variables and students' perceptions toward the use of the digital library for their learning. A predetermined level of significance chosen for this study was 5%.

7. Results

The results of the descriptive analyses indicated that overall students perceived that the use of the digital library for higher learning was a positive learning experience (Table 1)

Table 1

Item No	N	Min Score	Max Score	Total Score	Mean	SD
01	72	2	5	323	4.486	1.607
02	72	1	5	279	3.875	1.207
03	72	2	5	296	4.111	1.333
04	72	2	5	312	4.333	1.509
05	72	1	5	271	3.764	1.219
06	72	1	5	298	4.139	1.374
07	72	2	5	313	4.347	1.541
08	72	1	5	289	4.014	1.253
09	72	1	5	296	4.111	1.333
10	72	2	5	279	3.875	0.964

The results of ANOVA yielded a significant difference in perceptions between males and females using the digital library for higher learning ($F_{0.05}$=5.32, F_1=2.15). Mean scores are presented in Table 2. Female students scored significantly than male did.

Table 2

Sex	Score	Mean	N	Percent	SD
Male	1213	4.043	30	41.66	1.333
Female	1733	4.126	42	58.34	1.374

The results of ANOVA for the level of course indicated that a significant difference in perceptions between first year and second year students ($F_{0.05} = 5.32$ $F_1 = 0.0486$). Mean scores are presented in Table 3. First year students scored significantly higher than the second year students.

Table 3

Year	Total Score	Mean	N	Percent	SD
I	1543	4.170	37	51.39	1.397
II	1361	3.888	35	48.61	1.320

The results of ANOVA for various courses showed a significant difference among various courses of the respondents using the digital library ($F_{0.05} = 4.1028$ $F_1 – 3.666$). The mean score is represented in table 4. It showed that social science students scored significantly higher than science students.

Table 4

Course	Score	Mean	N	Percent	SD
Science	1275	3.984	32	44.44	1.271
Computer Science	1410	4.147	34	47.23	1.420
Social Science	435	4.35	6	8.33	1.466

8. Discussion

Digital libraries are rapidly gaining attention in digital learning communities, especially higher education. A digital library can provide learners with access to information anytime, anywhere and any place in any format. For this reason, digital libraries can be a great addition to higher learning.

The overall results suggested that students had positive perceptions toward the use of the digital library.

The results of ANOVA for gender showed that female students scored significantly higher than male. This means that female students had higher positive perceptions toward the use of digital library. It is reported that female students appear to dominate the on line learning environment. They are inclined to demonstrate a higher confidence level toward on line learning. It is also noted that male students may have had difficulty with lower confidence levels. The lower confidence level may create the barriers that limit the opportunities and choices for male students in achieving a positive academic learning experience. Therefore, this study recommends that students, especially males be provided with special hands-on training that is uniquely geared toward the learning styles when dealing with the use of digital libraries in their learning. This may help to increase their confidence levels toward the use of the digital library.

The results of ANOVA proved that there is a significant difference between first year and second year students of post graduate education and also there is a significant difference between the sciences and social sciences courses.

This finding points can influence the incorporation of digital library use in higher learning and teaching. The association of class room and digital library as parallel media help the students to connect the learning environment with the research environment. The online access has created the expectation for facility and case of information in all its uses particularly outside class room learning. Therefore this study recommends that the digital library be incorporated as essential part of higher learning assignments. Further more, the institutions of higher learning should support and implement digital libraries to acknowledge in a commensurate way, the inclusion of web based higher learning programmes, the relay on digital library access.

9. Conclusion

It is concluded that future research should focus on a different population sample and refinement of the instrument. Further the current digital libraries are not following a standard model for retrieving information and many of them have problems regarding system usability. Absence of standardisation and usability may influence students' negative perceptions toward the use of digital libraries. Therefore, these variables deserve attention in future research.

10. References

1. Blum, K (1999). Gender differences in asynchronous learning in higher education: learning styles, participation, barriers and communication patterns (http://www.aln.org/publiations/jaln/v3n1_blum.asp)
2. Carey, J and Gregory, V (2002). Students' perceptions of academic motivation, interactive participation and selected pedogogical and structural factors in web-based distance learning. Journal of Education for Library and Information Science, 431, p6-15.
3. Jones, S.(2002). The internet goes to college: how students are living in future with today technology. (http://www.pewinternet.org/reports/toc.asp)
4. Koohang, Alex (2004). Students' perceptions toward the use of the digital library in weekly web based distance learning assignment portion of a hybrid programme. British Journal of Educational Technology, 35, 5 p617-626.

About Authors



Mr. A. Manoharan is working as Lecturer (SG) in Library and Information Science, Bishop Heber College (Autonomous), Trichy, Tamil Nadu. He holds MA (Econ) and MLISc from Madurai Kamaraj University, MPhil from Annamalai University and PGDCSA from Bharathidasan University. He is the life member of ILA and IATLIS. He has contributed 48 papers in books and journals and edited 3 books.



Ms. Anu Vasanthi did B.Sc. Computer Science at Sri Sarada College for Women, Tirunelveli, currently doing MLIS at Bishop Heber College, Trichy, Tamil Nadu She has contributed three papers in books.



Ms. T. Deepa did B.Com. at Holy Cross College, Trichy, currently doing MLIS at Bishop Heber College, Trichy, Tamil Nadu. She has contributed three papers in books.

Is the Big Deal Mode of E-Journal Subscription a Right Approach for Indian Consortia ? A Case Study of Elsevier's ScienceDirect Use at Indian Institute of Technology Roorkee

Yogendra Singh T A V Murthy

Abstract

Big deal or the consortia site licensing is the most preferred way of e-journal subscription for Indian Consortia be it INDEST or the UGC Infonet. In the big deal model all the journals published by a publisher or hosted by an aggregator on its web site are made available to the consortia members at a so called "highly reduced" price. It has been seen that the librarians throughout the world haves been raising objections to this mode since beginning. There are various concerns which have been identified such as monopoly of the publishers, use of a limited number of titles, effect of citation ranking of journals published by the small publishers and the fear of death of journals published by the developing countries. Though a number of articles have been published on this topic but most of them have been on the qualitative aspects of such deals. There are a few studies that have been conducted on quantitative aspects. In this paper a study of use of Elsevier's ScienceDirect at IIT Roorkee has been presented which clearly shows that a very limited number of titles are frequently used in the Institute. This data clearly indicates that the Big Deal mode of subscription is not at all in the favour of the consortia. Supports an alternate model for subscription which should be based on the fixed fee access to the limited set of journals which are frequently used and pay for use for the journals which are less frequently used.

Keywords : Consortia, Consortia Model, Survey, E-Resources

0. Introduction

Consortia site licensing model of subscription to electronic journals is the most common model or the so-called Big Deal being followed by the consortia throughout the world. In this model a publisher or an aggregator enters into the contract with the consortium for allowing access to the whole set of electronic journals being published or being hoisted by the said publisher or the aggregator. Most of the time the publishers offers a very wide range of subjects. A substantial portion of such collection may not be of any use for the consortia members. The publisher or the aggregator offers this access at a (so called) heavy discounted price. However, the librarians have through the world shown their concerns about the usefulness of this big deal to their respective libraries and continuously the voices are being raised to find the alternative models of subscription to scholarly journals. The main argument behind this thought is the fact that a major portion of the journals being offered by the publisher/aggregator is never used. This paper presents an analysis of the usage statistics of Elsevier's ScienceDirect (1) which bundles about 1800 journals together and make available to Consortia members.

1. Indian Institute of Technology Roorkee (IIT, Roorkee)

The IIT Roorkee is one of the oldest engineering educational Institute in the world. It was established in the year 1847 as Roorkee College of Civil Engineering and became the first ever such college in the whole British Commonwealth (2). The Roorkee College was renamed as Thomason College of Civil Engineering in 1854. After Independence of the country in 1947, a wider role was envisaged for the

Thomason College and therefore, in 1949 it was made a first ever Engineering University in the country and University of Roorkee came into existence. Recognizing its national importance the University of Roorkee was converted into Indian Institute of Technology Roorkee in 2001. The Institute imparts education and research in most of the branches of Science and Technology. It has a separate department for Humanities and Social Science also(3) . Further information about the Institute may be obtained at www.iitr.ac.in or www.iitr.ernet.in

1.1 Library and Information Services (LIS) at IIT Roorkee

The library services of the Institute has its origin with the establishment of College Library in January 1848 as a subsidiary department of the College. Later on the collection of Adiscombe College London and also of the Ganges Canal Library was merged into it and it came to be known as Central Library (4). After the conversion of Thomason College into University of Roorkee, departmental libraries were also established which contains the core collection to meet the day to day requirement of the various departments. The information requirement of the departmental library is also met through Central Library. Therefore, the Central Library has the overall responsibility for the development of LIS in the Institute.

To provide effective service through the use of latest available technologies has been the 'Mool Mantra' of the Central Library since beginning. It has started using computers in the early 1990s and by the year 1994 it has started providing CD-ROM search services and e-mail services to its users. The Central Library had computerized most of its functions by 1997 and established its own LAN in 1999 where all the CD-ROMs and OPAC were made available. In the year 2000 the Institute Fibre Optic Network (IFON) was established and the Library LAN was integrated with it. Thus all the electronic resources of the Central Library became available throughout the Institute campus. At present Central Library itself maintains a network of 52 modes with 5 servers. It has its own web portal and all the services are available through Intranet as well as Internet. For more information http://www.iitr.ac.in/resources/library/ may be visited (5).

1.2 E-journal Subscription at IIT Roorkee

The Central Library subscribes to about 7000 electronic journals through following two modes (6).

1.2.1 Through INDEST

IIT Roorkee being level one member of INDEST Consortium all the resources being subscribed to by the INDEST are available. To it. The major full text packages include about 1800 journals through Elsevier's ScienceDirects, more than 500 titles through Springer's SpringerLink, the whole database of ACM's ACM Digital Library, IEEE'S IEL, fulltext collection of ASME and ASCE. The aggregator services include EBSCOhost and ABI/Information. Secondary resources available through INDEST are Compendex, Inspec, MathScinet , ScifinderScholar etc.. There are factual data bases available such as Euromonitor, Capitaline etc. The access to all these services are authenticated through the IP address.

1.2.2 Own Arrangement

Besides the INDEST Consortium the Central Library also subscribes to a number of e-journals. All the journals which give free online with print are available to the users. Further most of the journals where online is available at an extra payment are also subscribed on line along with the print subscription. Some of the major publishers whose journals have been subscribed by the Library are Institute of Physics Publishing (IOP), American Physical Society (APS), Royal Society of Chemistry (RCS), American

Chemical Society (ACS), American Institute of Chemical Engineers (AICHE), Science Online etc. The access to these resources are IP authenticated in most of the cases except a few cases where access is ID and pass word authenticated.

2. Review of Literature

A search on big deal or consortia site licensing in Library and Information Science Abstract retrieved 31 records of which ten were found to be relevant to the topics likewise search on the same topic in ABI/ Inform a full text database in management related information retrieved four relevant articles. A other management database Emerald also specialized management information produced 5 relevant articles. This clearly shows that the debate on big deal is on.

It appears that a detailed discussion on "big deal" took place first time in January 1999 at Mid Winter Meeting of American Library Association (ALA where first of all the issues like a panel of these speakers addressed the topic of Electronic Journal Pricing : What is the Big Deal? The issues which were discussed were economic pricing and current perspective and preferred practices for the selection and purchase of electronic information. (7,8).

The issue of monopoly in the general subscription policy over the UK's National Electronic site Licensing Institution (NESLI) by one of the four served agents i.e. SWETS was raised wayback in 2000 (9). This agreement forced the participating libraries to terminate the arrangement with other subscription agents in order to access NESLI.

Frazier (10) has suggested that "academic library directors should not sign on to the Big Deal or any comprehensive licensing agreement with commercial publishers". He gave the different reasons for that. The push to build an all electronic collection can not be undertaken at the risk of (a) weakening that collection with journals we never need or want and (b) it will increase our dependence on publishers who have already started sharing their determination to monopolize the market". He further suggested the alternative like subscribing access to only those journals which are most needed by us.

Bergstrom and Bergstrom (11) infers that in the process of shift from paper to electronic format societies and not for profit organizations may transfer the savings i.e. publishing to users in the form of reduced pricing but the commercial publishers may not do so. Indeed many commercial publishers have placed their electronic versions at par with print versions. They have stressed that the scientific community will only be benefited by licensing the sites on the basis of pay per view basis. The success of "Big Deal" has serious repercussions for smaller society publishers which make scholarly communication very hostile for these publishers. They will only be able to survive by changing their business models like Open Access (12).

Friend (13) argues that purchasing models so called Big deal is not in favour of small publishers and for large libraries even is in short term may be good for large publishers and small libraries. He further stresses that both the publishers and libraries should find an alternative models for small publishers and large libraries.

Quint (14) discusses the targeting of large institution, particularly academic libraries. In this respect she emphasizes on the librarians support on open access and reports that Association of Research Libraries has made a serious commitment to moving its members to open access scholarly models. She has also said that Big Deal may come to an end very soon.

Helfer (15) has presented the excerpts of statement from Cornell University Libraries (CUL, New York which explains why the CUL has decided to cancel over 200 titles from Read-Elsevier. Pickering (16) has stated that there is a growing revolt among academic libraries unhappy with the Big Deal schemes that force them to take many periodicals that force them to take them many periodicals that are seldom used. He (has reported that in ULC an investigation out the scientific publishing is being conducted by the Members of Parliament as a backlash against the escalating academic periodicals subscription costs. The investigation would focus on publishers pricing policy for scientific journals, particularly the Big Deal schemes and the impact of open access initiatives. He foresees that the out come of the investigation will have major impact on main publishers including Elsevier, Springer, Wiley, Wolt Kluwer etc. and it will encourage open access projects (17).

Ball (18) examines the Big Deal in the light of fundamental market conditions and suggests alternative models for electronic resources. He has defined the role and strength of various players in information supply chain. Special emphasis has been laid on the dangers of such big deals mainly monopolistic position of the publishers. He has also suggested ways to minimizing these dangers – such as consortia, alternative publishing models and new economic models to promote competition.

Dyer(19) has stated that the several of the United State's most prestigious universities are threatening to cancel their subscriptions to scientific journals published by Elsevier, in protest at what they call exorbitant pricing. Unversities are advising their faculty to consider placing their research in "open access" journals. Other universities to pass similar resolutions in recent months include Harvard, Massachusetts Institute of Technology, Duke, Cornell, the University of Connecticut, the University of California, and North Carolina State University. University librarians say that journal price hikes combined with a weak dollar and falling budgets leave them no choice but to cancel subscriptions. Several other US universities threaten to cancel subscriptions to Elsevier journals.

While there are a number of articles on the pros and cons of the big deal, on its quantitative analysis at micro level i.e. up to the use of individual titles seem to be a few. Hamaker (20) did an analysis of the use of Elsevier' ScienceDirect on 864 titles available online to the Universities in North Carolina and found that 28% (102) titles accounted 47% usage. There were 274 titles that were accessed only 5-times or less. Similarly Nicholas and Hurtington (21) have found that in case of Emerald 43% of the subscribers viewed only one and 40% of the subscriber viewed only 2 to 5 titles out of 118 licensed. Thus 83% of the users used only less than 5% of the titles available. They further argues that why to pay for 95% and why not to revert to the basic core collection which is alive in electronic format also.

3. Data Collection and Methodology

The main reason for selecting the Elsevier's ScienceDirect was that it covers about 25% of the total e-journals available to the library being the largest STM (Science Technology and Medicine) publisher of the world. In terms of expenditure also a major portion of the Consortium Budget is spent on Elsevier. Further it has put a condition that no consortia member will drop the print subscription below a level that was being subscribed to during 2002. Thus the Elsevier's ScienceDirect is a major stake holder in the whole process.

The data for this study was taken from the usage reports of Elsevier's ScienceDirect for the year 2003 as this is the year for which whole year's data was available. The usage reports provided by Science Direct are very exhaustive and are available in COUNTER Compliance format (22,23). These reports may be easily manipulated with Microsoft Excel worksheet for further analysis. The data so obtained was down loaded into the Microsoft Excel worksheet and sorted according to the ascending order of the full text requests made to a particular title. Thus titles were arranged according to the ascending order of the usage. Data was also available as per the monthly access. The further analysis was also done using Microsoft Excel.

4. Analysis of Data

4.1 Average Monthly Access

A total number of 95,787 requests for full text access were made from IITR to Elsevier's ScienceDirect during 2003. The minimum of requests were made during February (2949, 3%) and the maximum requests were made during July (12,156, 12.2%). The average monthly access was 2982 requests. The usage was slow in the first few months. The reason for slow usage in the first half the year may be attributed to the fact that the service has started but it pictured up gradually. July being the month of new students and research scholar has shown maximum number of requests.(figure1, Table1).

Table 1. Monthly access of ScienceDirect during 2003 in IIT Roorkee

Month	Requests recd	%age of total requests recd	Month	Requests recd	%age of total requests recd
January	5272	5.50	July	12156	12.69
February	2949	3.08	August	8463	8.84
March	8109	8.46	September	11318	11.82
April	5234	5.50	October	9253	9.66
May	4328	4.52	November	9427	9.84
June	8474	8.85	December	10804	11.28
Total	34366	35.88	Total	61421	64.12

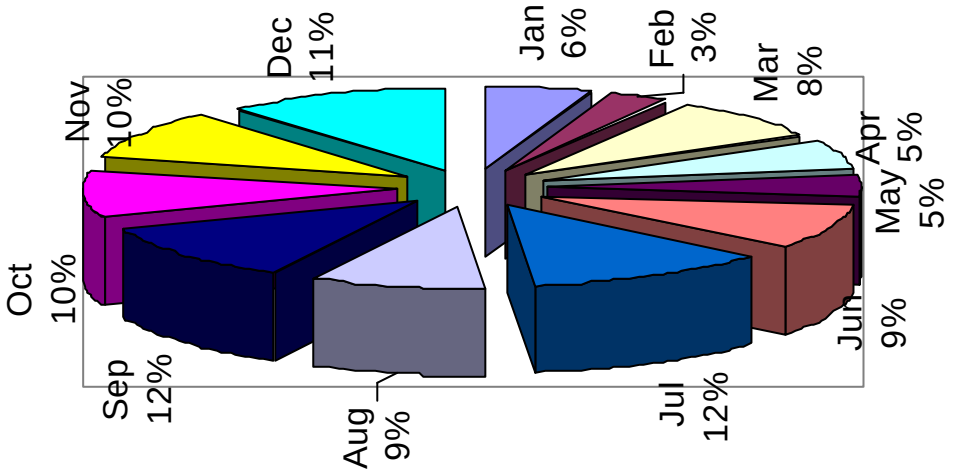


Figure 1. Monthly fulltext requests made to Sciencedirect during 2003

4.2 Analysis of Titles Accessed

The further analysis of titles has shown very interested trend as it was found that the requests were mostly centered around a limited number of titles. The number of requests ranged from zero to 3974. Whole data was divided into four groups. The group I contained the data about the journals that received

access in single digit in whole year i.e. for zero to nine. Group 2 contains the data receiving requests in two digit i.e. for 10 to 99. Likewise group 3 contains the data about titles receiving requests in three digits (100 to 999) and group 4 in four digits i.e. 1000 onwards. Group 1 contains 785 titles responsible for 1.99% access, Group 2 contains 476 titles responsible for 17.35% access, Group 3 contains 213 titles responsible for 60.34% of access and Group 4 contains 12 titles responsible for 20.39% access.(Fig.2)

4.2.1 Titles in Group 1

It was found that there were 251 titles which received no request at all which constitute about 16.89% of the total titles (1486) available online during 2003. There were 145 titles which received single requests each. The number of title started declining and the number of requests per title made started increasing except in case of nine requests made per title where it was slightly higher than its predecessor i.e.number of eight requests made per title. It was found that this group contains 785 titles

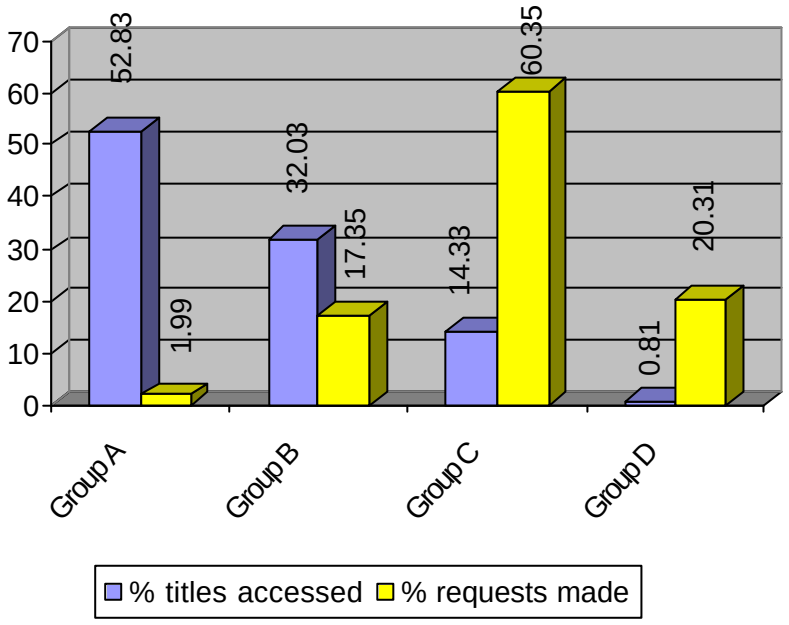


Figure 2. Requests made in different groups

which is about 52.82% (more than half) of the total tittles available but contributed to only 1.99% (less than two percent) of the usage. (Table 2 and Figure.3).

Table 2. Details of requests made and titles accessed in Group 1

No. of Requests Made	Titles accessed	% age of total titles available	Requests made	% age of total requests made
0	251	16.89	0	0.00
1	145	9.76	145	0.15
2	101	6.76	202	0.21
3	64	4.31	192	0.20
4	53	3.57	212	0.22
5	39	2.62	195	0.20
6	43	2.89	258	0.27
7	35	2.36	245	0.26
8	25	1.68	200	0.21
9	29	1.95	261	0.27
Total	785	51.00	1910	1.99

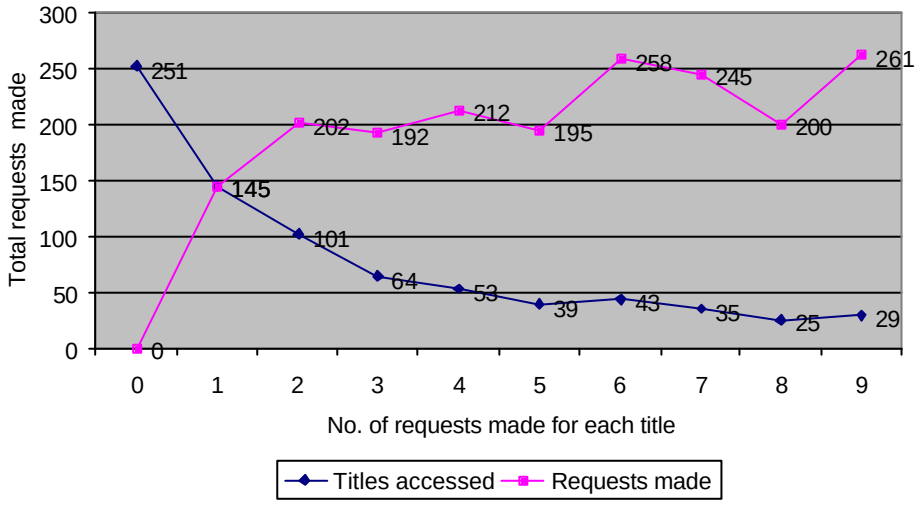


Figure 3. Detail of requests made and titles accessed in Group 1

4.2.2 Titles in Group 2

The same phenomenon i.e. decreasing of number of titles and increasing of number of requests was shown by group 2. There were 175 titles receiving requests between 10-19 and 12 titles receiving requests between 90-99. The total 476 titles of this group which is about 32.03% of the total titles available received 16622 requests which is about 17.35% of the total requests made. The total 82% of the titles in the group 1 and 2 received about 19.34% requests. The famous 80-20 rule looks to be prevailing here (Table 3 and Figure 4).

Table 3. Details of request made and titles accessed in Group 2

Range of Requests Made	Title Accessed	%age of tota title available	Requests made	% of total Requests Made
10-19	173	11.64	2477	2.86
20-29	79	5.32	1895	1.98
30-39	58	3.90	2008	2.10
40-49	45	3.02	2002	2.09
50-59	39	2.62	2113	2.21
60-69	34	2.29	2165	2.26
70-79	23	1.55	1724	1.80
80-89	13	0.87	1103	1.15
90-99	12	0.81	1135	1.18
Total	476	32.03	16622	17.35

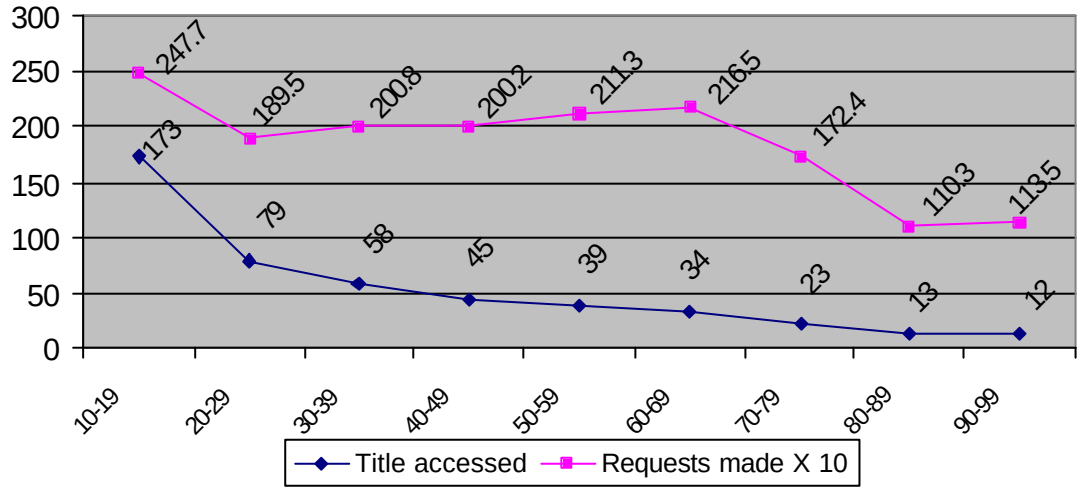


Figure 4. Detail of request made and titles accessed in Group 2

4.2.3 Title in Group 3

Group 3 has also shown the same pattern as shown by Group 1 and 2. There were 99 titles in this group that received requests between 100 and 199. The number of titles started decreasing as the number of requests started increasing except a small variation as the requests made to 7 journals were in the range of 700-799 while only 5 titles received requests in the range of 600-699. The 213 titles (14.33% of titles) received 57800 (60.34% of the total) requests. (Table 4 & Figure 5). This was the group receiving maximum number of requests.

4.2.4 Titles in Group 4

Titles in Group 4 received 1000 or more requests. It was found that there are only 12 titles in this group that received 19455 requests. Thus 0.82% of the titles in this group received 20.39% of the total requests. (Table 5 and Figure.6). Individual requests received by the top 12 journals are shown in figure 7.

Table 4. Details of request made and titles accessed in Group 3

No. of Requests Made	Title Accessed	%age of total titles available	Requests made	% of total requests made
100-199	99	6.65	14098	14.72
200-299	48	3.23	11651	12.16
300-399	28	1.87	9510	9.93
400-499	14	1.44	6306	6.58
500-599	8	0.54	4405	4.60
600-699	5	0.37	3211	3.35
700-799	7	0.46	5141	5.37
800-899	3	0.20	2572	2.69
900-999	1	0.07	906	0.95
Total	213	14.33	57800	60.34

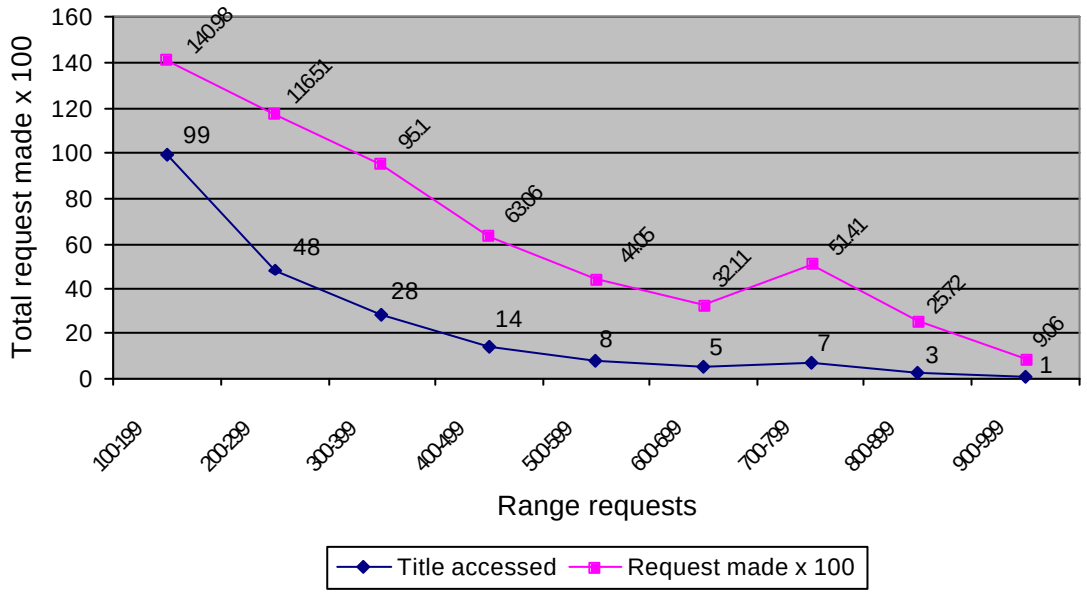


Figure 5. Details of request made and titles accessed in Group 3

Table 5. Details of requests made and titles accessed in Group 4

No. of Requests Made	Title Accessed	%age of total titles available	Requests made	% of total requests made
1000-1999	10	0.67	13123	13.79
2000-2999	1	0.07	2358	2.46
3000-3999	1	0.07	3974	4.19
Total	12	0.81	19455	20.31

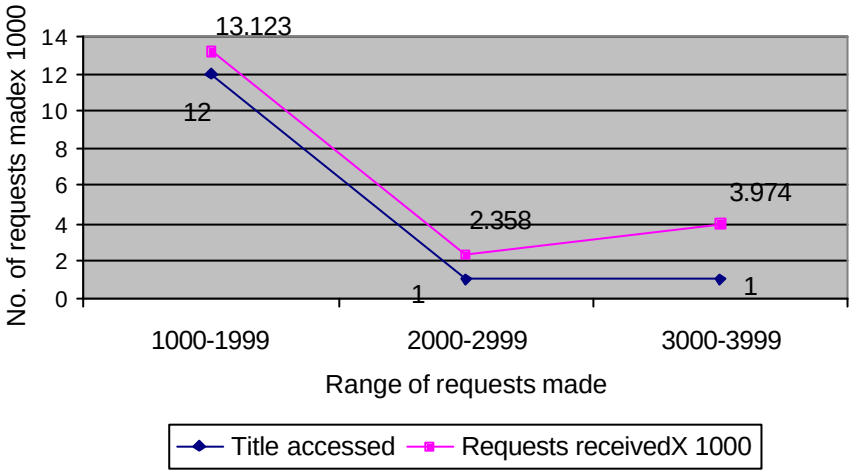


Figure 6. Details of requests made and title accessed in group 4

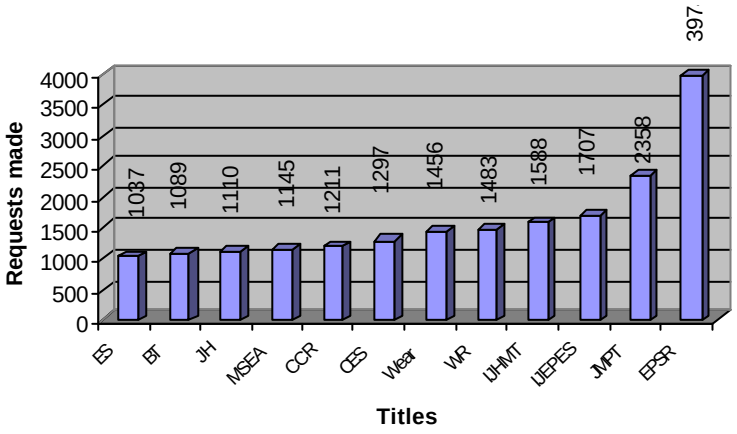


Figure 7. Top 12 Journals receiving more than 1000 requests during 2003

ES= Engineering structures	Wear=Wear
BT= Bioresource Technology	WR = Water Resources
JH=Journal of Hydrology	IJHMT= Int. Jl of Heat & Mass Transfer
MSEA= Material Science Engineering A	IJEPES = Int. Jl. of Electrical Power & Energy Systems
CCR=Coordination Chemistry Reviews	JMPT= Jl. Of Material Processing Technology
CES=Chemical Engineering Science	EPSR= Electric Power System Research

4.2.5 Distribution of titles in different quarters .

For the further analysis, four quarters containing the number of titles receiving 25% of the total requests were made. It was found that top 25% of the requests were made to only 17 titles i.e. 1.14% of the total titles available. Next 25% of the requests were received by 49 titles (3.3% of the total). Third 25% requests were made to 111 titles i.e. 7.47% and the last 25% of the requests were made to 1058 titles (71.20% of the total. 251 titles (16.89%) of the total titles available received no request at all, (Figure 7, Table 7).

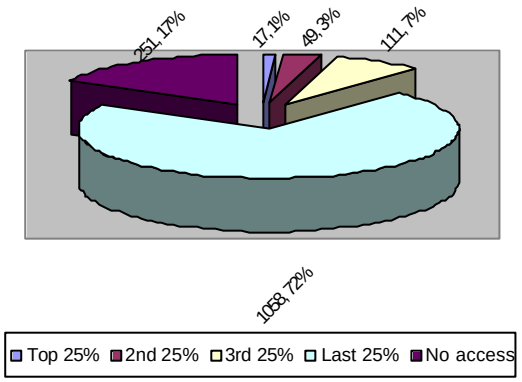


Figure 8. Titles accessed in different quarters

5. Discussion

The analysis of data in this study clearly shows that a very small fraction of the titles available are being heavily used and there is a very large portion which either not being used at all or being used rarely. The concentration of the requests around a limited number of titles clearly shows that the core collection is very much alive and active. If IITR subscribe to only 17(1.14% of the total available) titles than its 25% requirement can be met. Subscription to 66 (4.4% of the total available titles) alone can meet 50% requirement and subscription o 177 titles (only about 12% of the total) can meet its 75% requirement. Thus 88% of the titles are being subscribed to meet 25% of the total requirement. It is thus clear that by this arrangement consortia members are the losers and the publisher is winner as he gets the payment for the information which is never used. It clearly shows that there is an urgent need to look into the subscription model of big deals. Besides more revenue generation for the publisher the big deal model has certain other disadvantages for the information domain as a whole as panted out by Ball (18). On the top is the issue considering archival rights and licensing as the information being made available is licensed and not sold to the Consortia members. Another issue which is worth considering is that the big deal increases the citation ranking of the big publishers as after spending on big deal, hardly any budget will be left for the small publishers which will eventually lead to their death. Librarian will have no role in decision making about the subscription and last but not the least it will definitely lead to the monopoly of big publishers.

6. Conclusion

The highly core centered access to ScienceDirect in IIT Roorkee clearly points out that the renegotiation with the publishers is necessary as no mutually agreed contract can be successful if it is not equally balanced. At present the Big Deal arrangement seems to be in favor of publishers. Librarian have already started opposing it and with valid apprehensions. Some may argue that there is always 80-20 rule prevalent in the libraries but it may have been valid when the libraries have to keep stock in print. In case of electronic resources we have constant access to the information, which can always be accessed on payment. There is no need to pay in advance for information which is never used. A viable model will be that the both libraries and publishers are benefited equally. This can only be brought by increasing the use of the information and subscribe to only a core set. Rest of the information can be accessed in pay for use method.

Though the present study is based only on the data for one year that too for one institution which may not be representative of all members, but the data gives a sufficient insight into the state of affairs. It at least is sufficient enough to initiate further studies on this issue. It also gives the sufficient grounds to negotiate with the publishers.

Elsevier's ScienceDirect covers very broad spectrum of STM and therefore, may not be a true representative of all the publishers which are specialized such as AIP, IOP, ACS etc. Therefore individual study are necessary for individual publishers/aggregators.

7. References

1. Science Direct Usage Reports available at http://usagereports.elsevier.com accessed on 25th October, 2004.
2. Mittal, K.V.(1996). History of Thomason College of Civil Engineering. Roorkee, University of Roorkee.
3. IIT Roorkee at a Glance (2004). Roorkee, Indian Institute of Technology Roorkee, 2004. Also available http://www.iitr.ac.in/utilities/iitr_at_a_glance.pdf
4. Saxena, R.S.(1982). A history of Central Library of University of Roorkee. Roorkee, Unversity of Roorkee..
5. Singh, Yogendra.(2004). A profile of Central Library, Indian Institute of Technology Roorkee. In Souvenir International Workshop on Webometrics, Informetrics and Scientometrics, held at Central Library, Indian Institute of Technology Roorkee, India 2-5 March 2004.pp13-19.
6. Indian Institute of Technology Roorkee(2004). Annual Report 2003-2004. Roorkee, Indian Institute of Technology Roorkee, India.
7. Davis, S.(1999). Journal costs in libraries discussion group, ALA Midwinter Meeting, 1999. Serials Review; 25 (3):103-4.
8. Roth, A C.(2000). Electronic journal pricing: what's the big deal? A report of the ALCTS Serials Section discussion group meeting. ALA Midwinter Meeting, 1999. Technical Services Quarterly; 17 (3):2000, p.67-73.
9. Ball, D., Wright, S.(2000). The information value chain: emerging models for procuring electronic publications. Online information 2000: 24th International Online Information Meeting: Proceedings.Learned Information Europe, Oxford, pp. 213-223. available at: www.lib.umich.edu/libhome/peak/.
10. Frazier, Kenneth.(2000). The librarians' dilemma: comtemplating the costs of big deal. D-lib Magazine. 7(3).

11. Bergstrom, Carl T. and Bergstrom, Theodore C.(2004). The cost and benefit of library site licenses to academic journals. Proceedings of the National Academy of Sciences of USA. 101(3): 897-902.
12. Prosser, David C (2004). Between a rock and a hard place: the big squeeze for small publishers. Learned Publishing; 17 (1): 17-22.
13. Friend, F J (2003). Big Deal: good deal? Or is there a better deal? Learned Publishing; 16 (2) :153-5.
14. Quint, Barbara (2004). The end of the 'big deal' era. Information Today; 21(1):7.
15. Helfer, Doris (2004). Leading libraries. Is the big deal dead? Status of the crisis in scholarly publishing. Searcher; 12 (3):.27-32.
16. Pickering, Bob.(2004). Consortium signs up Elsevier. Information World Review.199:2.
17. Pickering, Bob(2004). MPs launch journal pricing inquiry. Information World Review; 198:1.
18. Ball, David (2004). What's the "big deal", and why is it a bad deal for universities? Interlending and Document Supply.; 32 (2):117-125.
19. Dyer , Owen (2004) US universities threaten to cancel subscriptions to Elsevier journals. British Medical Journal. 328:543.
20. Hamaker, C.(2003). "Quantity, quality and the role of consortia", paper presented at What's the Big Deal? Journal Purchasing - Bulk Buying or Cherry Picking? Strategic Issues for Librarians, Publishers, Agents and Intermediaries, ASA 2003 conference, available at: http://www.subscription-agents.org/conference/200302/ chuck.hamaker.pps
21. Nicholas, D., Huntington, P.(2002) "Big deals: results and analysis from a pilot analysis of web log data: report for the Ingenta Institute", in The Consortium Site Licence: is it a Sustainable Model?, In Proceedings of the Meeting held on 24 September 2002 at the Royal Society, London, Ingenta, Oxford, pp. 121-159, pp. 149, 151
22. Shepherd, P.T. (2003).COUNTER: from conception to compliance. Learned Publishing, 16(3):201-205.
23. COUNTER: Counting Online Usage of NeTworked Electronic Resources. Available at http:// www.projectcounter.org

About Authors



Dr. Yogendra Singh is Librarian at IIT, Roorkee. He holds PhD in LIS; MSc(Zoology), MLISc. Prior to joining IIT he has worked in DESIDOC. He is a Fulbright Fellow with University of Maryland, USA from July 1999 to February 2000. He has published and presented several papers in journals and conferences. He is also a member of many professional bodies. His areas of interest are Library management, Library automation, Library Networking, Digital libraries, Internet searching.

Email : yogi@iitr.ernet.in



Dr. T.A.V. Murthy is currently the Director of INFLIBNET and President of SIS. He holds BSc, MLISc, MSLS (USA) and PhD. He carries with him a rich experience and expertise of having worked in managerial level at a number of libraries in many prestigious institutions in India including National Library, IGNCA, IARI, University of Hyderabad, ASC, CIEFL etc. and Catholic University and Case Western Reserve University in USA. His highly noticeable contributions include KALANIDHI at IGNCA, Digital Laboratory at CIEFL etc. He has been associated with number of universities in the country and has guided number of PhDs and actively associated with the national and international professional associations, expert committees and has published good number of research papers. He visited several countries and organized several national and international conferences and programmes.

Email : tav@inflibnet.ac.in

Familiarity and Use by the Students' of Digital Resources Available in the Academic Libraries of Medical Science University of Isfahan(MUI), Iran

Asefeh Asemi

Abstract

An attempt has been made to determine the present status of familiarity and use of Digital resources. It was felt that use of digital resources is still poor among the medical students of the Universities in the developing countries. This paper presents survey to investigate the familiarity and use of Digital resources by students through online and offline Information Databases of the Central Library, "Central Library Books & Journals Database (CLBJD), and the CD-ROMs databases available in the academic libraries (MUI). The subjects of this study were the students of the Isfahan Medical University. For evaluating study questions and data collection, the questionnaire was distributed to a random sample of 250 students. The result of this survey are presented and discussed in the paper.

Keywords : Digital Resources, Electronic Resources, Academic Libraries, Isfahan Medical Science University (MUI), Iran

0. Introduction

Today, we live in exciting times. Digital resources, whose history spans a mere dozen years, will surly figure amongst the most important and influential institutions of this new century.

The information revolution not only supplies the technological horsepower that drives Digital resources, but fuels an unprecedented demand for storing, organizing, and accessing information. If information is the currency of the knowledge economy, Digital resources will be banks where they are invested (Hewitson, 2002).

There are more reasons today than ever before, which have necessitated students to use the Digital resources through the Central library Information Databases and the databases available in the academic libraries in MUI. Digital resources provide access to much richer content in a more structured manner and allow us to search for any word or phrase in the entire collection. The need is to provide online easy, ceaseless access, with multiple user access facility to electronic collections from researcher's desktop or from remote Internet computers. This saves the time of researchers in terms of access to online resources of his choice. Access to online E-journals is possible much before the library receives the journals in paper form. Electronic Document helps minimizing processing time for providing access to the users (Deb, Kar, and Kumar, 2003).

These are some of the reasons for the trust in the use of Digital resources by students. In order to exploit the current information explosion, familiarity and use of the Digital resources in the libraries for rapid development is necessary and important. Digital resources can be used for efficient retrieval and meeting information needs. This is very important for university libraries since most of them call for more and more research work. This important fact is convincing many libraries in Iran that computerization is no longer a thing of the past (Davarpanah, 2001).

2. Objectives

This study was designed and carried out with the view to achieve the following objectives:

- To assess the amount of familiarity and use of the MUI students of the Digital resources through the central library online and offline information databases.
- To assess amount of familiarity and use of the MUI students to the "Central Library Books & Journals Database" accessible through Central Library LAN Network in the academic libraries in MUI and also accessible through central library homepage on the MUI website.
- To assess amount of familiarity and use of the MUI students of the Digital resources on CD-ROMs available in the academic libraries.
- To determine the percentage of students, who have had educational program about use of digital resources in MUI and also, indicate retrieval of students' information needs via these resources.

3. Methodology

The study was conducted by survey method. The design of the study called for the MUI students as the subject. At the time of survey, based on the directory of MUI (MUI, 2003), there are 7 faculties and 7000 undergraduate and postgraduate students in this university.

In the Medical University of Isfahan there is a central library. In this study the use of three central library information and the CD-ROM databases available in the academic libraries by students was surveyed based on the following considerations:

In each faculty there is a library and these libraries are active in using various types of Digital resources on the CENT-LIB databases, databases on the CD-ROMs, floppy discs and etc.

These are three kinds' Digital resources on the CENT-LIB network in MUI. These Digital resources include:

- Online databases accessible through MUI website.
- Offline databases available through CENT-LIB LAN Network. These databases are accessible in the university campus and in the faculty libraries and Alzahra hospital library, only.
- "Central Library Journals and Books Database" available on the Central Library homepage and CENT-LIB LAN Network.

Questionnaire technique was used for collecting data from users. The questionnaire was distributed to a random sample of 250 students for measuring their familiarity and use of Digital resources.

Survey of literature, personal visits, interviews, field observation, and opinions of experts of library professionals were considered. Documents and records available in the libraries related to the subject were referred. Thus measurement & quantification, questionnaire, and interviews were used as the research instruments in this study.

The data thus collected have been analyzed, classified and tabulated for representation in the paper. The graphical soft wares and statistical methods employed in the study were basically the Excel, SPSS.

4. Results and discussion

Online databases available on the MUI website are:

Springer Journals, Oxford Journals, Ovid Journals, Ingenta, Proquest, Blackwell, Elsevier Science, EBM Review, Ovid Medline, ERL 5 Medline, MDC Consult, Up-to-date Software, Images- MD.

Offline databases available on the CENT-LIB LAN Network. These databases are accessible in the campus university and in the faculty libraries and Alzahra hospital library. Include:

Medline (Silver platter), LCB, ISA, ERIC, IPA, MELI, Serfile, CINAHL, Ulrich's, BIP.

"Central Library Journals and Books Database (CLJBD)" are available on the MUI website and CENT-LIB LAN Network.

4.1 Student's familiarity with Digital resources

The data analysis pertains to 250 students, 7 academic libraries and Central Library.

"Figure 1" Shows that maximum number of respondents were from Medicine College (35%) whereas minimum (2%) number of respondents were from Rehabilitation College. Of course distribution of questionnaires was based on random sample.

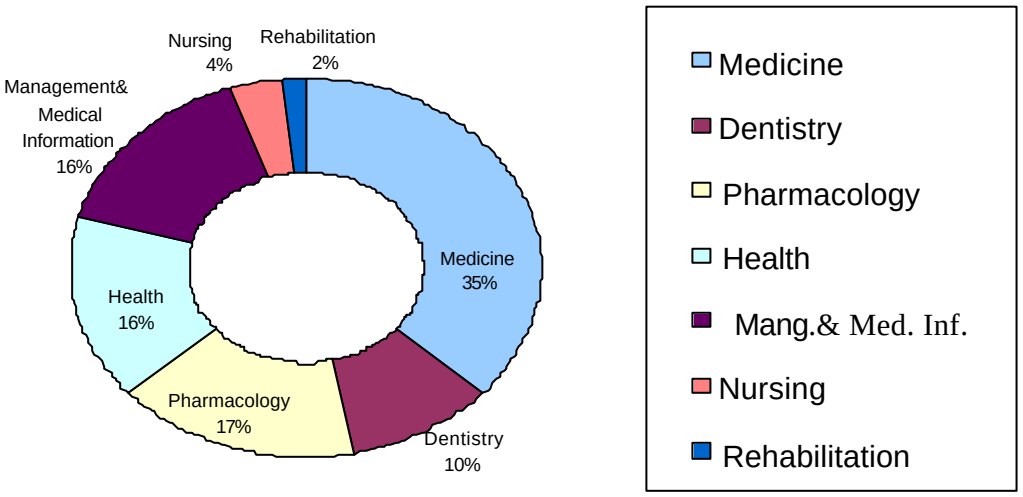


Figure 1. Percentage of despondences based on Colleges

"Figure 2" Shows maximum number of respondents were Master and MD students (53.4%) and minimum (6.9%) number of respondents were PhD and Associate students.

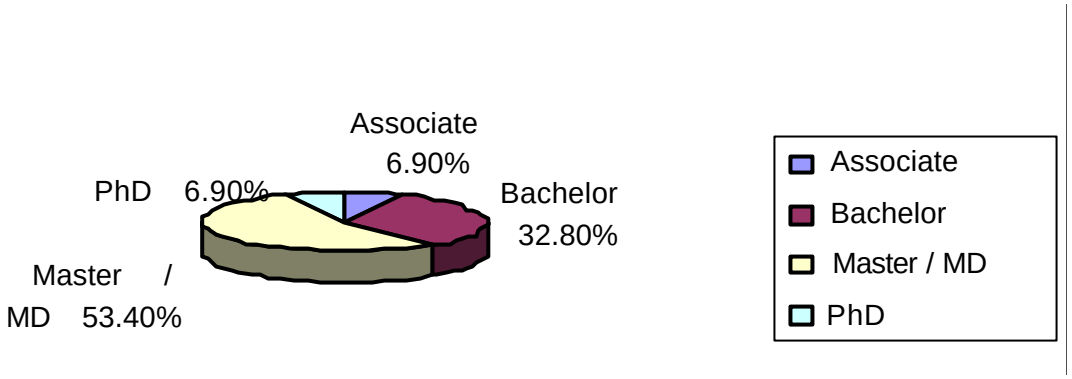


Figure 2. Percentage of respondents based on graduation

"Figure 3" shows that 31% of respondents were familiar to digital resources in academic libraries via personal communication. 19% of respondents via directories, 17% of them via library and 14% of them via other means. 47(19%) of respondents did not answer this question.

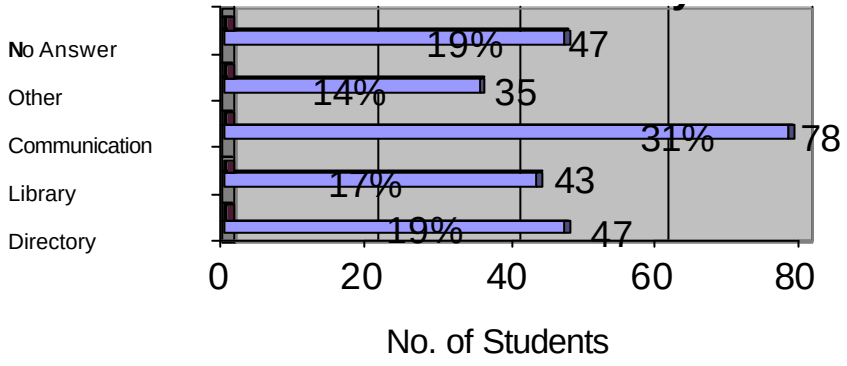


Figure 3. Quality of students familiarity with digital resources available in academic library

"Figure 4" shows that 175 (70%) of respondents are familiar with digital resources. 51 of them (20.4%) replied in negative saying that they are not familiar with digital resources. 24 students (9.6%) did not answer this question.

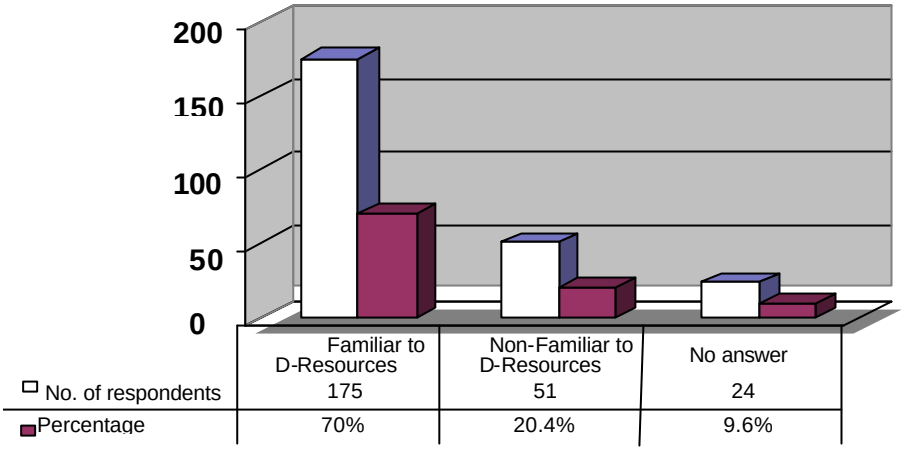


Figure 4. The familiarity of students with digital resources

"Table I" shows that 31.2% of the students are familiar with offline databases through CentLib LAN network in MUI. Familiarity of 22.8 % of them to these databases is less. According to the data of this table 16.8 % of students have less familiarity to online databases accessible through CentLib homepage and 27.2% of them are more familiar to these databases. 58 of the respondents (23.2%) are more familiar with "Central Library Books & Journals Database" through library homepage and 8.4% of students have less familiarity to this database. Also this table shows 30.4% of the students have less familiarity to CD-ROMs databases available in the academic libraries and 7.2% of them are more familiar to these databases. In total, 62% students have familiar to offline resources, 69.6% students to online resources, 50.8% students to CLBJ Database, and 43.6% of them have familiar to CD-ROM databases available in the academic libraries in MUI.

Table I. Amount of familiarity of the students to the kinds of Digital resources available in MUI

Resource kind	**Less**	**Middle**	**Many**	**No answer**
Offline Databases	57 People =22.8%	20 People=8% =8%	78 People =31.2%	95 People =38%
Online Databases	42 People =16.8%	64 People =25.6%	68 People =27.2%	76 People =30.4%
CLBJ Database	21 People =8.4%	78 People =31.28%	58 People =23.2%	93 People =37.2%
Resources on the CD-ROMs (available in the academic library)	76 People =30.4%	15 People =6%	18 People =7.2%	141 People =56.4%

Use by the students of Digital resources in academic libraries (MUI)

"Table II" shows that use of academic libraries by more than 32% of the students (82) is Fair. About 14% of the students (35) have less used these libraries quite less.

Table II. Amount use of the academic library and use of the d
igital resources by students in MUI

Amount of use / Case	**Less**	**More**	**Middle**	**Many**	**So many**	**No Answer**
Use of Academic Library	35 People =13.8%	17 People =6.9%	82 People =32.8%	77 People =31%	39 People =15.5%	-----
Use of Digital Resources	30 People =12%	35 People =14%	17 People =6.8%	80 People =32%	12 People =4.8%	76 People =30.4%

Also, from the data of table 2 it can be observed that 32% of the respondents (80) have frequently used the digital resources in the academic libraries in MUI. 12 people of them (4.8%) have also used these resources. 76 respondents (30.4%) did not answer this question. In other word 174 students (69.6%) have used digital resources.

"How do you use the Digital resources, if you need them?" This is the question that we asked students. "Figure 5" shows that 48% of students, themselves use digital resources in the academic libraries (MUI). 32% of them have used these resources with librarian help. Also, this figure shows that 20% of students have used digital resources by other ways.

"Figure 6" indicates that 17% of the students have frequently used "Central Library Books & Journals Database". 16% of the students have used this database less and 7% have used this database in mid range. This figure shows 4% of respondents have used Central Library Books & Journals Database through central library homepage to maximum extent and also 10% of the respondents have used this database to a large extent. 46% of the students did not answer this question.

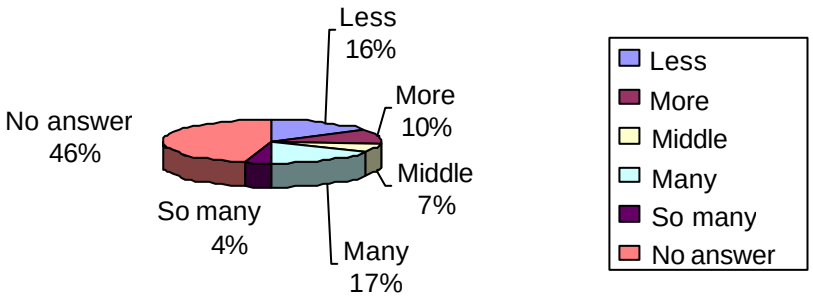


Figure 6. Percentage use of Central Library Books & Journals Database by the student

"Table III" shows usage of offline databases through Central Library LAN Network in the campus of MUI. The students have least used MELI database. Only 3% students have used MELI database. This database includes bibliographic information of the collection of the Iran National Library.

Table III. Amount of Use by the students of offline databases on the Central Library Network (MUI)

Database Name ⇩	No. of Users	Percentage
Medline	187	74.8%
LCB	24	9.6%
ISA	35	14%
IPA	40	16%
ERIC	75	30%
MELI	7	2.8%
Serfile	13	5.2%
CAS	50	20%
BIP	14	5.6%
CINAHL	50	20%
Ulrich's	20	8%

It appears that almost 75% students' use Medline database through offline network of the central library.

After Medline database, about 30% students have used ERIC database. In fact, MEDLINE, ERIC, CAS, and CINAHL are databases which student use of them more than other offline databases.

As "Table IV" shows that online databases through central library homepage have been used more than offline databases. The students have least used Image-MD database. Almost 20% students have used Image-MD database. It appears that almost 85.5% students' use Ovid Journals database through online network of the central library. After Ovid Journals database, about 82% students have used Ovid Medline database. In fact, all of the online database which central library is subscribing, are used by students of MUI. In addition to Ovid Journals and Ovid Medline databases, Proquest, Elsevier Science, Springer Journals, Ingenta, and Blackwell databases have been used more than ERL5Medline, Oxford Journals, MDC-Consult, Up-to-Date, and Image MD online collection.

Table IV. Amount of Use online databases on the Central Library Network (MUI) by the students

Database Name	**No. of Users**	**Percentage**
Springer Journals	151	60.4%
Oxford Journals	87	34.8%
Ovid Journals	213	85.5%
Ingenta	153	61.2%
Proquest	202	80.8%
Blackwell	150	60%
Elsevier Science	176	70.4%
EBM Review	73	29.2%
Ovid Medline	204	81.6%
ERL5Medline	89	35.6%
Up-to-Date	63	25.2%
Image-MD	50	20%
MDC Consult	86	34.4%

"Did you have formal educational program about use of Digital resources in the library?" This was the question, which was asked to students. "Figure 7" shows, that 9% of respondents replied in affirmative while 74% said that they did not have any educational course about use of digital resources. 17% of respondents did not answer this question.

"Please indicate retrieval of your information needs via digital databases in the academic libraries in MUI?" This is the question that we asked students.

"Figure 8" indicates that 106 (42.4%) of the respondents thought that available digital resources met information needs of them at average level. The satisfaction level of information needs of 32 students (12.8%) was poor. 47 students (12.8%) replied in affirmative saying that they were satisfied with retrieval of them information needs of digital resources.

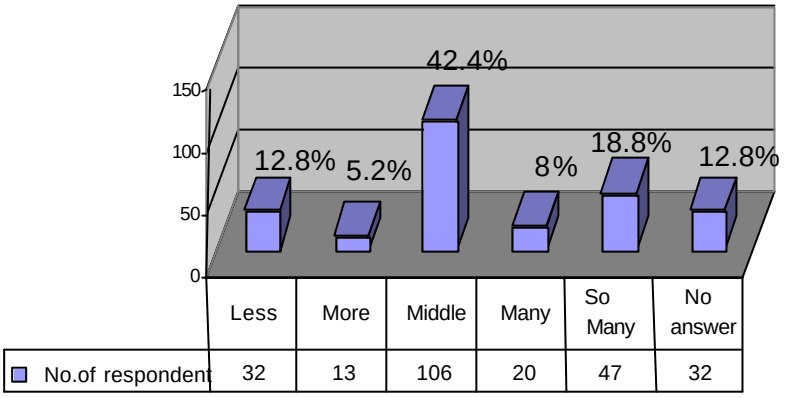


Figure 8. Amount of retrieval of users information needs via digital resources available in the academic libraries

32 students (12.8%) did not answer this question. On the average 218(87.2%) think that available digital resources meet their information needs.

Of the 250 students interviewed, 175 (70%) were aware of digital resources

(62%) were familiar with offline databases available through Central Library LAN in the academic libraries (MUI), 174 (69.6%) were familiar with Online databases accessible on the central library homepage, 160 (64%) were familiar with "CLBJD" accessible through Central Library LAN in the academic libraries in MUI and also accessible through central library homepage on the MUI website, 109 (43.6%) were familiar with CD-ROM databases available in the academic libraries.

All of the students (100%) had used academic libraries and only 173 (69 %) had used digital resources. More than half of the students i.e., 135 (54%) had used "CLBJ Database" accessible through Central Library LAN or Central Library homepage. Also on the average 18.72% had used all of the offline databases through Central Library LAN in the campus of Medical University of Isfahan (MUI). On the average 52.24% had used all of online database accessible through Central Library homepage on the MUI URL (http:// www.mui.ac.ir).

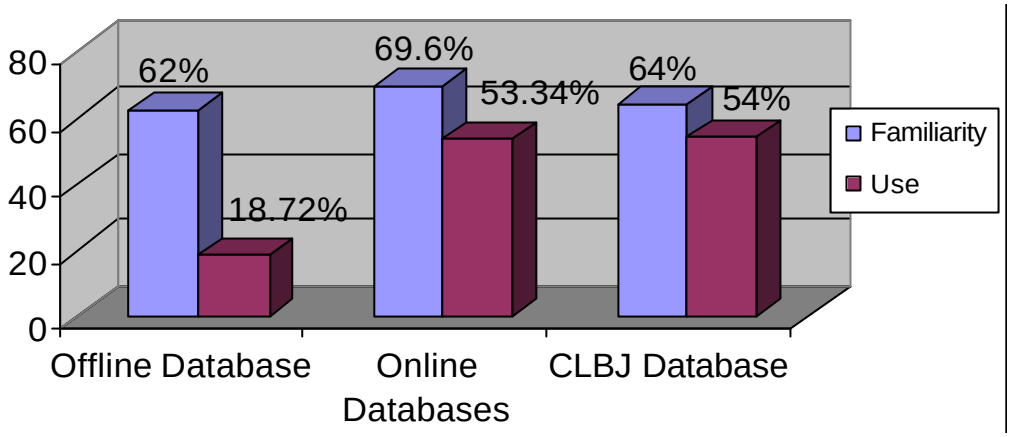


Figure 9. Compare between familiarity and use by students of digital resource in MUI

4.2 Discussion

Though the above factors, it was concluded that a relationship between the scales of familiarity and use of digital resources exists. The level of use by students as compared to their familiarity is good but the students have used poor of offline databases. The reasons for the little use of offline databases through Central Library LAN are multifactored. Poor use of these databases could be attributed to these critical underlying important factors: infrequent periodic orientation and education to use of offline databases by students, infrequent power of these databases on meet information needs of students. In other words these databases are poor in the retrieval information needs of the students, and infrequent of networked personal computers in the academic libraries, which should be connected to a central server in the Central Library.

It was therefore suggested that the offline databases on the Central Library LAN should be supported. The scale of students' familiarity with digital resources should be increased and that the staff of academic library should periodically train the preclinical and clinical medical students in searching available databases. These steps would enable the medical students to benefit from digital resources searching when it becomes fully operational in the future.

5. Conclusion

The difference between the scale of familiarity of the students with Digital resources and use of them available in the academic libraries (see "Figure 9"), and the relationship between the scale of students' use of Digital resources and scale of the retrieval of users information needs through D- Resources through the Central library Information Databases and the databases available in the academic libraries in MUI, have been discussed. The paper concludes that the use of digital resources by students are needed to solve the user problems. Also these resources need to carry up the level of user familiarity and literacy for use of them. Also, the paper concludes that the use of these resources in the academic libraries in MUI by students needs to have formal programme of educating course for users with special references to students (Ogunyade and Oyibo, 2003).

6. References

1. Chen, H. (2001). Medical Text Mining: ADLI-2 Status Report. India: Banglore.

2 Davarpanah, Mohammad Reza.(2001). Level of information technology application in Iranian university libraries. Library Review, 50, 9, 444-50.

3. Deb, Subrata, Kar, D. & Kumar, S. (2003). Electronic Library: A Case Study With Reference to TERI. India: Banglore.

4. Hersh, W. R., Detmer, W. M. & Frisse, M. E. (2000). Information Retrieval systems. Edward H. Short & Leslie E. Perreault. Medical Informatics: Computer Applications in Health Care and Biomedicine. Newyork: Springer.

5. Hewitson, Andrew. (2002). Use and awareness of electronic information services by academic staff at Leeds Metropolitan University -a qualitative study. Journal of Librarianship and Information Science, 34, 1, 43-52.

6. Hunt, R. H. & R. G. Newman. (1997). Medical Knowledge Overload: A disturbing trend for physicians. Health Care Management Review, 22, 1, 70-75.

7. Ogunyade, Taiwo O. & Wellington A. Oyibo. (2003). Use of CD-ROM MEDLINE by Medical Students of the College of Medicine, University of Lagos, Nigeria. Journal of Medical Internet Research, 5,1.

8. Peterson, Michaael W & et al. (2004). Medical Students' Use of Information Resources: Is the Digital Age Dawning? . Academic Medicine, 79, 1, 89-95.

9. Sridhar, C. B.(2001). Internet: Apower Full Tool Disseminating Medical Knowledge in Urban and Rural India.. India: Banglore.

About Author

Asefeh Asemi is a Lecturer at the Department of Lib.& Inf. Sci., Isfahan University, Isfahan, Iran and a PhD student of Library & Information Science, Department of Library & Information Science, Pune University, Pune, India ,
Email : asemi@dnt.mui.ac.ir

UGC-Infonet E-Journals Consortium an Indian Model Bridging the Gap between Scholarly Information and End User

T A V Murthy
V S Cholin
Suresh K Chauhan
Raghavendra Patil

Abstract

Any educational system must have to depend on authentic, factual, fast and up to date information. Indian educational system is one of the largest in all over the world but due to financial limitations large number of libraries have not been able to subscribe to quite a good number of journals required for research and academic community, University libraries could play a major role to further improve the status of higher education system of India. After analyzing the situation the University Grants Commission initiated two important projects viz. UGC-Infonet providing connectivity to universities and UGC-Infonet, E-Journals Consortium to provide scholarly access to electronic journals and databases. Probably this is the golden era in the history of higher education system in India. The total program is funded by UGC and ERNET (Education and Research Network) has been entrusted to establish infrastructure within member universities on a turn key basis and the overall monitoring and execution of the project is being done by INFLIBNET. Through this program large number of e-resources subscribed and provided access to faculty and research scholars working in universities. To make people aware about the use of e-resources good number of user awareness training programs and also conducted 5 national seminars at five different places. Usage statistics provided by different publishers are also very interesting and encouraging.

Keywords : Consortia, E-Resources, UGC-Infonet

0. Introduction

The history of higher education in India has been started way back to ancient and colonial time which used to restrict up to oral and written communication only. But modern higher educational system in India has been started 147 years back when foundation of three major universities in Calcutta, Madras, and Bombay was led by the British. Today, we have more then 310 universities and 14000 colleges affiliated to these universities and about ten million students are studying in these institutions. All this makes Indian educational system as one of the strongest and largest educational systems in the world. No doubt that it is very complex and complicated to manage this whole educational system in systematic and qualitative manner.

Library is the essential and most important part to drive this whole educational system in systematic way. Students, research scholars, faculty members all have to deal with information only and this is the responsibilities of the library to collect, manage, and disseminate the information to its user according to their need. But in true sense university library in India started making progress after 1924 when *Dr. S R Ranganathan,* the father of library science in India, chaired the *Madras University Library*. He did several efforts to improve the status of university and college libraries. Establishment of University Grants Commission (UGC), 1953 as the apex national organization concerned with the setting up and maintenance of standards in higher education throughout the country was a mile stone in the history of higher education in India. Right from very beginning UGC has been giving importance to the development of university and colleges libraries and taking significant steps to improve the higher education.

We have such a big educational system but not able to generate and acquire scholarly information with in the time. More than 11,000 students are being awarded Ph. D every year. After serious thinking of experts and analysts, one reason has found that our university libraries are unable to fulfill their obligations to the objectives of higher education. The biggest reasons for this are inadequate funds and inadequate information resources. There are many other factors which are working as constraints to free flow of scholarly literature in the university libraries like explosion of information resources, cost of books and journals, which is going higher and higher every year and other side due to shrinking budget library has to cut the subscription of important journals, users demands are also changing day by day, duplication of resources in university libraries, changing nature of collection, etc., and its not possible for an individual library to keep pace with these changes and diversified demands of users single-handedly. The constraints could not stop here, emergence of digital or online information makes the job of librarian and information scientist tougher and more complicated. Our higher educational system has weakened day by day in terms of collection, management, generation, and dissemination of scholarly literature. Some of the reasons are:

- Subscription of small number of scholarly journals from overseas by universities.
- Repetition of subscribed titles in different universities.
- Research conducted by Indian universities is not being considered up to the international level.
- Most of the scholarly literature is coming online and back volumes of that literature are also being made available online by publishers.
- Digital Divide in Indian universities.
- It is difficult to the publisher to deal with many subscribers and in the same way it is difficult for a university to manage the different publishers at a time.
- In broader sense we can say increasing cost of information, shrinking financial sources and information explosion leads us to many challenges.

Keeping all these things in view University Grants Commission (UGC) has launched two ambitious programs namely “UGC-INFONET” and “UGC-Infonet: E-Journals Consortium” to facilitate the research and academic community of the country in terms of scholarly information with the help of state-of-the-art technology. The foundation of this program was not so easy because it was needed good amount of recurring and non-recurring expenditure with systematic planning. A Joint Technical and Tariff Committee (JTTC), comprises well known scholars and academicians of the country, has been established by the UGC to guide and monitor the entire project which has given the guidelines for proper implementation of the project. In this time of digital environment most of the scholarly literature is being published electronically and accessible on the Internet. So consortium model of subscription has finalized to serve the higher educational system of the country. To pursue this program establishment of a broadband network was important. At this juncture, UGC analyzed the problem of digital divide in different universities and to get rid of this, it was decided to modernize the university campuses with state-of-the-art technology by establishing a nationwide communication network. UGC entrusted this task of establishing infrastructure within the member universities to ERNET (Education and Research Network) India on a turnkey basis. Infrastructure in more than 130 universities has been established so far. The facility subsequently be extended to colleges, therefore, in near future each university will become a hub for colleges affiliated to it.

After a successful trial of more than three months this program was launched by His Excellency Hon’ble Dr. A P J Abdul Kalam, the President of India at Vigyan Bhavan, New Delhi on 28th December 2003 in the Golden Jubilee Celebration of UGC.

1. Why consortium only?

"Library is a growing organism", one of the Five Laws of Library Science given by Dr. S R Ranganathan, leads whole world to the flap of Consortium. Consortium is the joint venture of homogeneous institutions working for the same objectives. Being a part of consortium an individual library can spread its wings all around the world with more resources and more services. In today's scenario consortium is the cutthroat need of the hour, especially for libraries. Library consortium is the virtual way to cope with the different problems of libraries through proper coordination and cooperation. Apart from these, duplication can be avoided as the situation calls for optimum use of resources by rational use of funds and it can be worked as platform for training and workshops for providing strength to the information professionals as well as users.

2. Why E-Resource only?

'E-resource' is "A term used to describe a variety of resources in electronic format e.g. databases, 'the web', e-journals". No doubt the print mode of information is still dominant medium but now it becomes the secondary mode due to the innovations in information and communication technology (ICT's) and its involvement in manage, manipulate and disseminate the information. Today, user needs latest and authentic information with minute time lag. S/he cannot wait for weeks or months to get needed information. Here electronic resources are enabling to fulfill these needs, therefore the whole world is shifting from print resources to electronic resources. Some of the advantages of e-resources are:

- Speedy Information: One can get the information very quickly through e-resources even many weeks before as compare to print issues.
- E-mailing: One can e-mail the output or important article at his/her e-mail account or any researcher or academicians e-mail account but that one should be limited to copyright issue.
- E-mail alert: user can get information according to his/her desired journals or topics by registering oneself in the publisher's site.
- Hyper links: An E- resource contains the links to other cited articles, e-journals, and other supporting material making like audio, visual aids etc.
- Maintenance: E-resources have not the problem of wearing and tearing, stolen, binding and shelving etc.
- Multi access: E-resources have the facility to change the concept of single user to multi user access at one time. In simple words more than one user can access the same information simultaneously.
- Search facilities: E-resources have the different search options like simple or quick, advance search, which is totally based on Boolean- logic, and search within the search result.
- No time limit: there is no limit or restriction in respect of time, e-resources can be used on the terms of 24 x 7 hours.
- Economical: The access to electronic resources can be provided with 85-90% discount compared to print collections. This means access to more resources for less money.

Apart from above quoted benefits of e-resources, downloading facility is there one can download the data in different formats, e-resources needs less space, users can keep up to date themselves by availing different services like article alert service, table of contents (TOC), As Soon As Published (ASAP) articles, etc.

3. UGC-Infonet E-Journals Consortium

As we discussed earlier that "UGC-Infonet E-Journals consortium" has launched by UGC, is one of the biggest and ambitious program in the history of higher education in India. The main objective of this program to facilitate the research and academic community of the country by providing them latest, authentic and scholarly literature from all parts of the world with the help of state-of-the-art technology. The program came into existence with the cooperation and co-ordination of UGC, Education and Research Network (ERNET), Information and Library Network (INFLIBNET) Centre, and many prominent national and international publishers. Here UGC, is the main funding body, ERNET India is responsible for designing and maintaining the infrastructure within the each member university and, INFLIBNET Centre has the responsibility to execute the program in systematic and planned mode. The e-journals consortium has received 60% to 90% discount on subscription of these e-resources and varies from publisher to publisher. This effort enable the faculty and research scholars access to not only current issues of the journals but also getting access to 5 to 10 years back volumes and publishers like IOP and ACS are providing access to whole archive of their collections from volume 1 issue 1 almost the full text access from 1874. This will bridge the gap between the scholarly information, which was denied 5-10 years back and now they will have access to all of their collections. The access to scholarly literature has been made available from January 1st, 2004 to 50 universities start with. Additional 50 universities were given on trial basis and from 2005 the access to all collections are extended to these universities. Presently, 2000+ scholarly e-journals with 8 databases and 2 portals are being provided under this program and about 100 universities across the country are accessing these e-resources, some of them are accessing under trial period. The list of full text e-journals and bibliographic databases subscribed under this program are:

- Chemical Abstract Services (CAS): CAS is the most important, strong and costly service tool in the areas of chemical sciences. This e-resource needs a specialized training to access. Therefore, in 48 universities specialized training program on STN access to Chemical Abstract has been conducted for proper and systematic utilization. The CAS access is given to 10 universities through Sci-Finder and other Universities having the Chemistry subject are getting this access through STN service. The archival access is made available since 1907 onwards.
- American Chemical Society (ACS): Full text access to ACS titles is giving access to the prominent 31 full text e-journals from volume no.1, issue no. 1.
- Royal Society of Chemistry (RSC): Full text access to RSC titles is given access to 23 journals and 6 databases and the archival access is made available from 1997- onwards.
- Institute of Physics (IOP): Access is provided to 36 full text topmost journals in the area of physics, and the archival access is made available from Vol.1 issue.1 of all 36 IOP titles.
- Nature Journal: World famous 1 full text un-limited simultaneous access for Nature Weekly from 1997 onwards is available.
- Cambridge University Press (CUP): Access to 72 prominent full text e-journals of CUP are being subscribed under the scheme. and archival access is made available since 1997 onwards. Access to social science and humanities package of CUP titles are given from January 2005.
- Project Muse: The access is made available to 222 full text journals in the area of humanities and social sciences with the archival access mostly from 1999- onwards.
- Biological Abstracts (BA): One database for biological sciences more than 7.7 million archival records are available back to 1969.

- American Institute of Physics (AIP): Access to 19 Full text journals with Archival access from 1997 onwards.
- American Physical Society (APS): Access is made available to 8 Full text journals from 1997 onwards.
- Encyclopedia Britannica: This Encyclopedia Britannica is one of the popularly used reference tool and can be used by Faculty and research scholars including colleges across the country. INFLIBNET has the National Site License for this reference tool.
- Science Online: This is a popular science magazine with 52 issues a year and provides access to the full text of all SCIENCE contents published from October 1996 through the latest issue.
- Annual Reviews: Access is made available to all 29 full text journals and archival access is provided up to 10 years back issues.
- Kluwer Journals: Access is made available for approximately all the 650 journals of Kluwer online for one year and after that members can access top 100 Kluwer journals. The archival access is provided from 1997 onwards.
- Springer Online: Access is made available for around 550 journals from Springer Link for one year and after that member institution will be continued to make access for 100 top Springer journals. The archival access is provided from 1997 onwards.
- Emerald Journals: Under UGC-Infonet e-journals consortium access is made available for 28 e-journals from Library and Information Science full text database and archival access is varies from journal to journal (mostly 1994- onwards).
- Elsevier Science: One can access the 34 full text e-journals of health sciences from Cell Press, Current opinions and, Trends. The access is made available through thewww.sciencedirect.com and archival access is provided from 1995 onwards.
- J-STOR: The most awaited e-resource which deals with back volumes of social sciences, humanities and to some extent with natural science subjects. For accessing JSTOR each university perform the Network Performance Test (NPT). This test is essential for getting access to the JSTOR. JSTOR access is presently given to 24 universities selected by the committee. Member universities can access to 319 full text e-journals from Vol.1 issue 1- onwards up to last two-three years gap depending on the moving wall period of original publisher rights.
- Ingenta and J-Gate, 2 subject gateway portals are also available and both of these are gateways for more than 20,000 journals. One of these portals is provided to each university in the first year, however five major universities have been provided both the portals to get the feedback on these portal services. The portals provide one stop solution to all publications and get access to full text access for the above titles from single window.

Apart from these subscribed e-resources negotiation with prominent publishers is being considered and the number of scholarly journals is likely to be doubled from January 1, 2005.

In the era of digital divide academic and research community are in the crucial period of transformation especially in India. They have to be dependent upon the electronic based collections rather than print based resources due to faster and quicker means of searching, browsing and interlinking facilities. Their expectations have been growing tremendously in this electronic age. But the computer literacy rate, which is a prerequisite in accessing electronic resources, in India, is comparatively less then many developed countries. Along with this it brings out challenges like copyright, archiving and how to exploit these available e-resources are some of the major aspects one should be aware about. So, there is need for awareness among the Indian academic and research community for proper utilization of subscribed resources.

5. User Awareness Programs

Role of user awareness program : After getting such an enormous amount of scholarly literature one should have to be aware about all these e-resources. All the publishers are providing many facilities to the users, some facilities are listed below:

The main and foremost motive of these awareness- cum- training programs is to ensure the proper utilization of e-resources by the users.

- How can user make effective use of available e-resources?
- What e-resources users are getting?
- How to make access these e-resources?
- How to approach abstract, full text of needed article?
- How to download?
- How to search a particular article/page?
- How to use different search options?
- How to avail e-mail alerts from the side of publisher, etc.?

Now, it is important for the user to utilize these resources in systematic and exhaustive manner. Therefore, the training, orientation and awareness are very much important to inform the user. To keep proper usage of these e-resources INFLIBNET Centre has already been conducting different training, awareness and orientation programs all over the countries. These programs can be divided broadly in four different groups.

5.1 One day user awareness program

This type of training program has been conducted at more than 35 universities across the country. The users of these programs are faculty members and research scholars and library staff from the host university.

5.2 STN Training Programs:

The same way access to Chemical Abstract through STN needs specialized training and it has been also carried out by INFLIBNET in different places in collaboration with STN staff from Pune. Under this program, participants of more than 45 universities have been trained. This training is especially for subject experts in chemistry from various universities viz. faculty and research scholars and the librarians / asst librarians of these universities. Around 2 to 3 chemical experts and 2 to 3 library staff are trained under the program.

5.3 E-Resource Management Using UGC- Infonet Training Program

INFLIBNET has been providing 5 days training to a person (mainly from the library science background) who is nominated by the Vice-Chancellor of that particular university to carry out and co-ordinate the UGC-Infonet project within his or her university. Six training programs have already been conducted, each programs contains 20 to 25 library and information professionals. Till now, professionals from 96 universities have already been trained.

5.4 National Seminars on E-resources

INFLIBNET has conducted five national seminars at five different places across the country to make the academic and research community aware about UGC-Infonet E-Journals Consortium and to solve the problems and doubts of the users in big gatherings. Vadodara (MS University of Baroda, 25-26 Oct., 2004), Goa (Goa University, 1-2 Nov., 2004), Bangalore (Bangalore University, 1-2 Dec., 2004), Kolkatta (Jadavpur University, 10-11 Dec., 2004) and, New Delhi (University of Delhi, 14-15 Dec., 2004) were the five places and huge number of faculty, research scholars and library science professionals working in the universities funded by UGC have attended and taken benefit from these seminars.

Enhancement of systematic approach, proper utilization of e-resources is the main objective of these user awareness programs. Usage statistics is the important tool to evaluate the usages of particular journal or particular subject area. But for this kind of usage statistics we have to depend on the publisher and publisher has to depend on particular tool/software to maintain the statistics. The usage statistics are coming up is very much encouraging to the authority.

Table I

This table is showing the statistics provided by some publishers regarding downloaded hits by the member university under UGC-Infonet E-Journals consortium (January-July, 2004)

Publisher	**Jan**	**Feb**	**March**	**April**	**May**	**June**	**July**	**Total**	**Avg. based Total**	**% of Total Download**
IOP	5,521	6,276	4,494	5,823	4,074	—	—	26,188	36,659	08.64
Elsevier	7,225	5,666	3,977	6,814	—	—	—	23,682	41,440	09.76
Springer	__	__	__	2,267	4,083	3,776	—	11,146	26,005	06.13
Kluwer	__	__	__	9,194	11,158	9,904	11,778	42,034	73,556	17.33
Kluwer (2nd ph univ)	__	__	__	__	924	2,646	9,668	13,238	30,891	07.28
ACS	15,617	20,985	26,611	24,647	24,487	24,696	—-	1,37,043	1,59,880	37.65
Sci-finder (CAS)	2,201	4,146	4,412	4,312	3,737	4,783	5,189	28,810	28,805	06.78
Nature	3,585	3,337	4,796	4,817	3,105	3,749	__	23,389	27,286	06.43
Total	34,149	40,410	44,290	57,874	51,568	49,554	26,635	3,05,530	4,24,522	100.0

Note

- Total number of articles downloaded through some publishers only (month- wise)
- Statistics from many publishers are still awaited
- Springer and Kluwer online has been started in April 2004
- Statistics awaited
- Statistics of universities which are coming under 2nd phase of this program

Table I: The usage statistics in table I are showing that in the months of March and April of 2004 is very much high the reason is INFLIBNET Centre has conducted 36 one day user awareness programs in various universities in the months of March and April of 2004. There is continuous increase in the article download from January 2004 to April 2004. The main reasons for slightly decrease in the use in the month of May-June due examination period in most of the universities. In this duration students are busy preparing for their examinations, admissions and vocations. Almost half of the search has been carried out for chemistry subject only.

Chart - 1

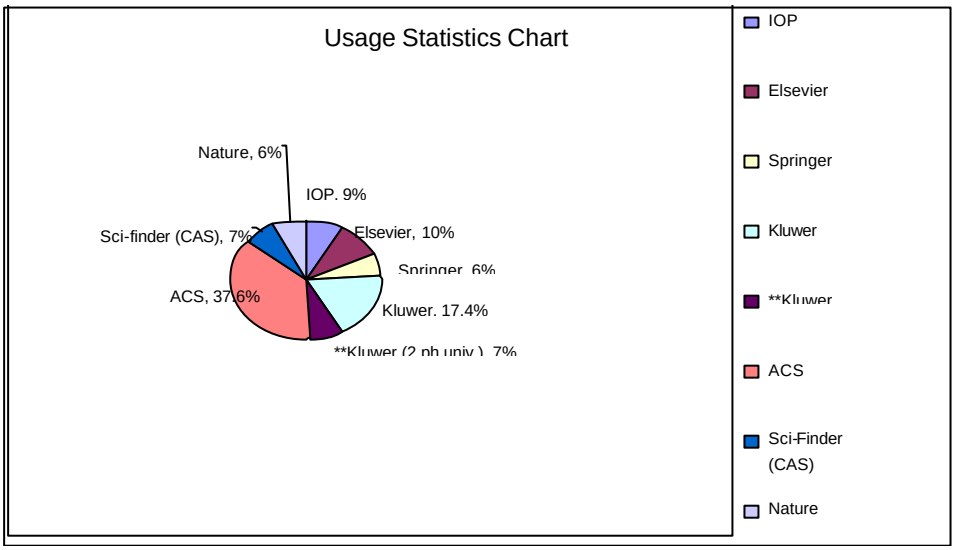


Chart for usage statistics: publisher-wise (average based) under UGC-Infonet (January- July 2004)

The usage statistics chart is showing that 45% of the total downloads are for Chemistry, 37% of the total downloads are from American Chemical Society alone. Users gave 24% importance to Kluwer online.

5.1 Information Service in Informatics Lab at the Centre:

Information service of Informatics Lab is completely a virtual service has been started by the Director, INFLIBNET. The Informatics Lab has wireless connectivity with the state-of-the-art technology using WI-FI. This lab is open to the scholars from different universities across the country. The faculty and research scholar visit the lab take the benefit of search and download facility of the lab and also photocopy of the print journals of the required article from the national archive set up by the centre. The undertaking has been collected from the users regarding fair and academic use of information. The user can come to the lab and can search the information from different online journals himself/herself under the assistance of staff of the centre.

6. Conclusion

Probably for the first time in the history of higher education system in India, higher education has been given prominence and access to many scholarly journals is made available from the support of University Grants Commission. Presently the system allows universities covered under the purview of UGC but gradually it will be extended to colleges and different R & D Institutes of the country. The ultimate goal of this program is work on the virtual philosophy of libraries i.e. right information to the right user at the right time with the help of state-of-the-art technology. Access is more important rather than collection development, whatever you have, that should be accessible. User awareness programs are started working as a tool to achieve the goal of qualitative and authentic research output from the side of Indian universities with the help of scholarly and updated information. This consortium bridging the gap between information and it's end user. Now we have latest information and it is responsibility of the end to utilize and exploit it in systematic and appropriate way. It is expected that e-subscription initiative of UGC-Infonet will bring remarkable change in the academic environment in the country.

7. References

1. INFLIBNET Centre, Ahmedabad, http://www.inflibnet.ac.in

2. Murthy, T A V: Kembhavi, A and Cholin, V S (August 23-29, 2004), 'Access to Scholarly journal and databases: UGC-Infonet: E-Journals Consortium' University News: a weekly journal of higher education, Volume 42, Number 34; ISSN 0566- 2257.

3. Cholin, V S (2003); Consortia for Libraries and Information Centres, retrieved 19th April 2004 from www.alibnet.org

About Authors



Dr. T.A.V. Murthy is currently the Director of INFLIBNET and President of SIS. He holds BSc, MLISc, MSLS (USA) and PhD. He carries with him a rich experience and expertise of having worked in managerial level at a number of libraries in many prestigious institutions in India including National Library, IGNCA, IARI, University of Hyderabad, ASC, CIEFL etc. and Catholic University and Case Western Reserve University in USA. His highly noticeable contributions include KALANIDHI at IGNCA, Digital Laboratory at CIEFL etc. He has been associated with number of universities in the country and has guided number of PhDs and actively associated with the national and international professional associations, expert committees and has published good number of research papers. He visited several countries and organized several national and international conferences and programmes.
Email : tav@inflibnet.ac.in



Dr. V S Cholin, Scientist - B is working with INFLIBNET Centre for last 11years. He was the coordinator for INFLIBNET Training courses and workshops. Awarded Ph. D in Library and Information Science from Karnatak University, Dharwad. Awarded Fulbright professional fellowship 2004-2005, ASIS & T International Paper Award 2004 and also awarded SIS-Professional Young Scientist-2004. He has received the best paper award in the Raja Rammohun Roy Library Foundation RRLF for writing the best professional article 2002. Attended 69th IFLA International conference held during August 1-9, 2003 at Berlin, Germany. He has contributed a course block to Post Graduate Diploma in Library Automation and Networking (PGDLAN) course of Indira Gandhi National Open University (IGNOU). Coordinating Bachelor and Master's Degree and PGDLAN courses of IGNOU. He has more than 25 papers to his credit. He has visited several Universities to deliver lectures training etc. He is currently heading the Informatics division of the centre and looking after the prestigious UGC-Infonet E-Journals Consortium.
Email : cholin@inflibnet.ac.in



Shri Suresh K Chauhan, is currently working INFLIBNET Centre as a Project Scientist for the last one year. Did his MA(English) and M L I S from Panjab University Chandigarh. Actively involved and assisting in UGC-Infonet E-Journals Consortium initiative of the Centre. Delivered lectures, presentations in the national seminars and workshops conducted by the centre time to time. As a faculty visited universities for user awareness programs and also SOUL installation programs. Presented three papers. Prior to joining INFLIBNET worked as Library Assistant at the Centre for research in Rural and Industrial (CRRID) Chandigarh.
Email : chauhan@inflibnet.ac.in or sureshbabal@yahoo.com



Shri Raghavendra Patil is currently working as Library Trainee (Library Science)at INFLIBNET Centre, Ahmedabad. He has done his Masters Degree in Library and Information Science from Karnatak University Dharwad with first Class. He worked for Library and Documentation Centre of Centre for Environment Education (CEE), Nehru Foundation for Development, Ahmedabad for six months. Presently he is working with INFORMATICS group of INFLIBNET centre since last 1 and ½ years as a library trainee. He participated in various National level seminars. His areas of interests are E-Journals consortium, Digital Library, Online and Offline Information Services, Man power training in effective use of Electronic Resources etc
Email : raghav@inflibnet.ac.in

Potential Role of Subject Gateways, Portals and OPAC's in Electronic Journals Access

K Prakash V S Cholin
T A V Murthy

Abstract

Outlines the access methods and new technologies in accessing local and global electronic resources in the libraries. The subject gateways, portals, search engines and Library OPAC's (Online Public Access Catalog) are an important method of providing current and reliable information in a variety of disciplines and research areas. This paper describes various access points that disseminate information to researchers, librarians, and other web users in the various disciplines. Summarizes some of the issues, and explore the potential role of World Wide Web portals in helping library consortia to fulfill their objectives. At the end the paper highlights the importance of portal service to the Indian academic community in the light of UGC-Infonet E-Journals consortium.

Keywords : Aggregators, Portals, Gateways, E-Resources

0. Introduction

Over the last decade there has been an information revolution using digital methods of publishing and online access. This has led to certain print based journals being available online and the frequent appearance of new e-journals. Libraries can benefit greatly from this publishing phenomenon, particularly with the utilisation of high-speed Internet access. Busy library staff is faced with the task of selecting e-journals appropriate to the needs of teachers and students. However, with this change, the professional staff will have new challenges and responsibilities for librarians to manage this environment in terms of wise expenditure of funds, obtaining reasonable contracts for use, ensuring ongoing access to digital content (archiving), preserving the rights of authors/publishers/users, maintaining rights of privacy, and offering the best possible access.

Will portals be the answer to managing and providing access to resources available for academic libraries as well as other content needed by the academic community? Will consortia help move libraries towards a different way of providing access? What about our role in the academy? Many academic institutions have selected a portal as the way to provide access to a wide range of information to current and prospective students, alumni, and donors. Institutional portals can help support the educational mission of the institution and can develop new constituencies. Library portals can also do the same.

Libraries are searching for a Google - like tool backed by authoritative citings and numerous options. The ideal portal will help users overcome the information overload that is besetting libraries and will combine powerful searching with the diverse resources and services that users find when they use a library. Portals should provide library experience of that quality without actually requiring people to come to the library.

1. Consortia Initiatives in India

Several library consortia around the country have been formed on different lines and objectives. Presently many consortia are functioning viz. FORSA (Forum For Resource Sharing In Astronomy & Astrophysics),

CSIR Consortia, HELINET (Health Sciences Library & Information Network), INDEST (Indian National Digital Library In Science And Technology) and UGC-Inofnet E-Journals Consortium etc.

All these consortia offers access to multiple electronic resources from different reputed publishers. The resources subscribed under the consortia can be divided in to following three categories.

- Full text Journals and Databases: JSTOR, Project Muse, Society publications, etc.
- Bibliographic Databases: Chemical Abstracts Service, Biological Abstracts etc.
- Gateway Portals: Ingenta, J-Gate, Ebscohost etc.

1.1 UGC-Infonet E-Journals Consortium

The University Grants Commission (UGC) has initiated a programme to provide electronic access over the Internet to scholarly literature in all areas of learning to the university sector in India. The programme is wholly funded by the UGC. All universities which come under UGC's purview will be members of the programme, and it will gradually be extended to colleges as well. The programme is being executed by Information and Library Network (INFLIBNET) Centre, Ahmedabad, which is an autonomous institution under the UGC. Access to various E-Journals has started from January 1, 2004.

The programme will increase in a very fundamental way the resources available to the universities for research and teaching. It will provide the best current and archival periodical literature, from all over the world, to the university community. The programme will go a long way in mitigating the severe shortage of periodicals faced by university libraries for many years, due to the ever widening gap between the growing demand for literature, and the limits of available resources.

The E-Journals programme is a cornerstone of the UGC-INFONET effort, which aims at addressing the teaching, learning, research, connectivity and governance requirements of the Universities. The E-Journals programme demonstrates how communication networks and computers can be used to stretch and leverage available funds in furthering these aims. The UGC provides funds for the programme, which will be cost free for the universities.

The E-Journals programme aims at covering all fields of learning of relevance to various universities including:

- Arts, Humanities and Social Sciences
- Physical and Chemical Sciences
- Life Sciences
- Computer Science, Mathematics, Statistics, etc.

The literature made available includes journals covering research articles, reviews and abstracting databases. Access is provided to current as well as archival literature. Portals are provided which will enable users to navigate easily through all the literature that is made available under the program.

2. Improving the Usability of Electronic Journals

Encouraging the use of electronic journals is in the long-term interests of vendors and libraries. Publishers and librarians should cooperate in removing barriers that discourage legitimate use. Uncertainty about what is contained within a particular electronic journal or journal aggregation package can cause users

considerable trouble and discourage use. Such uncertainty can be eased by providing libraries and their users with clear and consistent information about the e-journal content and the rules for its use. The libraries should understand the transition period and make its best efforts to change the mind sets of its users with number of awareness programs, training etc.

2.1 Finding and accessing electronic journals

Electronic Journals are electronic versions of print journals or, in some cases, journals available in electronic form only. Electronic journals are made available most often via the Web. Whilst electronic journals offer a whole range of benefits for users, such as desktop delivery, multiple access, distribution around a university campus, keyword searching etc., their arrival presented new challenges for libraries in terms of access and management. It was often necessary for libraries to work with each individual publisher to arrange electronic access - obtaining and distributing passwords, IP authentication, handling registration and licence arrangements, etc. Users themselves had to learn a variety of different publisher's systems, interfaces and search engines. These administrative and access difficulties became significant as numbers of electronic journals have grown. Libraries were keen therefore to find services to streamline and simplify the management of, and access to, electronic journals - using intermediaries or one-stop shops, which offer a comprehensive all-round service viz. portal service.

2.2 Problems in accessing e-journals: Finding the right e-journal

The task of locating appropriate e-journals can be exhausting. A good starting point is to use a search engine such as Google. Results can be rewarding, but searching often requires time, thoughtful selection of search terms and patience to obtain the most suitable resources. Gateway searches through university-subscribed databases are often of the required academic level. Many university libraries started developing portals and journal search sites such as provide users with a searchable database, giving details of e-journals. Annotated lists of journals provide a précis of e-journal sites, which can aid in the search for appropriate journals. By shifting through the multitude of publications and subscribing, or selecting, only the best material, library staff can introduce their library users to a range of timely, cost-effective and relevant e-journals. Library staff can add significantly to their own knowledge base with access to relevant library, education and technology e-journals. The role of intermediary is essential here to provide the end user a better access to his/her required resources.

3. Electronic Journals Access Made Easy

Subject gateways, Portals, Aggregators, Search Engines and Library OPACs are an important method of providing current and reliable information in a variety of disciplines and research areas. The general features of such a system are described in the next section.

3.1 Search Engines

A search engine is a collection of software programs that collect information from the Web, index it, and put it in a database so it can be searched. Search Engine is automated keyword searching tools, it use piece of software, usually known as a 'spider' or 'crawler' to gather the information from web and other servers and generate indexes. Search engine crawl the networks continuously to update their databases. It usually indexes the full-text of web page, meta data and holds lot of information in the databases. They are quite comprehensive and freely available but not complete. The key difference between subject gateways and the popular automated large-scale Web indexing systems such as AltaVista is the quality of the results, which the end-user receives. This is dependent on the nature of the cataloguing process.

Web search leader Google Inc. unveiled Google Scholar, a new search product aimed at helping users search scholarly literature such as technical reports, theses and abstracts. Google Scholar, at http:// scholar.google.com, searches a specific subset of Google's index and covers a wide range of fields, from medicine and physics to economics and computer science.

3.2 Subject Gateway

A subject gateway, in the context of network-based resource access, can be defined as some facility that allows easier access to network-based resources in a defined subject area. The simplest types of subject gateways are sets of web pages containing lists of links to resources. Subject gateways are online services and sites that provide searchable and browseable catalogues of Internet based resources. Subject gateways will typically focus on a related set of academic subject areas viz. SOSIG, EEVL, etc.

3.2.1 Basic gateway facilities

Most subject gateways allow the end-user to either search or browse the database of resource descriptions. For example, the SOSIG gateway consists of a browsable multi-level menu of sub-areas and resources, as well as a WAIS-based search mechanism. In addition, most gateways allow the user the options of case sensitive searching and stemming, where resource descriptions containing variations of a term are located. SOSIG incorporates a thesaurus containing social science terminology. This gives users the option of generating alternative terms/keywords with which to search the resource catalogue. SOSIG also allows users to search on resources that are located in distinct geographic areas, such as in the whole world, just in Europe or just in the UK.

3.3 Subject Portal

The Joint Information Systems Committee (JISC) defines a portal as "A network service that brings together content from diverse distributed resources using technologies such as cross searching, harvesting, and alerting, and collates this into an amalgamated form for presentation via a web browser to the user. For users, a portal is a, possibly personalised, single point of access where searching can be carried out across one or more than one resource and the amalgamated results viewed. Information may also be presented via other means, for example, alerting services and conference listings or links to e-prints and learning materials."

3.4 Online Public Access Catalog (OPAC)

OPAC is online public access catalog. In traditional library it provides information about what is available in the library. Present day OPACs mostly reflect now a day's print as well as electronic collections. Research library catalogs serve as authoritative sources of access. The phrase, "if you can't track it, you don't own it," is quite real for the library that is trying to monitor thousands or millions of items. From the last few years, libraries have started retrospective conversion projects, bringing metadata about all their monographs and other collections into one place this has helped librarians and users to know about what is available in the library. But relatively few libraries track some of the newest, and most popular, resources they provide: the electronic journals available through database aggregators and online publishers.

4. Role of Aggregators

A company who specializes in selling content from multiple sources via the Web. Generally, the aggregator's site is focused on a particular subject matter. Although aggregators are most common in the Scientific, Technical and Medical (STM) world, many are now popping up in other fields such as

Libraries, Technology Management, Education and other areas. These are companies, which have come into being as part of the electronic services environment to offer libraries a range of bibliographic and full-text services through a common search interface. Libraries have placed great faith in aggregators such as Ebsco, Ovid Technologies, Ingenta etc. because of their proven ability to package and present services in a reasonably transparent fashion to end users. The role of such aggregators, however, has been challenged over the past couple of years by the large commercial publishers who, for the time being, deny the aggregators access to valuable information content. This has resulted in libraries to have caution before they get on to access the journals through aggregator products.

There are, however, significant opportunities for content providers and aggregators to make strategic alliances, where both parties can benefit without eroding each other's market. From the customer viewpoint such strategic alliances are highly desirable as it works out economical. The interfaces and access which aggregators provide to libraries cover a range of bibliographic and abstracting services which are highly valued by customers. The value-added component from the publisher viewpoint is that the customer can be directed in a transparent way to the content service whenever a citation to a journal in that content service is picked up through the bibliographic search on the aggregator platform. Such 'arms-length' arrangements are already a developing feature in the evolution of the market place.

The larger subscription agents developed services to address these issues, offering a broad range of functions and benefits, including:

- Single interfaces and search engines for accessing a range of titles from multiple publishers;
- Keyword searching across all the titles offered in the service, usually searching within the tables of contents and abstracts;
- Browsing the contents of selected or favorite journals, as new electronic issues become available;
- Simplified password administration, usually on the basis that one password, or a small no. of passwords, allows access by all members of an institution to all titles in the agent's service;
- Library management functions, including the facility to input information on holdings of print titles;
- Regular data for libraries on the usage of the electronic journals;
- Services to alert end users to new tables of contents as well as more comprehensive SDI services.

4.1 Ingenta

Since its launch in May 1998, Ingenta has developed and grown to become the leading Web infomediary empowering the exchange of academic and professional content online. Ingenta provides publishers of academic and professional content with technology-driven solutions for leveraging the Web as a profitable distribution and marketing channel. Ingenta provides libraries and researchers with access to the most comprehensive collection of academic and professional content available online. More than 8,000 academic, research and corporate libraries, institutions and consortia, from around the world, currently rely on Ingenta for managed access to academic and professional content.

The acquisition of Catchword helped Ingenta to give publishers access to the most comprehensive set of online publishing solutions available in the market.

Ingenta has created a specialized gateway for users of UGC-Infornet consortium. This Gateway is a searchable database of more than 11 million citations from over 20,000 journals. The Gateway is a powerful, easy to use service as a means of expanding access to current, scholarly research. Electronic,

fax and Ariel document delivery is also available for millions of articles with fees. The purpose of subscribing to Ingenta services is to provide one stop solution to the faculty and research scholars through single window access to large number of journals. The end users gets full text access to subscribed titles and up to the abstract level information for all other collections which are not subscribed under UGC-Infonet E-Journals consortium. In simple terms users instead of searching each of the 20-25 publishers individually, they search in a single window through this portal and get full text access to all the collections subscribed.

4.2 J-Gate

The features of J-Gate, which provides an integrated search engine to the contents of all journals subscribed by the consortium. Other features include: abstracts and contents of over 12,000 journals, access to free e-journals, email addresses of the authors, better coverage of Indian journals, list of libraries which subscribe to the journals etc. Facilities are also available for browsing by subject, publisher or journal title. Full text access to all the free journals is also available. Another important feature of this portal service is that the user can restrict their query to only UGC-Infonet subscribed journals while searching. It also offers Table of Contents (TOC) service for about 12,000+ e-journals. It has a comprehensive searchable database with good number of articles, with 4,000+ articles added every day.

4.3 SwetsWise

SwetsWise is a journal subscription management tool for information specialists. The service offers a straightforward interface simplifying online access to electronic publications as well as allowing to control and organize subscriptions efficiently and effectively.

4.4 EBSCOhost Electronic Journals Service

EBSCOhost Electronic Journals Service (EJS) is a gateway to thousands of e-journals containing millions of articles from hundreds of different publishers, all at one web site.

5. E-Journal Holdings Data Services

Keeping track of the specific holdings available through all of a library's subscriptions to electronic journals can be a daunting task. This is especially true for the products that combine a large number of electronic journals, such as Elsevier Science, Kluwer etc. The titles and holdings may vary over time, and it may be difficult to determine the specifics of the beginning and ending dates of each title. The number of historical issues may change during the license period as the publisher digitizes additional material. Some titles may vanish from one aggregated service and turn up in another as the aggregators compete for access to content.

The volatile nature of the aggregated electronic journal products plus the sheer number of titles involved create an enormous amount of work for those who maintain the serial holdings in the library's catalog. To help the users INFLIBNET is maintaining a serials union catalog and it can be accessed by web from http://www.inflibnet.ac.in under the database category. For each electronic journal, the corresponding record would have a URL that allows users to link to that journal on the Web.

5.1 Accessing Electronic Journals Through INFLIBNET Website

The UGC-Infonet E-Journals Consortium has its website providing access to various types of information such as number of titles, subjectwise arrangement, publisherwise arrangement, contact details etc.If

you have a specific journal title which you wish to access, or if you want to check whether a particular title is available electronically, the simplest method is to go to the A-Z list of e-journal titles.

You may notice that some journal titles appear more than once in the subject list. This is because these titles are multidisciplinary subjects. It is useful to list all available links for each title because in the event of difficulties accessing the title via one service then the alternatives may still be working. Another common problem with all e-journals is that coverage is not always as complete as it may seem. For example, when you access a journal's web page you may notice that one or more issues of a particular volume may be unavailable. Sometimes individual articles from issues are missing. This comes about for various reasons involving the supply chain between publishers and service providers. However, if you have an alternative service to try you may find that your article or issue is available there. For most e-journals Adobe Acrobat (PDF) reader is required to view the articles in full text. This software enables you to view and print the articles in the same format as they would appear in the printed version.

6. Emerging Technologies

The online catalog of a library provides one means for accessing electronic resources. Through title searching and subject headings, users can find any electronic journal the library subscribes to and go to that journal through the link provided. The main limitation of this approach is that it works only to find the journal itself, not the individual articles. Now many technologies are coming to help the librarian in providing access to the e-journals.

6.1 E-Journal Locator Resources

Many western country libraries maintain an electronic finding aid that consists of lists of electronic databases and e-journals on their Web site apart from the main online catalog. These e-journal locaters work as good navigational tools for researchers who want a quick way to get to an e-journal without the complexities of the online catalog. These lists of e-journals may in fact be database-driven applications that also offer significant information about each e-journal, including the dates of coverage and a description of the types of material available, in addition to the title and URL. Like the online catalog, this approach takes the researcher to the e-journal itself, and not to individual articles. Keeping these journal locater applications up-to-date also requires significant effort. Rather than relying on manual work, many libraries will extract data from their online catalog or rely on an e-journal holdings service to automatically populate the e-journal locater.

6.2 Linking to Full Text

Library users, however, might not care about finding an e-journal, but might want to read the full text of articles on their research topics. This process typically involves searching an Abstracting & Indexing(A&I) resource that yields lists of citations of the articles that contain the information. Finding good ways to link the user from that citation to the full text is one of the key challenges in the development of a library's information environment. Within self-contained, aggregated products like EBSCOhost and those from ProQuest, the process is simple and automatic. Yet, the scope of these products is limited to a specific set of disciplines. The real challenge lies in connecting the user that searches an A&I database with the full text in an e-journal that's located elsewhere. Citations in A&I resources are increasingly able to provide links directly to the full text of the article they describe. Through the efforts of CrossRef, an initiative of over 200 publishers, citations include digital object identifiers (DOIs) that can be used to provide links to full text. It is also important to provide links to full text from references within an article, allowing a researcher to easily navigate among resources.

6.3 OpenURL and CrossRef - Based Link Resolvers

In order for information providers to equip their products for optimal integration with library linking systems, they are being asked to implement the OpenURL. This has caused some confusion concerning primary and secondary publishers who use the CrossRef/DOI system for cross-publisher links to full-text, because of the mistaken perception that the OpenURL and the DOI are competing technologies. They are not. CrossRef and the DOI provide persistent identification of scholarly content and centralized linking to the full text and other resources designated by the publisher. The OpenURL enables library-controlled links to a multiplicity of resources related to a citation and is designed for localized linking.

Yet the linking that's possible through the publisher-provided links of A&I resources or in article citations isn't always effective. These links may point to resources that the local library doesn't subscribe to. Given that many resources are available through multiple sources, knowing which version to link to is a problem. It would be unfortunate for the link to point to the article in one resource when the researcher would have been able to access it through another. This scenario has grown to be called the "appropriate copy" problem. A growing genre of products has emerged in response, both to address this problem and to offer additional services and options to searchers as they navigate among library-provided electronic resources. The basis of these products is link resolvers that rely on a database of the library's profile of subscriptions to determine the appropriate links that a library user should be presented with in a citation. Through a standard syntactical construct called the OpenURL, the producers of A&I databases, the publishers of electronic information, and the developers of link resolvers are able to create an environment where all the components work together. If the local library uses a link resolver, a citation in an A&I resource would have a button for the user to press that would then launch a menu that presents the various options available, usually the link to the full text from the appropriate source. But since not all information is available electronically, other options might include a search in the online catalog to see if the library has a print version, or an option to request the item through interlibrary loan or document delivery.

Following are some of the major linking products available today:

- SFX from Ex Libris
- LinkSource from EBSCO
- LinkFinderPlus from Endeavor Information Systems
- WebBridge from Innovative Interfaces, Inc.
- Sirsi Resolver from Sirsi Corp.
- Article Linker from Serials Solutions
- 1Cate from Openly Informatics

6.3.1 Some basic definitions

The OpenURL is a mechanism for transporting metadata and identifiers describing a publication, for the purpose of context-sensitive linking. The OpenURL standard is currently on the path to NISO accreditation.

A link resolver is a system for linking within an institutional context that can interpret incoming OpenURLs, take the local holdings and access privileges of that institution (usually a library) into account, and display links to appropriate resources. A link resolver allows the library to provide a range of library-configured links and services, including links to the full-text, a local catalogue to check print holdings, document delivery or ILL services, databases, search engines, etc.

CrossRef is a system for the persistent identification of scholarly content and cross-publisher citation linking to the full-text and related resources using the DOI. CrossRef DOIs link to publisher response pages, which include the full bibliographic citation and abstract, as well as full-text access (for authenticated users or at no charge, as determined by the publisher). The publisher response page often includes other linking options, such as pay-per-view access, journal table of contents and homepage, and associated resources. CrossRef is a collaborative membership network, and not a product for purchase.

DOI stands for Digital Object Identifier and is an open standard. A DOI is an alphanumeric name that identifies digital content, such as a book or journal article. The DOI is paired with the object's electronic address, or URL, in an updateable central directory, and is published in place of the URL in order to avoid broken links while allowing the content to move as needed. DOIs are distributed by publishers and by CrossRef, and there is no end-user charge associated with their use. As an identifier, the DOI can be incorporated into many different systems and databases.

6.4 Federated Search

Another major area of interest is in applications that allow users to search multiple sources simultaneously so they don't have to decide which resource might have the information they need. This approach goes by various names: federated searching, cross searching, or metasearch.

A number of products with differing technological underpinnings are available in this category. The products are based on a mechanism that knows how to send a query to each individual resource behind the scenes, and then receive the results. When the user enters a search request, the system translates it into the form needed by each of the selected targets, gathers and collates results as they are returned, and then presents the orderly results. These metasearch applications typically involve presenting a set of broad subjects or disciplines, removing from the user the burden of knowing what kind of information is contained within each of the brand-name resources. As part of the configuration of the metasearch application, the library would maintain a profile of the electronic resources to which it subscribes.

These are some of the major products in this category now:

- ENCompass from Endeavor Information Systems
- MetaLib from Ex Libris
- Sirsi Single Search from Sirsi
- WebFeat Prism from WebFeat
- MuseSearch from MuseGlobal
- ZPORTAL from Fretwell-Downing

6.5 The value of persistent links

Static URLs are not a persistent linking mechanism. If a URL is published as a link and the content it points to is moved, then that link will no longer function. DOIs address this problem. For instance, the publisher may need to migrate content from one production system to another (pre-print to post-print), or content may move from one publisher to another if a journal, or the publisher itself, changes hands. In these cases the publisher simply updates the DOI directory; the DOI itself never changes, which means that all the links to that content that have already been propagated still function. An OpenURL link that contains a DOI is similarly persistent.

Among the range of linking options they might display, local link resolvers frequently contain links to full-text at the publisher's website, as when the library subscribes to the e-journal in question or otherwise wishes to provide its patrons with access to publisher services and access options. While OpenURLs without DOIs can function persistently if the relevant metadata is updated within the institution's link resolver, this process is greatly streamlined via access to the CrossRef system. CrossRef provides a single source for linking reliably to hundreds of publishers without the need to track varied metadata-based linking schemes. Therefore, link resolvers benefit from using the DOI wherever linking to publisher-designated resources is appropriate.

7. Conclusion

Accessing electronic resources shows that librarians and users are facing a complex set of challenges. While a number of products have evolved for each aspect of the problem, the question is, how can they all be designed and implemented in such a way that they all work together, providing a clear and seamless interface for library users and avoiding redundant work for library staff? To date, no single product exists that provides comprehensive management of electronic resources. Will portals be the answer to managing and providing access to resources available from academic libraries as well as other content needed by the academic community? Will consortia help move libraries toward a different way of providing access? And, what about our role in the academy? The portal is one of the services offered by digital libraries. Many academic institutions have selected a portal as the way to provide access to a wide range of information to students, scholars and teachers. Institutional portals can help support the educational mission of the institution and can develop new constituencies. Library portals can do the same. Librarians welcome the development of CrossRef and other publisher-based linking systems. However, to be fully effective, publisher-based systems must be linked to local library systems and to non-commercial sites. Future use will be heaviest for those publisher sites that offer the greatest variety of links and linking capabilities. At a minimum, all the applications that a library employs to manage its electronic resources should draw from the same knowledgebase of its electronic holdings. A library should not have to maintain the same information in multiple ways. If the library catalog, linking environment, electronic resource management system, and metasearch engine cannot all share the same physical knowledgebase, then it should at least be possible to have a master copy of the data that is automatically distributed through these applications.

8. References

1. Falk, Howard. 2004. Open Access Gains Momentum. The Electronic Library, 2004, vol. 22, no. 6, p. 527-530
2. INFLIBNET Centre http://www.inflibnet.ac.in/
3. Ingenta http://www.ingenta.com/
4. J-Gate http://www.j-gate.informindia.co.in/
5. Jackson, Mary E.; Preece, Barbara G.; Peters, Thomas A. 2002. Consortia and the Portal Challenge. Journal of Academic Librarianship, Vol. 28 No. 3, p160-62
6. McCracken, Peter. 2004. The OPAC Reborn (netConnect). Library Journal, July, 2004
7. Moyo, Lesley M.2004. Electronic libraries and the emergence of new service paradigms The Electronic Library, vol. 22, no. 3, p. 220-230
8. Open URL and CrossRef http://www.crossref.org/03libraries/16openurl.html

9. Prior, Albert. 2001. Acquiring And Accessing Serials Information - The Electronic Intermediary. Journal: Interlending & Document Supply. Vol.29 No. 2 p. 62-69.

10. Shouse, Daniel L; Crimi, Nick; Lewis, Janice Steed. 2001. Managing journals: one library's experience. Library Hi Tech. Vol.19 (2) p. 150 - 155

About Authors



Mr. K Prakash is working as Scientific/Technical Officer-I with INFLIBNET Centre since 1995. He has his basic degree in Science and Masters Degree in Library and Information Science from Karnatak University, Dharwad. He has qualified SLET. Pursuing research in Library Automation. He has done specialization course in "Information Technology Applications to Library and Information Services" from NCSI, IISc Bangalore. Before joining to INFLIBNET, he has worked in academic and industrial libraries. Presently he is working in Serials union database development & managing, content creation and management and involved in training and other activities of the centre. He has contributed more than 20 papers in seminars, conferences & journals. He is a life member of professional bodies like ILA, IASLIC, KLA, SIS etc. and he is managing Digilib_India discussion forum also. His areas of interests are Library Automation, Database Management, Information Retrieval, Organisation of e-resources, Digital Libraries and Training etc
Email : prakash@inflibnet.ac.in



Dr. V S Cholin, Scientist - B is working with INFLIBNET Centre for last 11years. He was the coordinator for INFLIBNET Training courses and workshops. Awarded Ph. D in Library and Information Science from Karnatak University, Dharwad. Awarded Fulbright professional fellowship 2004-2005, ASIS & T International Paper Award 2004 and also awarded SIS-Professional Young Scientist-2004. He has received the best paper award in the Raja Rammohun Roy Library Foundation RRLF for writing the best professional article 2002. Attended 69th IFLA International conference held during August 1-9, 2003 at Berlin, Germany. He has contributed a course block to Post Graduate Diploma in Library Automation and Networking (PGDLAN) course of Indira Gandhi National Open University (IGNOU). Coordinating Bachelor and Master's Degree and PGDLAN courses of IGNOU. He has more than 25 papers to his credit. He has visited several Universities to deliver lectures training etc. He is currently heading the Informatics division of the centre and looking after the prestigious UGC-Infonet E-Journals Consortium.
Email : cholin@inflibnet.ac.in



Dr. T.A.V. Murthy is currently the Director of INFLIBNET and President of SIS. He holds B Sc, M L I Sc, M S L S (USA) and Ph.D. He carries with him a rich experience and expertise of having worked in managerial level at a number of libraries in many prestigious institutions in India including National Library, IGNCA, IARI, Univ of Hyderabad, ASC, CIEFL etc and Catholic Univ and Casewestern Reserve Univ in USA. His highly noticeable contributions include KALANIDHI at IGNCA, Digital Laboratory at CIEFL etc. He has been associated with number of universities in the country and has guided number of PhDs and actively associated with the national and international professional associations, expert committees and has published over 110 research papers. He visited several countries and organized several national and international conferences and programmes
Email : tav@inflibnet.ac.om

Security for Libraries in the Digital Networked Environment

Manoj Kumar K Haneefa K M

Abstract

Libraries are using Information and Communication Technologies (ICT) for their operations and services by making huge investments and spending vast amounts of staff time for the selection, acquisition, retrieval, and dissemination of digital information. But the proliferation of computers, widespread acceptance of computer networks, explosive growth of Internet, increased reliance on electronic databases and the move from dedicated mainframe environments to client-server environments make libraries vulnerable to security threats. The moment user connects the computer to a Network or Internet, is the moment that the security of data has been compromised. Even the most secure systems, shepherded by the most intelligent and able system administrators, and employing the most up-to-date, tested software available are at risk every day. It is very essential to take all measures to protect the ICT infrastructure from security threats. However, libraries are lagging behind in realizing the need to protect their ICT resources and services from misuse, damage, theft, sabotage, mistake, etc. This paper deals with the issues related to the security of libraries in the present digital networked environment and makes recommendations for protecting ICT resources and services. The paper discusses security risks, strategies for security, security policy, personnel security, physical security, software security, network security, Internet security, access control, protection against computer viruses, protection of public terminals and backup information. This paper also discusses the need for professional assignments for library security and the importance of security training for library professionals.

Keywords : Security Risks, Security Policy, Internet Security, Access Control, Network Security

0. Introduction

With the advent of the Information and Communication Technology(ICT), the paradigm for libraries has dramatically changed due to the penetration of internet, communication technologies and the consequent elimination of the constraints based on the geographical boundaries. The conventional library system is undergoing rapid changes; it has transformed from secured physical location to less secured public domain systems. As the application of Information and Communication Technologies (ICT) in libraries is widely accepted, Libraries are using these ICT facilities for their operations and services by making huge investments and spending large amounts of staff time for the selection, acquisition, retrieval, organization and dissemination of digital information. But in the absence of sufficient security ICT may not be used to with its full potentials. The proliferation of computers, widespread acceptance of computer networks, explosive growth of Internet, increased reliance on electronic databases and the move from dedicated mainframe environments to client-server environments make libraries vulnerable to the security threats. It is very essential to take all measures to protect the ICT infrastructure of libraries. However, libraries are lagging behind in realizing the need to protect their ICT resources and services from misuse, damage, theft, sabotage, mistake, etc.

Information is the most valuable resource of libraries. This information should be stored in such a way that its integrity and availability is maintained. ICT allow libraries to store, preserve, index, retrieve and

disseminate information more easily and quickly. But it gives a lot of scope for misuse and abuse of electronic information. There are risks of loss from unauthorized access, use, modification or destruction of information, which may be caused accidentally or intentionally. If it is damaged or lost due to misuse, mistake, theft, sabotage, the very purpose of these resources is not achieved.

Digital networked environment of a library may include hardware, library management software, computer programs, electronic databases, data, information, etc. Now libraries are web accessible and increasingly handling web accessible information. Threats to these resources may arise from the failures of computer and communication hardware or software, malfunctions caused by bugs and viruses, overload or other operational or quality problems. Misuse or abuse of digital information may arise from the unauthorized access for the purpose of mischief, vandalism, sabotage, fraud or theft, etc. Digital resources of a library should be stored, retrieved and disseminated in an authorized manner and should be disclosed to authorized users only. It should be protected from threats to gain its availability, integrity and confidentiality. ICT infrastructure should be protected from physical threats. System and application software should be protected from non-physical threats. Computer networks should be protected from unauthorized access, interruption and manipulation. Access to digital information should be controlled through authorization.

1. Security Risks

The first task is to determine the type and level of security risks associated with the library. The assets to be protected should be determined. Usually more insidious threat to security comes from internal sources. The explosive growth of Internet had increased the security risks from external sources also. E-mail is one of the most used Internet service or tool, but E-mail became the most dangerous medium for security threats by spreading computer viruses through attachments. Downloading of programs and files using www or ftp is also threatening dangerously. In order to evaluate the security risks associated with the network of a centre, proper security audit methodology should be adopted.

In a security audit, a company's IT infrastructure and associated process is tested for reactions to known attacks. Those reactions are then analyzed to identify possible security weaknesses. These weaknesses are measured and prioritized so appropriate controls can be deployed

The assessment will be carried out in following steps

1. Data collection
2. Assessment of Existing Systems & Processes
   - ✍ Vulnerability assessment
   - ✍ Threat assessment
   - ✍ Security Process assessment
3. Information security policy document
4. Risk analysis & management
5. Recommendations

The information gathering for the purpose of this analysis should be done by the following methods by a security audit team.

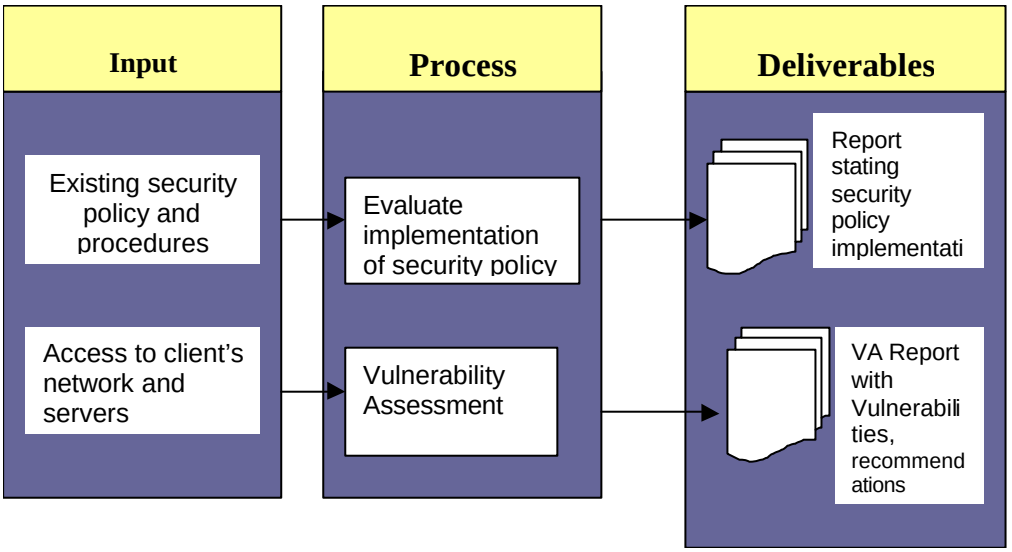
1.1 Questionnaires

A Security audit team will conduct interviews or administer detailed questionnaires that will require the client to list key business processes and assets in the order of criticality they perceive and link them to the business outputs. The team would ask the department heads to detail the departmental, interdepartmental and inter organizational dependencies of these processes and assets

1.2 Consultative discussions with the IT Team of Customer

The team will review the IT related policy documents that can provide good information about the security controls used and planned for the IT system. An organization's asset criticality assessment provides information regarding system and data criticality and sensitivity.

Assessment of Existing Systems & Processes



As part of this exercise, the task is carried out to evaluate the current operational posture of the Perimeter Architecture and servers of the organization in order to check various servers against their standard functional and non-functional requirements

2. Security Policies and Procedures

It is very important to establish well-formulated policies and procedures to protect ICT resources and access tools from security threats. A security policy defines the resources and services to be protected, discusses the technologies to be used for protecting the resources and explains how these tools should be deployed. The policy should include the purpose, scope, rules, standards and specific activities. It should cover the use of ICT resources, marking of sensitive information, movement of computing resources, disposal of sensitive wastes and security incident reporting. Enforcement of these policies is very essential to their effectiveness.

3. Personal Security

Personal security consists of management constraints and operational procedures to provide an acceptable level of protection for ICT resources. It includes procedures established to ensure that all personnel who have access to electronic resources have the required authorization and appropriate security clearance. It also includes personnel-oriented techniques for controlling people's behavior to ensure the confidentiality, integrity, and availability of digital resources.

4. Physical Security

Physical threats to ICT resources may come from extreme environmental events or from adverse physical conditions or from purposeful activity. Extreme environmental events include earthquake, fire, flood, lightning, excessive heat, humidity, etc. The digital networked environment of a library is susceptible to rough treatment and inexperienced users as well as knowledgeable vandals and thieves. Hardware can be stolen, damaged or destroyed. Peripherals can be moved, stressed, overused or damaged. Threat assessment should be carried out in a proper manner.

During this stage it is mandatory to explore threats to the assets that have been identified and will arrive at the likelihood of occurrence of adverse events due to these threats. For identification of threats threat analyse team should use multiple sources for enumeration of relevant threats. The sources that will be consulted for threat exploration must include the client representatives involved with the IT dept and threat database with details regarding post identification of threats (who and what causes the threat) and threat agents (who and what elements of the organization cause the threat). Most of the physical security threats are based on the following factors. Threat frequency: how often the threat might occur, according to experience, statistics. Threat Source Motivation and Skills: the motivation, the capabilities perceived and necessary, resources available for attacker etc. Geographical factors: proximity to chemical factories, areas of extreme weather conditions etc.,

Many security threats can be avoided by protecting the ICT infrastructure. The library should be fireproof with fire alarms, smoke detectors, extinguishers etc. It should have intrusion detectors and guards and should use surge suppressors or electronic power filters for all devices in area of power fluctuations. Secure computer/server room is a very important component of a good security program. Smoke and water detectors are essential features of a good computer room. ICT infrastructure should not expose to extreme cold or warm temperature and should be kept in an air-conditioned environment. It is a good practice to keep significant resources separate from general access equipments. All electronic media for storing digital information such as diskettes, CDs, DVDs, tape, etc. should be secure. Library professionals should clearly understand who to contact if urgent equipment repairs are needed and how to contact them.

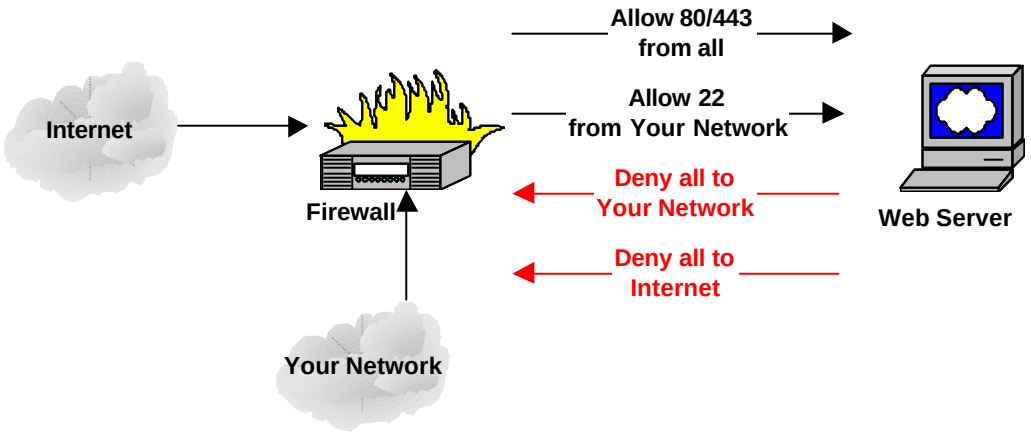
5. Software Security

Software security is very important for the smooth functioning of ICT resources and services. Software security threat includes the unauthorized access for breaking the nonpublic areas of the automated information system and viewing private records, changing files, or erasing records. Unauthorized access also includes introducing viruses to the system or using the system as a base for further unauthorized activities on any other system. Poorly programmed or configured software is always prone to hacking or cracking. Application software is usually very similar and therefore become well known to an intruder. Default locations for key files are often accepted so that an intruder can easily find the locations and tamper with them. The homogeneity of services and programming for different software reduces the time required for a person to find and alter key areas on the system.

Software should have adequate security measures and it should be protected from computer viruses. Library should be able to get local support from the software company and should be able to access a secure master copy in case the copy in use is corrupted or lost. It can be avoided by taking backup copies of major software. All software and new files should be regularly checked by reputable anti-virus software. The system administrator should keep a software toolkit for troubleshooting.

6. Network Security

The increased reliance on computer networks has made security a major issue. ICT infrastructure and access tools should be protected from unauthorized access, interruption and manipulation. Libraries are using Local Area Networks (LAN) to share resources, peripheral devices such as printers and scanners, to store or archive files and to exchange files through e-mail, ftp, telnet, etc. Most libraries use CD-ROM networks using CD-Net server or CD-ROM tower for sharing electronic databases. Once computers are connected to a network, they become vulnerable. Public access networked computers like OPAC can be used for other purposes if adequate security measures are not implemented. As libraries are forming library networks and library consortia, it is very important to verify the contents and origin of digital resources. The library networks and consortia should be protected from unauthorized access by various techniques such as encryption, remote access regulations, etc. Sensitive information should be encrypted, authenticated by digital signature, time stamps, sequence numbers and digital certificates. Libraries should ensure security by setting devices like routers, firewalls, proxy servers, etc. A firewall is usually a combination of hardware and software which will inspect network traffic, and allow or deny the flow of traffic based on some sort of rule set.



Firewalls filter the traffic exchanged between networks, enforcing each network's access control policy. Often, a firewall defends an inside "trusted" network from attack by "untrusted" outsiders. Firewalls ensure that only authorized traffic passes into and out of each connected network. To avoid compromise, the firewall itself must be hardened against attack. To enable security policy design and verification, a firewall must also provide strong monitoring and logging.

7. Internet Security

Library networks connected to Internet is the greatest risk to the digital environment of libraries. Internet is a virtual library, which revolutionized the ways of accessing, organizing, managing, retrieving and dissemination information. It provides different types of tools/services/utilities for accessing electronic resources all over the world. Information is disseminated openly over the Internet. In many cases the parent organizations provide Internet connectivity to their libraries. Once connected to the Internet, libraries become vulnerable to outside attempt to break into the library systems. It can facilitate undesired access to internal systems, unless systems are appropriately designed and controlled. The open architecture of the Internet also makes it easy for system attacks to be launched against systems from anywhere in the world. Library Internet systems can even be accessed and then used to launch attacks against other systems. Confidential information that sends over the Internet could be viewed, intercepted, or stolen. Any information accessed, stored, retrieved or disseminated on a web server may be susceptible to compromise if proper security measures are not taken. If proper access controls are not maintained data integrity could also be compromised. Web servers and internal networks can be secured automatically by using software programs. These software are effective to check unauthorized access to the system. It is the responsibility of library to ensure that all data is maintained in its original or intended form.

8. Access Control

Only authorized users should have access to ICT resources. Usernames and passwords are the most common way of authenticating users and monitoring their access. Most of the library management software provides several levels of password access for different modules. A password should not be easy to guess and it should not be a nickname, common word, a film title, or a character in a film or in a book, birth date or popular work. The best passwords are at least eight characters in length and a mix of uppercase and lowercase characters, numbers, and special characters. It should be changed regularly. All vendor-supplied passwords should be changed. Usernames and passwords may be deactivated when they are not required. Library should have additional security measure for public access workstations. It should be monitored through passwords. It is a good practice to remove, hide or rename dangerous or unnecessary programs from such workstations. User privileges on public access workstations should be limited.

9. Protection Against Computer Viruses

Library professionals and users should work in accordance with safe computing practices to minimize the risks associated with computer viruses. A computer virus is a program, which disrupts the normal operations of a computer, leaving unwanted messages, erasing data, or scrambling a system's configuration. There are two major types of computer viruses, file infectors and boot sector infectors. File infectors are attached to executable programs. Boot sector infectors are restricted to diskettes, hard drives, and other storage media. These media contain a boot sector that holds specific information about the formatting of the storage medium and data stored there. When the boot sector program is read and executed, the virus goes into memory and infects the hard drive. A third type of virus, the hybrid, infects both sectors and files. An important feature of any virus is that it replicates itself, usually by attaching itself to program files. To prevent the spread of viruses, library staff should be made aware of the potential sources of infection. The best defense against viruses is running anti-virus programs. Computer system should have up-to-date anti-virus software that checks for virus and repair them. A good anti-virus program scans a system for files that match its database of known viruses, and will also watch for the generic symptoms of virus infection. It is a good idea to scan for viruses after transferring a file into a system. Scanning should take place before the transferred file is executed or used. All hardware and software should be scanned at periodic intervals. The anti-virus software should be updated regularly to ensure its effectiveness. If the anti-virus software detects a virus from an incoming file, inform the people who introduced that file so they can ensure it does not happen again.

10. OPAC

The Online Public Access Catalogue (OPAC) of library should be protected from misuse and abuse. It can be used for unauthorized access to the digital resources of the library. Physical risk of theft or damage of the public terminal, and risk of unauthorized access to resources outside the library are the major security issues associated with OPAC. It is very important to consider any accessible item to be subject of tampering, theft, damage, sabotage, etc.

11. Backup Information

Backing up electronic information, software and other important electronic documentation is a reliable security technique. An accident or a severe environmental condition may destroy these resources. Errors and omissions may occur during accessing, creating, processing, storing, managing, retrieving and transmitting data and information. If the information is destroyed or corrupted, there must be copies of the information that can be restored to the system. So it is very essential that backup copies should be readily available. Regular backups must be performed to ensure that no data are lost in the event of equipment failure. If files have not been backed up, the library may incur significant expense in time and money in recreating them. Backup information should have the same level of protection as the active files of this information. It should also be kept in a secure location physically separate to the one in which the computer system is located. It should not be kept near any magnetic fields, extreme heat or cold and should have adequate protection against fire and other physical hazards. Backup logs should be examined on a daily basis to check that backup has been completed satisfactorily. There should be written procedures outlining all aspects of backup procedures. Generally the system administrator is responsible for backups. It is very important that in addition to the system administrator one or two other professionals should know how to backup and access backed up information.

12. Professional Assignments and Security Training

All library professionals should be aware of their responsibility in relation to security. All major information related procedures should be documented so that important procedures can be followed when concerned library professionals are not available. More than one staff member should know important procedures. Library computer system should be assigned to a system administrator who is responsible for the maintenance and security of the system. Maintenance of digital information is very essential in order to avoid inevitable decay due to interaction with the environment. To be diligent about the security of the systems, library personnel with specific security related job descriptions are a necessity.

Libraries should give training to library professionals in security. Security awareness and education for library professionals and users are critical to good security practice. It is a preventive measure that helps users and library professionals to understand the benefits of security. Technical training in the form of emergency fire drills for library personnel can ensure that proper action will be taken to prevent such events from escalating into disasters. Security related magazines, mailing lists and newsgroups should be subscribed for more information, suggestions and warnings about security. A Network Security Administrator should be able to do the following jobs after proper training.

- ✍ Assimilate information gathered on the network and classify the criticality of the information. Reading documents and e-mails as they flow in the network.
- ✍ Sniff the network to capture network traffic. Put agents on the network to scrutinize and store all the network traffic logs.
- ✍ Assimilate clear text passwords. Collect passwords that can easily be read by any user on the network

- Deploy Snooping Agents – placing agents on network. Test and record current state of affairs inside the network through automated processes controlled by agents to analyze network devices and services offered;
- Identifying security lapses on the network based on the findings. Define the factor of risk associated with each lapse
- Identify open ports
- Identify various services running on the workstation
- Map the above services to the application used and user requirements
- Identify unnecessary services
- Search and identify configuration errors, which form potential Vulnerability
- Identify the patch level on various applications
- Run vulnerability assessment tools on the servers to find the Vulnerability
- Conduct manual vulnerability research
- Evaluate the vulnerability based on the potential of exploitation and impact

Based on these above findings, the Administrator should be able to give proper recommendation for Redesigning the Security Architecture, Redesign the Information security policies.

13. Conclusion

Digital networked environment of libraries should have adequate security. Security relates to the techniques, policies and strategies used to protect the availability, confidentiality and integrity of electronic information. Security includes personal security, physical security, software security, network security, Internet security, etc. Libraries need to have policies and protection measures in order to protect their ICT resources. Libraries should also have well-established backup policies. Public terminals like OPAC should be protected from the internal system with adequate security measures. Libraries should take necessary steps to safeguard their digital resources and access tools from various threats like damage, misuse, mistake, theft, sabotage, etc. Library professionals must be assigned security related tasks. Due to the explosive growth of Internet with various tools, which provide unprecedented access to digital information and resources ongoing diligence is required to keep the digital networked environment of library secure.

14. References

1. Brandt, D Scott. (1998) Insecurity on the net. Computer in Libraries : 34-37.
2. Breeding, Marshall., 1997 Designing secure library networks. Library Hi Tech 57-58 (15:1-2) : 11-20
3. Camp, Jean. (1999) Web security and privacy: an American perspective. The Information Society, 15 : 249-256.
4. Gladney, Henry M. (1997) Safeguarding digital library contents and users. D-Lib Magazine <http://www.dlib.org/dlib/june97/ibm/06gladney.html.> (Accessed on 22 th December 2000).
5. Lavagnino, Merri Beth. (1997) System security in the networked library. Library Hi Tech 57-58 (15:1-2) : 9-10.

6. Morgan, Eric Lease. (1998) Access control in libraries. Computer in Libraries : 38-39.

7. Rasmussen, Audrey. (2002) Desperately seeking security. Information Technology 11(3) : 14-23.

8. Schuyler, Michoel. (2002) A serious look at system security. Computers in Libraries : 36-39.

9. Vince, Judith. (1996) Information security- protecting your assets. Aslib Proceedings 48 (4) : 109-115.

About Authors



Mr. Manoj Kumar K working with INFLIBNET Centre as Scientist-D (Computer Science) after having more than 10 years of wide experience in the entire gamut of Information Technology which include 5 and half years of service in Indian Institute of Management, Kozhikode(IIMK) as an Officer in Computer Centre. Involved in setting up of state-of-the-art IT infrastructure in IIMK from scratch, which comprise of a multi layered architecture with File servers, Database servers, Web server, FTP server, Email server and other high-end servers/computers. He holds BSc from University of Calicut and MCA from Government Engineering College, Thiruvananthapuram. He has worked in Coal India Ltd, Ranchi, Bihar as Technical Secretary to Director(Finance) and CEDTI, Calicut as Asst Engineer. He has involved in setting up wireless LAN based on WiFi in e-journals lab and he is looking after training programmes such as DSpace workshop, Network Administration etc. for different kind of professionals including Librarians at INFLIBNET Centre. He has contributed number of papers in seminars and conferences.
Email : manoj@inflibnet.ac.in



Mr. Mohamed Haneefa K is Senior Research Fellow in DLISc at Calicut University, India. He holds BSc, MLISc, PGDCA and awarded JRF from UGC in 1999. He worked with NIT Calicut, IISR Calicut and Calicut University. He has published few research papers in professional journals and participated in many national and international conferences. His current research interests include application of ICT in libraries, library consortia and digital library.
E-mail: haneefcalicut@yahoo.com

Transition in Information Services : A Digital Experience

Kalyani Accanoor

Abstract

Web based resources have become increasingly important to the Academic community where the thrust is on on-going Research, Time bound projects and Consultancy. With information being added at a tremendous rate on the Net it becomes difficult to find the required information easily. The expectations from the users are its services. From the library and the librarian's point of view the services should reflect the users quest for information. Services are the strength of any Library as they are no longer mere store houses. Reference service is one service that still retains its content, glory and importance. Reference service concerns with the internal and external resources through the Internet and other networks. These services are provided not only personally but also through electronic means. Most users go for a Google search still thinking how could the search strategy be better. Most times the answers lie in their Libraries without them knowing. A module E-Ref Desk on the lines of Ask a Librarian is designed to answer such queries.

Keywords : Digital Information Services, Information Services, Digital Libraries

0. Introduction

The mission of the libraries i.e., providing excellent information service to patrons has not changed. Libraries are now more known as Information centres while adapting the vagaries of changing technology. Technology has changed the way the librarians serve their users and this change will continue in future also. While continuing to provide many traditional information services, librarians are developing new skills and taking on new roles that are necessary to support technology-based services.

Information Technology has its impact in all areas of library work such as providing user access to digitized resources in the local environment and also remote-access. It is in this context that libraries must develop individual solutions that are appropriate to local circumstances. Many libraries have added electronic materials and services to the traditional items associated with physical space. These resources available via the Web include e-books, online catalogs, licensed databases, e-journals, research guides and finding aids, freely available Web resources and local digital collections.

The need to provide value-added library services In Academic Libraries is all the m ore to help in Academic and Research pursuits. The user expectations have increased with the proliferation of electronic sources. And this has made them increasingly important in providing reference service to library users. Librarians have to make these information resources accessible on the campus and also link the libraries to other networks away from the campus.

Libraries have to redesign their services on context rather than form and delivered to the users. More so as the Information is available electronically through networks and consortia licensing agreements 24 hours a day. In fact Technological developments and the impact of Information Technology in libraries have led to the emergence of new services.

Reference services also known as Information Services are an important standard service in library and information centres.

Traditionally it is a one to one service with the user and reference librarian. The user is helped by the sources available to meet his needs. Libraries are a great place to discover answers to many questions. For most people, the best place to start looking is the library of the organization where one is associated with, or people do go to well-maintained public Libraries for consulting and referencing.

Information services are receiving more attention as librarians have taken a proactive role to make the users know what the activities and holdings of the Library are, all the more due to the vast and varied electronic information resources available now. In fact they are venturing out of the library, working with the faculty, computing facilities, telecommunications and campus wide activities. The new learning technologies have provided librarians with greater scope for a larger role in distant education programs also. In fact librarians should provide guidance to remote users to enable them to use information resources.

Many college and university libraries have a general reference desk staffed by a librarian.

However it has been observed that many Libraries abroad have a general reference desk, full-time reference librarians with subject specialties and off-desk responsibilities such as collection development and instruction, and a number of part-time librarians who work reference desk hours only. This type of an environment has to be cultivated in Academic Libraries where the thrust is on imparting education and pursuing Research and Consultancy.

Since the Web and digital libraries are meant for providing direct access to information sources and services without much help of human intermediaries, there is an ongoing thought process as to whether the reference services in digital environment will survive. The present day digital libraries focus more on access, retrieval of digital information, and also provide a number of services such as TOC, Alerts etc., It also includes the online public access catalog, electronic databases (both indexes and full-text databases), and the Internet.

Instruction in the use of library is normally the job of the libraries reference desk, or is available individually as users require the information. The users do get confused by the varied amount of information and so many other formats. There is a need for user-centered method of delivering Reference and Information services in this age when the digital media is gradually replacing the print media. When Library Automation started a few years ago it was felt that computers and related programs would replace the human touch in Libraries. But, this advent of computers and the growth in the electronic information resources by leaps and bounds has made human intervention more necessary than ever. Library Automation has not replaced traditional services, but rather run parallel, in fact requiring additional expenditure and skills. The traditional work of reference librarians has been greatly influenced by the access to electronic publications on the World Wide Web. So much so that the Reference librarians are creating HTML documents that provide access to Web and other electronic resources. The limitations of the physical library timings are extended for 24-hour access.

1. Functions of Reference Service

Samuel Swett Green better known as the father of Reference Services laid down four functions for Reference Librarian.

They are:

- ✍ Instructing the reader in the ways of the library
- ✍ Assisting the reader with his queries

- Aiding the reader in the selection of good works, and
- Promoting the library within the community.

Even today these four functions remain the core of reference service.

2. Present Scenario

Libraries of the present days have undergone a massive facelift although the above functions have remained the same such as

- Organizing print and electronic resources
- Directing library users to resources within the library
- Assisting users with locating the best sources of information
- Marketing reference services and resources
- Serving as a public relations representative
- Online searching
- Professional activities for professional development and growth.
- Referral process - forward the enquiry or provide the user with live links to authoritative web sites.

In addition to the print media librarians have now incorporated technology by way of PC's telephone, photocopier, microfilm, fax machine, television, printer, modem, disks, CD-ROMs, scanners, telecommunications, and the Internet. Most Libraries have now become hybrid Libraries. It is needless to say that even with all this the functions of Green holds good.

3. Background

Distance education programs and the growth of Departments, Schools, Centres increase in the student intake, at IITB have increased the demands on the Library and Information Services. The size of the electronic holdings have increased mainly because of the expanse of Digitized information sources through the Consortia efforts and also by exploring the Internet for related free digital resources. The emphasis is also on computer based learning which encourages the support of teaching and research by non-traditional means and beyond the walls of the Library.

Adapting to the changing digital environment has required the Library to be innovative, flexible and imaginative in their internal organization and in their relationship with users.

Services and Facilities in making information available directly to end user, to search in an easy manner and also to know what is available within the Library has been an ongoing activity. More emphasis is given to cater to users requirements and their expectations.

- The Library has a Reference & Information Section and renders traditional Short range and
- Long-range reference queries.
- Short Range such as Directional,
- Long Range such as Literature Search and other information through Machine Readables, CD-Rom's, E-journals, Online and Internet.

4. Scope / Purpose

This paper looks into these aspects and is providing personalised information services to,

- Educate users to new resources available freely on the net, or procured by the Library.
- Continue updating information for the existing resources.
- Attend to Comments and requests by faculty and students.
- Observe the User behavior & their changing interests.
- Find answers from sources within the Library and also those, which are not easily found.

5. Digital Collections at IIT, Mumbai

While the Library has a good collection of standard Reference Tools in the print media

Information Technology has taken roots and over the years the Library has modernized to a vast extent. All in-house operations are computerized and the regular features are the OPAC, E- journals consisting of current, backfiles and perpetual access, ETD, CD-ROM Databases, Audio-Video collection, Access to Internet etc.,

6. Digital Information Services

6.1 Intranet solutions

An intranet is many things together i.e., the network, the web servers, browsers and the databases that supply information to the server. Basically it is an information delivery system designed for use within an organization. It is the latest technology that offers perfect solutions to the growing demands in the electronic environment. Some of the applications are,

6.2 OPAC

- Database containing information that the users can consult before asking for human intervention. E.g. - In this the Library holdings is already in place along with the lent, claims, fines details of the varied print documents available in the Library. The holdings also contain data of material other than books such as Journals available in print/electronic/perpetual access or as backfiles.
- Adding links to web resources especially those available freely e.g. PubMedCentral has made available the full-text of the complete volumes of the 'Bulletin of the Medical Library Association' now published as 'Journal of the Medical Library Association' online. The archive begins with Vol.1, 1911. The full text articles are available in .pdf format.
- Information Gateways, pathfinders, help to organize the reading material from within the Library as well as compiling the relevant resources from the web and make it easier for users to find the information all at one place. It includes www gateways, online reference and educational resources. Gateways in fact serve as a ready reference tool. Similarly pathfinders that give the details of subject holdings in the Library in a nutshell helps the freshers immensely.

6.3 E-mail Reference

A. Instant messages are circulated to the user community such as trial access to an e-product.

B. Sending pages to end-users regarding digital resources. E.g. .SD top 10 articles cited The top - 10 of most downloaded articles in your area of interest.

C. Latest additions of the resources added such as books, Journal issues, standards etc.,

D. Sending customized news to end-users e.g. Useful articles picked up from popular journals subscribed by the Library and would be of interest to many users.

E. ILL through a web form. ILL and Document Delivery services are a natural extension of the references process. Reference services provide the user with instruction in the user of bibliographic tools and database searches. Since no one library is likely to own all of the material cited in a library database or index the need to obtain materials from other Libraries is a process, which has to be done by the reference personal that identified them. ILL-uses many of the same databases to verify bibliographical details before submitting requests for loans and copies.

F. Get material matching User's profiles such as TOC's, Alerts

G. Any query which could be answered within library resources or needing outside referencing.

H. Citation Alerts of faculty's latest articles. The need for academic research staff to quote their publications from refereed journals over a period of years along with the impact factor and citation analysis is rising. This involves the help, advice and compiling information by the library staff.

6.4 On chat and by telephone

One of the most efficient reference services is to know about the user's problems and demonstrate solutions for easy access. E.g. To make a search on IEL for any IEEE Transactions one need to search under the subject like Antennas and Propagation, IEEE Transactions and not the other way round. Other problems like connectivity, Links to e journals not working, some figures, and graphics are blurred, such queries are being solved by chat messages or by telephone.

6.5 E-Ref Desk System

In addition to this a help desk system, E-Ref Desk is designed to answer queries by electronic means. The emphasis is on the local resources that would otherwise be unknown to the users and thus lose their significance.

This module is designed to take care of such queries that do not come under the purview of the above categories. While the Library tries to see that all queries would be answered with the help of modern technologies, in addition to the traditional jobs such as assigning Subject Headings there will still exist some queries that need extra effort on the part of reference staff.

Say for query on Prevention of Food and Adulteration Act there were no books on it, but a full chapter was found in Manual of Central Acts & Important Rules. Maybe the subject heading may not have identified it, as the item was very small. Hence the question has to be assigned a serial or a Docket no. and referred to the classifier or, the reference staff should go out of the way in search of the pertinent information. With the answer ready the user is given specific details as to where he should look up. In simple words this type of query would undergo the process as below.

Query: Any book/journal article/material on Prevention of Food and Adulteration Act —— Assign a Docket No.—— Pass the query to Classifier / Reference staff. ———Get details and post back to the person who asked for it. If not post the query to other Sections to get a satisfying answer. If the answer is satisfactory and to the point the Docket is closed. If not the circle continues again.

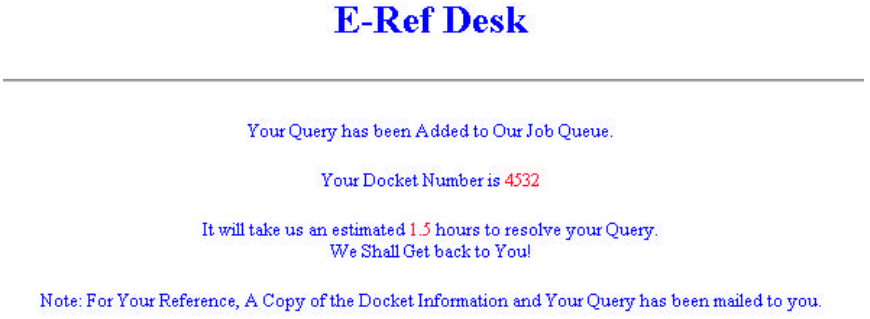
In a schematic presentation, In this E-Ref Desk, User can file in queries regarding Book Acquisitions, Technical Processing, Reference, Journals, and Pamphlets, etc.,

User writes his name and contact information in the form and selects a category that relates to his query. User has to describe and prioritize his query as either Low, Normal, High, or Urgent.

User can now add the query to the E-Ref Desk by clicking Add Job!.



Once the Job is added, User is given a Docket Number. A separate screen displays.



This Docket number refers to the query that the user has submitted and the user can use this docket number for any future communication relating to this particular query. Depending on the priority set and the number of jobs presently in Queue to resolve, E-Ref Desk will also display an estimated time that the department will take to resolve the added job.

E-Ref Desk also sends a mail to the user that contains:

- Information added by the user
- Docket Number
- Estimated Time

at the email address specified earlier.

E-Ref Desk now sets this docket to the designated section in the department. The designated section gets a docket number depending on the category chosen by the user.

So, With Category set to Reference, the docket would be assigned to reference section and added to that sections Job Queue.

Once the designated section resolves the query, it sends a communication to the user through mail, instant messenger, or phone.

The Communication now includes:

- Docket Number
- Answer to the Query
- Section Head Contact Information

Once a query is resolved and the communication sent to the user, the Docket is closed. If User is satisfied with the answer, he may simply carry on with his work. But if he is still not satisfied by the answer, he can file in another job at E-Ref Desk, with reference to the previous Docket. He may also contact the Section Head for further clarifications on his query.

Since all Dockets are categorized, queries are constantly tracked and statistics maintained by E-Ref Desk. The Library can now generate a report periodically to analyze the queries encountered. Hence, processes can thus be adopted for those queries that are encountered frequently.

7. Enquiry Services

Before sending or asking any question, one needs to check the following important information.

Have you:

- Checked our FAQ to see if you can find the answer there.
- Searched our OPAC to find the material you want.
- Checked the HOMEPAGE for all the options to know whether the answer expected might be there.

Types of questions expected

While going through some of the e-mail reference questions received, certain familiar questions arose such as

Library Details /Finding Books /Finding Articles /Finding text books/Reference Tools / Finding Other Research Materials/ Citing Resources /Getting Help Even with library automation one can now easily answer the following

- Provide the answers to questions on an enormous range of topics and in-house functions.
- Refer to other possible sources of information.
- Procure Books and Journal articles from other Libraries.
- Provide photocopies of articles to other Libraries on cost and reciprocal basis.
- Check with other Libraries to send our staff and students for referencing.
- Help users with the Literature Survey.
- Help in any area of the Library such as Issue /Claims etc.,
- Locate journal articles on a topic
- Search if the Library owns a document (periodical, theses, standard etc.,) if not where is it available
- Find the full-text of an article online.

8. Conclusion

In the library of the future customer service will remain as the primary objective. It is said that some believe that Reference is an activity suited to the paper collection of the past. In other words to use Reference Services users had to come to the section. It is a building-centerd old style make them come to us model. Librarians should begin to evaluate how consumers need information and tailor information sources to those needs. Librarians have to give the information to the user and not the other way round.

One may mention Dr.S.R.Ranganathan's five laws that still stand the test of changing times and changing media. Also Kuhlthan who identifies five levels of services,

Level 1 – Librarian / library is the organizer of the material

Level 2 – Librarian is the locator or ready reference

Level 3 – Librarian is the identifier, helps user identify tools for the information need

Level 4 – Librarian is the advisor.

Level 5 – Librarian is the counselor

These 5 Levels of service remain valid, even as users have less contact with traditional library support.

This study as stated above by various examples shows that users ask similar questions whether in person or via an e-mail reference service. Academic librarians especially those in charge of Information desks should be well prepared to answer a full range of questions – basic questions to the difficult ones.

Limiting digital reference service to "ready reference" questions alone does not adequately meet users' needs and may not even be understood by them. The virtual information desk services will become an integrated system to include live chat and telephone sessions for better functioning.

More newer technologies will come into existence. One does not know how long the digital environment would remain as it is, but may change into much more sophistication. Inspite of all the hype and use of Information Technology, the human factor is very essential. Libraries may change to Cybraries and Librarians to Cybrarians but the human, personal element will be present always.To quote Samuel Swett Green for promoting the idea of reference service, "The more freely a librarian mingles with the readers, and the greater the amount of assistance he renders them, the more intense does the conviction of citizens, also, become that the library is a useful institution," is absolutely true.

9. References

1. Biddiscombie Richard (1996) The end-user revolution, CD-ROM, Internet and the changing role of the information professional., Lib.Ass.Rec: 202
2. Choate Jennifer (Mar 1997) Microsoft librarians –Training for the 21 st century. Information Outlook Vol.1(3):27-29
3. Chowdhury Gobinda G (2002) Digital Libraries & Reference Services: present and future.Journal of Documentation.Vol.58(3):258-283
4. Courtney L Young; Karen R Diaz (1999) E-reference: incorporating electronic publications into reference. Library Hi Tech. Vol.17(1):55 - 62
5. Jobe Margaret M and Grealy Deborah S (2000) The role of libraries in providing curricular support and curriculum integration for distance learning courses. Advances in librarianship V.23:.239 – 267
6. Marshall Joanne Gard (2003) Influencing our professional practice by putting our knowledge to work. Information Outlook Vol.7(1):41-44
7. Moyo, Lesley M(2002) Reference anytime anywhere: towards virtual reference services at Penn State. The Electronic Library Vol 20(1):22-28
8. Welch, Jeanie M(1999) Laser lights or dim bulbs? Evaluating reference librarians' use of electronic sources. Reference Services Review Vol.27(1):73-77

About Author



Ms. Kalyani Accanoor is currently working as Assistant Librarian (SG) in the Central Library, IIT Bombay. Prior to this she has also worked at IISc and Times of India Library. She is the recipient of the ILA / Dr. C. D. Sharma award for the best presentation and best author at the 44th ILA conference Hyderabad, February 1999. She has given talks, and published papers in various Seminars/Conferences.
Email : accanoork@iitb.ac.in

Indian Academia on Copyright and IPR issues of Electronic Theses and Dissertations

J K Vijayakumar T A V Murthy
M T M Khan

Abstract

The idea of E- Theses and Dissertations (ETD) is coming up in International scenario, which can be easily located, readily accessible and delivered over the web. This paper analyzing the opinions of selected Ph D Researchers and Guides from selected Indian Universities on Copyright and IPR issues related to ETDs. On the basis of the output, the paper suggests that Universities can start collecting e-format of theses, creating a digital archive for easy access. But in terms of access, still only a minimum majority is favoring online global access to Indian research. This may be because of Copy Right Issues, Chances of Plagiarism and Poor Quality in Research, which may be solved through policy frameworks and enhancing standards through national agencies like UGC at national level.

Keywords : Copyrights, IPR, E-theses

0. Introduction

Digital libraries of electronic theses and dissertations (ETDs) offer an alternative to the waste of valuable academic scholarship in the form of Theses and Dissertations (TDs) and offer researchers and University Libraries in India opportunities to explore the possibilities electronic publishing trend in academic sector. The emergence of UGC Infonet, the aspiring and dream project of University Grants Commission, which also aims at Content Creation by Indian Academic Sector, will definitely boost this idea. Copyright issues related to university research output in the form of theses and dissertations, are discussed in great deal already with concerns of the researchers and academia. Theses and dissertations have long been regarded as the basis of university research. They represent the outcome of focused and extensive studies, involving intellectual labor over several years. Rapidly developing networking and digital library technologies are the reasons for ETDs (Electronic Theses and Dissertations) gaining momentum in university campuses worldwide. In recent years, many Indian universities have realized the importance of this new kind of digital resource and some local ETD programs have been carried out to increase availability of theses and dissertations. The adoption of electronic theses and dissertations in a university will require a number of alterations to the existing copyright agreements between the rights holders, usually the primary researcher, and those responsible for theses management, usually the university library. Before entering into any agreements it is critical to determine who actually owns the copyright to the work as there are a number of key stakeholders in the production of theses, including the author, host institution and perhaps the funding bodies. Copyright of ETDs has to be discussed separately in Indian condition and this paper tries to describe the ETD as a new concept in University libraries, its implementation and the opinions of Indian Academia about their acceptance. The opinions collected through a national level sample survey conducted among selected Ph D Scholars and Research Guides of Indian Universities receives funding from UGC, are analyzed along with a description of copyright and IPR issues related to theses and dissertations.

1. Electronic Theses and Dissertation

There is some variation between countries in the use of terminology e.g. some universities refer to doctoral theses and some to doctoral dissertations. The term ETD accommodates these differences and is becoming used internationally. It may be an electronic version of a printed thesis where the old document that has been scanned and converted into PDF. Alternatively it could be a recently completed piece of work produced and archived in Word or produced in Word and converted into PDF in order to be made available on the Web. ETDs allow more adventurous students to express their research results in creative and flexible ways that would not be possible if they were limited to paper based output. 'Born digital' theses may include audio and visual material and may not even be in a traditional linear format.

2. IPR and Copyright for ETDS

Creating conditions that favor the production of useful ideas introduces one of the most complicated matters associated with ETDs—that of intellectual property protections for authors. Any piece of Information will survive only through high accessibility and continued use, where new generations of scholars must access and incorporate the work of others into their own. They have to continually reproduce and develop the ideas society needs. Improved access to TDs through ETDs can contribute greatly to the dissemination and preservation of university research, but at the same time, intellectual property protection for researchers is also very important. Copyright protects the labor, skill and judgment that someone - author, artist or some other creator - expends in the creation of an original piece of work, whether it be a literary work, a piece of music, a multimedia programme, a Web page, a painting, a photograph, a TV programme, or whatever. The copyright issue involves two components: Protecting the information/work produced as part of the research program; and Granting license to University or to any ETD Programme to make the work available for use. This also includes obtaining permission to use parts of the work that have already been published in other sources.

In the world of scholarly publishing, authors create and intellectual output, which will be marketed or distributed by the publishers and the libraries will collect, preserve, organise and disseminate the information. The networking world really creates concerns on Copyright of digital documents, which can be easily downloaded and reused. In exceptional circumstances, where the thesis research has been particularly innovative, and there is potential for commercial exploitation, it may be desirable for the author to apply for a patent. A patent application may be successful only if the invention has never been made public in any way before the date on which an application for a patent is filed. It also must involve an inventive step and be capable of industrial application. These issues concerns the ETD promoters worldwide to take the necessary steps to safe guard the copyright issues for the real scholarly works done by the research scholars.

3. Literature Survey

In a recent survey conducted in China by CALIS, they let students and their advisors to determine the online accessibility of their ETDs. The majority of students allowed their ETDs to be viewable online soon after submission, while the others elect to protect their ETDs for a certain period of time. For example, of the 2,340 ETDs submitted by students of Shanghai's Jiao Tong University in 2003, 69 percent of wanted their ETDs be accessible online immediately, 22 percent gave permission for their ETDs be viewable after one to three years, seven percent agreed to their work being made available after four to five years, and only 2 percent wanted their ETDs to be protected for more than five years, or not to be accessed at all. The fact that most students give permission for their ETDs to be viewable online within five years of submission will greatly increase the use and accessibility of ETDs.

The Vidyanidhi Digital Library Project at Mysore University worked for developing a policy frameworks for creating an archive of theses and dissertations and identified most of the copyright issues related to scholarly communication apply equally to the world of theses and dissertations. It is argued that scholarly work should be freed from Copyright jargon, because the university researches usually supported by public funds and based on Collaboration. The Lack of formal publication practices result in the lost of scholarship and intellectual heritage. The tradition of a doctoral student defending the thesis in public implies that doctoral research works should be made publicly accessible.

Questions about intellectual property are often tied to concerns about whether electronic publication of a thesis or dissertation constitutes prior publication with respect to future efforts to publish student research as a book or a journal article. Much confusion surrounds these discussions, and because of Web technology and the publication opportunities it affords are so new, answers to questions that arise do not often appear simple or clear cut. A survey of Faculty and students in Virginia Tech University reveals that though it is obviously still an unresolved issue, an ETD would not preclude book or journal publication of research should be encouraging to students and their faculty advisors who are working in an increasingly electronic environment.

But, in another survey available at http://lumiere.lib.vt.edu/surveys/results/, 53 publishers were asked the following question “According to the editorial policy governing the journal(s) identified, under which circumstances would a manuscript derived from a WEB-BASED dissertation be considered for publication?” 25 publishers (47.1 percentage) welcomed this idea where 10 publishers (18.87%) suggested that it should be considered on an individual basis.

In a professional paradigm where the publication of original work is the coin of the realm, students and faculty advisors are naturally concerned about providing open access to dissertations that may or may not count as prior publication or that contain information considered sensitive in fields where competition for original credit is high. However, in a recent survey of journal editors and publishers, 82% said that an online thesis or dissertation widely available through a Web-based archive would not be considered prior publication according to their journals’ existing policies; only 4% said that an online thesis or dissertation with access limited to campus or institution would be considered prior publication. Yet, 40% of graduate students who publish ETDs are advised by faculty to restrict access in order to protect their professional interests. Such restricted access threatens to undermine the very purpose for which the ETD Networks like NDLTD was created.

In this context it would be appropriate to take the opinion of Indian academia, where the idea of electronic theses are getting much attention day by day. In the context of great difficulties facing by us for getting access to Indian University research in the form of Ph D Theses, the scenario of developed countries where they are able to procure the theses documents through variety of modes, will have to treat as special case.

4. Research Methodology

As part of the doctoral research work undertaken for proposing a model Electronic Theses and Dissertation for Indian Universities, a sample survey has been conducted at national level. The survey was focused on Ph D Research Scholars, Research Guides and Librarians of selected Indian Universities funded under UGC, and connected or getting connected to UGC Infonet Programme. Questionnaires were sent or distributed to the participants of INFLIBNET’s E-Resources awareness programmes consisting of Research Scholars, Guides and Librarians, who are familiar with latest IT developments taking place in Information and Communication. Separate questionnaires are also sent to University Librarian, but we are not taking that data for this article.

In the separate Questionnaire sent to Researchers and Guides, there were few questions related to their willingness to provide online access to their Ph D Thesis and copyright problems, which are being described below.

5. Data Analysis and Discussions

5.1 Research Scholars

163 Ph D Research Scholars were participated in this survey covering 26 Universities across the country. There were Four Questions related to the scope of this article, ie. copyright practices of Doctoral Research, and their responses are analysed below.

Table : 1

Q. 9.	Are you willing to provide an electronic format (soft copy) of your Ph D Thesis to your University?	Yes	No
		140 (163) (85.89 %) Group-A	23 (163) (14.11%) Group-B

Out of 163 Ph D Research Scholars, 140 said Yes (85.89%) and only 23 (14.11%) are not ready to provide an electronic format of their Ph D Thesis to the University. We will identify them as Group-A (Ready to provide e-format to University) and Group-B (Not ready to provide e-format to University), as shown in the Table-1, for data analysis.

Table-2

Q.10	Do you support online full-text access to your Ph D Theses through a Digital Library?	Yes	No
		135 (163) (82.82%) Group-C	28 (163) (17.18%) Group-D

From Group A, 135 Scholars (82.82%) are said Yes and 28 Scholars (17.18%) including 5 from Group A and 23 from Group B, said No. Group C represents the Scholars ready to provide online access to their work and Group D represent those who are not willing.

Table-3

Q. 11	If Yes, what can be the access policy?From	On Library Intranet, for users coming to library	On Campus Intranet for your University Use Only	On internet for global access
		22 (135) (16.30%) Group -E	18 (135) (13.33%) Group-F	99 (135) (73.33%) Group-G

Out of Group C, 99 scholars are ready for providing global access to their work through Internet (Group G - 73.33%) and 18 Scholars are ready to provide access on Campus network for using inside the University (Group F - 13.33%) and other 22 scholars are ready to provide access only on Library Intranet (Group E - 16.33%).

Table-4

Q.12	If No, what are the : reasons	Copy Right Problems	Chances of Plagiarism	Other Reasons
	Out of Group B	23 (23)100%	10 (23)(43.48%)	Not relevant for
	Out of Total Scholars	23(163)(14.11%)	10 (163)(6.13%)	this article

All 23 Scholars from Group B, who are not ready to provide electronic format of their Ph D Thesis, identified "Copy Right Problem" as reason for their unwillingness. 10 Scholars from Group B also identified "Chances of Plagiarism" as another reason. Only 2 scholars responded that they are not interested for a wider access to their Thesis. Other responses are not relevant to the topic of this article, and it was very few responds or nil response for other reasons.

There is an observation from one scholar that electronic copies should only be in readable format with controlled access, and not to be downloaded or printed. Another Researcher urged upon some kind of restrictions to prevent the duplication (Plagiarism). But it is very clear that only 14.11 percentage of the total Researchers surveyed identified Copyright Problems as a threat for providing online access to full text theses, and only 6.13 % feared about Chances of Plagiarism in a online environment. Scholars from Group E and Group F, who are not ready to provide Global Access to their thesis, have to be surveyed again to understand the reasons of their unwillingness.

5.2 Research Guides

75 Ph D Research Guides were participated in this survey covering 25 Universities across the country. There were Four Questions related to the scope of this article, ie. copyright practices of Doctoral Research, and their responses are analysed below.

Table-5

Q. 9.	Do you support in obtaining an electronic format (soft copy) of Ph D Theses of your scholars, by the University?	Yes	No
		68 (75) (90.67%) Group-1	7 (75) (9.33%) Group-2

Out of 75 Ph D Research Guides, 68 said Yes (90.67%) and only 7 (9.33%) are not supporting the idea of obtaining an electronic copy of theirs Researcher's theses by the University. We will identify them as Group-1 (Supporting the provision of e-format to University) and Group-2 (Not supporting the provision of e-format to University), as shown in the Table-5, for data analysis.

Table-6

Q.10	Do you support online full-text access of them through a Digital Library?	Yes	No
		65/75 (86.67%) Group - 3	10/75 (13.33%) Group - 4

From Group 1, 65 Guides (86.67%) are said Yes and 10 Scholars (13.33%) including 3 from Group 1 and 7 from Group 2, said No. Group 3 represents the Scholars ready to provide online access to their work and Group 4 represent those who are not willing, in the Table – 6.

Table-7

Q. 11	If Yes, what can be the access policy?	On Library Intranet, for users coming to library	On Campus Intr-a net, for your University use only	On internet, for global access
		8 (65) (12.31%) Group-5	12 (65) (18.46%) Group-6	45 (65) 69.23%) Group-7

Out of Group 3, 45 Guides support the global access to their Researcher's work through Internet (Group 7 - 69.23%) and 12 Guides support access on Campus network for using inside the University (Group 6 - 18.46%) and other 8 Guides support access only on Library Intranet (Group 5 – 12.31%).

Table-8

Q.12	If No, what are the reasons:	Copy Right Problems	Chances of Plagiarism	Others
	Out of Group 2	8 (10)(80%)	10 (10)(100%)	Not relevant
	Out of Total Scholars	8(75)(10.67%)	10(75)(13.33%)	

All 10 Guides from Group 2, who are not ready to provide electronic format of their Ph D Thesis, identified Chances of Plagiarism as reason to their unwillingness. 8 Guides from Group 2 also identified as Copy Right problem as another reason. Other responses are not relevant to the topic of this article, and it was very few responds or nil response for other reasons.

There is a suggestion from three Guides to provide only title and abstract/synopsis to avoid Plagiarism. Another Guide suggested that full Text could be made available only after two years of award, to give sufficient time to the scholar for publishing papers or book. Quality, Standard and Plagiarism were the concerns of few Guides. Only 10.67 percentage of the total Guides surveyed identified Copyright Problems as a threat for providing online access to full text theses, and 13.33 % feared about Chances of Plagiarism in a online environment. Guides from Group 5 and Group 6, who are not supporting Global Access to the theses literature, have to be surveyed again to understand the reasons of their unwillingness.

6. Findings and Suggestions

An attempt has been made to go into the nitty-gritty's of copyright and IPR issues relating to Electronic Thesis and Digitization in Indian context. Even though Majority of participants in the survey expressed their willingness to provide electronic format of their Doctoral Theses, a considerable number of them expressed apprehensions to provide Global online Access (Groups 5,6,E and F). This indicates that the hosting of Indian intellectual content in electronic form for the Global access is still an issue among a considerable academic fraternity. Enough protections on copyright, digital management rights, repositories' rights and responsibilities, digital preservation, access and distribution, Metadata, legal responsibilities etc need to be thoroughly worked out for protecting the scholar's contributions. It was also felt that efforts are needed to improve the quality of content on par with international standards with uniform pattern so that retrieval, establishment of open archives and other related technologies could be plotted for the benefit of worldwide ETD efforts. It is therefore essential that bodies like UGC should evolve a regulatory policy mechanism in maintaining standards, quality, proper submission and publication practices for Doctoral research both in print and online environment.

7. Conclusion

Mean while, our universities and librarians must take up the challenge to preserve and make available the key intellectual product of their institutions to the world, and the Internet presents a wonderful opportunity for us to do so. It is a fact that many of our Universities do not have full Internet access, but it should not deter us from collecting electronic files along with print copies of theses and dissertations produced in their institutions. It will give an opportunity for popularizing the idea of Digital Library and E-Publishing in particular in respective Universities (Vijayakumar, Murthy and Khan, 2004). INFLIBNET has already hosted an online database of Theses containing around 1.6 lakhs of bibliographic records of Ph Ds submitted to Indian Universities. Full text of existing theses collection can also be made available by converting them in to digital form. Metadata can be centralized in a common database at the coordinator institution site where, ETDs will be accessed through a single-web gateway, ie, INFLIBNET under UGC-Infonet (Murthy, Cholin and Vijayakumar, 2004). The authors hope that the issues and fear will go away in due course of time, and India will definitely contribute to the ongoing ETD efforts at International level.

8. References

1. Andrew, Theo (2004). Intellectual Property and Electronic Theses. http://www.jisclegal.ac.uk/publications/ethesesandrew.htm, 22 September 2004
2. Edminster, Jude and Moxley, Joe (2002). Graduate Education and the Evolving genre of Electronic Theses and Dissertations, Computers and Composition, 19 (1), April 2002, Pp 89-104.)
3. http://lumiere.lib.vt.edu/surveys/results/
4. Jin, Yi (2004). The development of the China Networked Digital Library of Theses and Dissertations, Online Information Review, 28 (5), 2004, Pp. 367-370
5. Murthy, TAV, Vijayakumar, JK and Cholin, VS. UGC-INFLIBNET initiatives in e-journal consortia and digital library of doctoral theses for Indian universities. Paper accepted for National Conference on Digital Library and e-Theses (NCDLET 2005) held during January 7-8, 2005 at Jadavpur University, Kolkata.
6. Oppenheim, Charles (2004). Recent Changes to Copyright Law and the implications for FE and HE, http://www.jisclegal.ac.uk/publications/copyrightcoppenheim.htm, June 2004.
7. Seamans, Nancy H. Electronic theses and dissertations as prior publications: what the editors say, Library Hi Tech, 21 (1), 2003, Pp. 56-61
8. University of Pittsburgh ETD Website (2004) http://www.pitt.edu/AFShome/g/r/graduate/public/html/etd/copyright.html
9. Urs, Shalini (2004). Copyright, academic research and libraries: balancing the rights of stakeholders in the digital age, Program: electronic library and information systems, 38 (3), 2004, pp. 201-207
10. Vijayakumar, JK, Murthy, TAV and Khan, MTM (2004). Accessing Indian University Research Literature: importance of ETDs in the verge of UGC-Infonet. In Chandra, Harish; Pichappan, P and Kundra, Ramesh; ed. Conference Papers of the 22nd Annual Convention and Conference of SIS (SIS-2004), Chennai, India. 22-23 January 2004: 53-57.
11. Vijayakumar, JK, Murthy, TAV and Khan, MTM (2004). Electronic Theses and Dissertations for Indian Universities: A Framework. In Murthy, TAV and others, ed. Conference papers of 2 nd PLANNER-2004, Imphal, India, 4-5 December, 2004: 65-70

About Authors



Mr. J. K. Vijayakumar holds B Sc in Mathematics from University of Kerala, MLISc with First Rank from Annamalai University and pursuing Ph D in the area of ETD from Bundelkhand University. He is working with INFLIBNET since 1998. He was selected for IFLA Travel Grant in 2004 and participated in IFLA Conference-2004 in Buenos Aires, Argentina and IFLA Pre-conference in Sao Paulo, Brazil during August 2004. He was also selected as IFLA/OCLC Fellow in 2002 and undergone specialised training programme based at OCLC in United States of America and visited so many Libraries in USA. He is the recipient of InfoShare Membership Award of ASIS&T, USA in 2003. He co-edited International CALIBER 2004 proceedings and widely published 25 papers at international and national level. He is life member of SIS, ILA, IASLIC, KLA, GGSS, ISOC and actively involved in various international and Indian library forums. His areas of interest are Digital Libraries, ETDs, Digital Repositories, Database Management, Internet Applications, Online Resources and Training etc.
E-Mail : vijay@inflibnet.ac.in



Dr. T. A. V. Murthy is the first LIS professional to become the Director of INFLIBNET with the status of a Central University Vice Chancellor and elected President of Society for Information Science. He is also the council member of IASLIC and Secretary of Ahmedabad Library Network. He holds B Sc, M L I Sc, M S L S (USA) and Ph.D ffrom Karnatak University and carries with him a rich experience and expertise of having worked in managerial level at a number of libraries in many prestigious institutions in India including National Library, IGNCA, IARI, University of Hyderabad, ASC, CIEFL etc apart from Catholic University and Casewestern Reserve University in United States. His highly noticeable contributions include KALANIDHI at IGNCA, Digital Laboratory at CIEFL, UGC Infonet E-Journal Consortium etc. He has been associated with number of universities in the country as visiting professor and has guided number of Ph.Ds and actively associated with the national and international professional associations, expert committees and has published good number of research papers and books. He visited several countries on professional activities and organized several national and international conferences and programmes. He is the recipient of SIS Fellowship, SATKAL Librarian Award and SALIS Harish Chandra Sushil Chandra Best Librarian Award.
E-mail : tav@inflibnet.ac.in

Dr. M. T. M. Khan is the Professor & Head in Institute of Library and Information Science, Bundelkhand University, Jhansi. He holds M.Lib.Sc. Degree from Aligarh Muslim University, Aligarh, he earned his Ph.D from Jiwaji University, Gwaliour amidst his academic process. He has rich teaching and guiding experience of 25 years and guided several scholars for research work. He widely published in the form of books, conference papers and journal articles and organized conferences and training programs. He is actively involved in professional associations and organized several conferences and programmes.

Use of Information Sources in Digital Environment : A Case Study

D Rajeswari

Abstract

Rapid advances in information processing, storage and communication technologies have revolutionized a role of worldwide libraries in disseminating information services to their users. Libraries are consolidating their positions, building digital collections, redesigning their services and information products to add value to their services in order to satisfy changing information needs of users. In this research paper the author covers the profile of Sri Padmavathi Mahila Visvavidyalayam, objectives of the study, use of electronic resources by the faculty, research scholars and students, suggestions and findings. Further, studies and research are suggested in application and implications of e-classrooms, e-teaching and e-learning should be the source of knowledge in future.

Keywords : E-Resources, Information Services

0. Introduction

The University plays a significant role in the development of the society. The main function of any University is to seek and cultivate new knowledge by way of Research and extend higher education to the youth, to encourage academic investigations into the problems of the society and for advancement of civilization. The university library plays an important role in the achievement of this objective. Electronic sources plays a vital and viable role to cater to the needs of research and faculty in the process of advancement of society in the present environment.

1. Profile of Mahila University Library

Sri Padmavathi Mahila University is the second Women's University in India, established in Andhra Pradesh in 1983 by the Late Chief Minister N.T. Ramarao. Under the UGC inflibnet programme University library has computerized all of its operations using SOUL software. Through INFONET programme the University gets e-journals(550 from Springer link and 650 from Kluwer online). The library has nearly 51,500 books, 230 periodicals, 50 CDs on books, 300 floppies on indexing and abstracting journals .

2. Objectives

The aim of the present study is to make "An analytical Study of the use of Electronic Resources and Services by Faculty, Research Scholars and Students of SPMUL, Tirupati". The main objectives of the study are

- to ascertain the requirements of the users;
- to identity the various channels of electronic sources through which information is accessed by users of SPMUL;
- to assess the purpose of using the Internet and OPAC.;
- to Analyze how the users are benefited from the INFLIBNET programme;
- to identify the problems and difficulties faced by the users in searching the information through Electronic Resources and
- to seek the suggestions from the users for the overall improvement of the library.

3. Methodology

The present study has been made by surveying the different user groups of the Sri Padmavathi Mahila University Library using the questionnaire method. The questionnaire has been distributed to the users following stratified sampling techniques. The respondents are stratified into three categories belonging the Teaching Staff, Research Scholars and P.G.Students, The sample respondents chosen for the study consists of 36 Teaching Staff, 42 Research Scholars and 58 P.G Students. In this study, the mode of collection of data, its presentation, and analysis and interpretation are presented in Table 1.

Table – 1
Category of Users

Sl.No.	Users Category	Users		Response	
		QD	%	QR	%
01	Teaching Staff	36	26.5	28	25
02	Research Scholars	42	30.9	38	34
03	P.G.Students	58	42.6	46	41
		136	100	112	100

QD : Questionnaire Distribution QR : Questionnaire Response

Table 1 Shows that of the total 136 questionnaire distributed, 112 users returned the duly filled in questionnaire 82 percent response. The respondents from P.G. Students are high (41%) followed by Research Scholars and Teaching Staff respectively. Most of the user community of the SPMUL covered under the survey are Computer and Internet Literate, therefore questionnaire was distributed only among computer/internet literate persons.

Table-2
Category of Users Subject wise

No.	Subject	Teacher		Research		PG. Student		Total	
		No.	%	No.	%	No.	%	No	%
1.	Education	5	17.86	3	7.89	4	8.7	12	10.71
2	English	3	10.71	4	10.52	4	8.7	11	9.82
3.	Social Work	3	10.71	5	13.16	5	10.9	13	11.6
4.	Women's Studies	4	14.29	5	13.16	4	8.7	13	11.6
5.	Computer Science	4	14.29	5	13.16	11	23.9	19	16.07
6.	Mathematics	3	10.71	6	15.79	9	19.6	18	16.07
7	Microbiology	4	14.29	6	15.79	6	13.0	15	13.39
8.	Sericulture	2	7.14	4	10.52	3	6.5	11	9.82
	Total	28		38		46	100%	112	100%

Table – 2 Reveals Respondents discipline wise and their percentage.

Table – 3.
Use of Electronic Rseources

S.No.	E-Resources	Teacher		Research Scholar		PG. Students	
		No.	%	No.	%	No.	%
1.	E-mail	26	24.5	35	23.2	38	25.2
2.	OPAC	22	20.8	34	22.5	46	30.5
3.	Internet Access	24	22.6	35	23.5	40	26.5
4.	Books Access	16	15.1	21	13.4	15	9.9
5.	Access to E-Journals	18	17.0	26	17.2	12	7.9
	Total	106	100	151	100	151	100

It is clearly seen that Internet Access, E-mail and OPAC are used by the almost all staff users. It is also found that teaching staff and Research Scholars are very much benefited from access to e-journals through Infonet. E-books and e-journals are found to be less used by P.G. Students. This may be due to e- books and e-journals are mostly used for reference by Research Scholars and semester system does not give scope for spending lot of time in the library. The user rate can be increased if Internet is established in the hostel premises.

Table – 4.
Purpose of Use of Internet

User	Email	%	Academic Website	%	E-Journal	%
Teachers	26	26.3	28	29.8	20	34.5
Research	35	35.5	34	36.2	24	41.4
Scholars						
PG Students	38	38.4	32	34.0	14	24.1
Total	99	100	94	100	58	100

Internet can be used for accessing the electronic resources, such as bibliography records, full text electronic journals with images, links to local and remote indices. It is useful as search utility to access information stored on millions of computers world wide. It also facilitates information book reviews that could enhance research and journal publications. It disseminates all kinds of data and information by keeping touch with the latest developments in various disciplines.

The survey further revealed that Internet service is used for different purposes by different groups of users. It is also observed that E-mail dominates over other purposes for which they use Internet-Use of e-journals through web is higher among the research scholars for their research work.

Table – 5.
INFLIBNET Services Meeting the Information needsof users.

Users	**To Great Extent Extent**	**%**	**To Some Extent**	**%**	**To title Extent**	**%**
Teachers	20	21.3	6	50	2	33.3
Research Scholars	34	36.2	2	16.7	2	33.3
Students	40	42.5	4	33.3	2	33.3
Total	94	100	12	100	6	100

Table – 5 reveals that most of the users one benefited from INFLIBNET Services available in SPMU Library. Thousands of check e-journals are going to be made available for seamless access via campus network very shortly through the UGC Infonet under the INFLIBNET Programme. Now of e-journals 550 from Springer Link and 650 from kluwer Online by users i.e Teaching Staff and Research Scholars under the UGC Infonet Programme.

Table – 6.
Use of E-Sources Subject wise.

S.No.	**Subject**	**E-mail**		**Internet**		**OPAC**		**CD Ram Data Bases**	
		No	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
1.	Education	10	10.3	10	10.6	12	10.9	—	
2.	English	10	10.3	8	8.5	11	10	—	
3.	Social Work	11	11.3	12	12.6	13	11.9	1	5.6
4.	Women Studies	12	12.4	10	10.6	12	10.9	2	11.1
5.	Computer Science	17	17.5	15	16	19	17.3	5	27.8
6.	Mathematics	16	16.5	14	14.9	18	16.4	4	22.2
7.	Microbiology	13	13.4	15	16	14	12.8	4	22.2
8.	Sericulture	8	8.2	10	10.6	11	10	2	11.1
	Total	97	100%	94	100%	110	100%	18	100%

Majority of the users are utilizing the OPAC system in library. Respondents from all the subjects are using Internet facility for their academic and research purposes. C.D ROM database is being used only rarely by staff and student when analyzed subject-wise

Table – 7
Use of E-Resources by Area of Research

S.No.	Subject	E-mail		Internet		OPAC	
		No.	%	No.	%	No.	%
1.	Child Labour and Parent-Child Relation	12	11.5	11	10.4	12	11.1
2.	Feminist Writings	11	10.6	10	9.4	11	10.2
3.	Gender & Development Women Empowerment	10	9.6	12	11.3	11	10.2
4.	Women in Higher Edu. Women and Health	12	11.5	11	10.4	12	11.1
5.	Graph Theory, Discrete Mathematics	18	17.3	17	16	18	16.7
6.	Speech Recognition Cryptography	17	16.3	19	17.9	18	16.7
7.	Mulberry Physiology	10	9.6	11	10.4	12	11.1
8.	Microbial Technology-Enzymology, Virology	14	13.5	15	14.2	14	13.0
		104	100	106	100	108	100

The distance is shortening due to the hi-tech development. The University has a larger percent of research scholars who are part timers and working in other places even outside the state. E-mail has enabled the Research guide and the research student to have easy interaction through E-mail. This can be seen from the Table-7. Though all the three OPAC, Internet and E-mail are being used OPAC is slightly on the higher side especially as it is used for bibliographical information.

Table-8
Information Source Provided in the Library

S.No.	Item	Teaching Staff		Research Scholars		P.G.Students	
		S	US	S	US	S	US
1.	Text books	26 (15.4)	2 (3.6)	38 (15.3)	0 (15.3)	46	0
2.	Ref. books	25 (14.8)	3 (5.5)	34 (13.7)	4 (7.1)	44 (14.7)	2 (2.9)
3.	Periodicals	22 (13.0)	6 (10.9)	32 (12.9)	6 (10.7)	36 (12)	10 (14.7)
4.	Bibliographical Services	24 (14.2)	4 (7.3)	35 (14.1)	3 (5.4)	42 (14)	4 (5.9)
5.	Photocopying	20 (11.8)	8 (14.5)	25 (10.1)	13 (23.2)	38 (12.7)	8 (11.8)

6.	Computer lending	16 (9.5)	12 (21.8)	34 (13.7)	4 (7.1)	46 (15.3)	0
7.	OPAC	26 (15.4)	2 (3.6)	38 (15.37)	0 (14.6)	46 (2.9)	2
8.	C.D. ROM Search	10 (5.9)	18 (32.7)	12 (4.8)	26 (46.4)	4 (1.3)	42 (61.8)
		169	55	248	56	300	68

98% and 91% of the users in the sample studied were highly satisfied with the sources available on textbooks and reference books respectively. The Online Public Access System is appreciated by 98% of the respondents. 95% of the P.G Students are satisfied about the computer lending service . 76% of the users are not aware of the CD ROM databases.

Opinions expressed by the respondents about the Soul OPAC in the Library

- Presently books/journals are traced or immediately without much strain and in certain new areas books are quickly located and identified.
- Lot of time and effort are saved
- Most of the respondents expressed that through OPAC users can access the records by name of the author, title, subject, publisher etc.
- Easy to use and search because, it has user friendly s/w.
- Online cataloguing is very useful to the users.
- Required books/journals can be picked in a short time.
- Status of the document can also be known immediately i.e available or issued.
- Most of the respondents expressed the process of computerization has helped them for their search.

Suggestions given by the respondents:

- In future digitalization of library is required
- Use of documents in the form of electronic media may be useful to Scholars and students.
- Most of the respondents suggested to conduct awareness programme about OPAC and also use of e-sources.
- Internet facility should be provided to more number of users in the library.
- If UGC infonet is connected i.e LAN then every staff and scholars are more benefited and services can be utilized from their departments itself and relevant material may be copied into CD or Floppy for leisurely reading at home or staff room.
- Space can be saved through storage of information in digital environment.
- Resource sharing is more feasible in a digital environment.

4. Conclusion

The role of library and its viability in the electronic publishing environment pose serious problems. There is no doubt that electronic resources are expanding rapidly. Electronic resources clearly allows for the rapid distribution of information at a reduced cost. Electronic or digital environment demands multiple websites for variety of information sources. The websites recommended for users must be current and easily accessible through internet. Today any academic library must equip digital collection on par with print media. Further the library should also provide browsing facility by providing number of terminals with internet connectivity and also it should provide offline and online facility using VSAT. Thus the University libraries in developing Countries like India should provide pinpointed information to the right user at the right time at their doorsteps. As language was the medium of communication for transformation of knowledge, e-class rooms and e-teaching should be the source of knowledge in future. Therefore, the curriculum of all the disciplines should teach e-library use to all students henceforth.

5. References

1. Borgman, Christine L. (2000) " Digital Libraries and the Continuum of Scholarly Communication". Journal of Documentation 56 : 412- 430, July 2000
2. Falk, Howard (2003) " Developing Digital Libraries". The Electronic Library 21, No.3 : 258-261.
3. Greenstein, Daniel (2000) "Digital Libraries and Their Challenges" Library Trends 49 : 290-303.
4. Ramesh Babu, B and Gopalakrishnan (2004) Information , Communication, Library and Community Development. Delhi, B.R.Publishing Corporation.

About Author



Dr. D.Rajeswari is Librarian I/C at Sri Padmavathi Mahila University, Tirupati, Andhra Pradesh. She holds MA (Political Science) MLISc and PhD in Library and Information Science. She also worked with S.K. University, Anantapur. She has over 20 research publications in her credit and attended more than 40 Conferences/Conventions/ Workshops etc. Her research interests are digital libraries, web based information services.
E-mail : rajeswari_dondapati@yahoo.co.in

Role of Telecommunication and Networking Technology in the Development of Digital Libraries

Mamata P K G Gopal Reddy
Praveen Kumar Kumbargoudar

Abstract

Many observers have praised the Internet for its omnipresent nature and argued that this global medium is revolutionizing the nature of modern communications. The rise of the internet, is challenging the telecom infrastructure, management and accessibility in India. Telecom in developing countries faces a distinct challenge as compared to developed countries. As the Internet continues to grow, questions of accessibility and infrastructure equity persist. This paper provides an overview of the developments in telecommunications and networking in India.

Keywords: Internet, Telecommunication, Networking

0. Introduction

Invention of computers and Networks (especially internet) are great milestones in the history and development of Libraries. There has been a convergence of a number of developments in computer technology in the last few years, which has significantly affected the way computers can be used in libraries. Emergence of compact disks(CDs), digital versatile disks(DVDs) and high speed processors have large information storage capacity. The internet technology made it possible to access information from any part of the world easily and quickly.

Larson defined digital libraries as not single, stand-alone, repositories of digital data. Instead they are a heterogeneous collection of network based repositories using a variety of protocols for user (and repository) interaction, data encoding and transmission.

Digital library is a logical extension of the networked environment and the development triggered thereof and provides the users with coherent access to a very rare, organized repository of information and knowledge. In a sense it is a global virtual library- the library of thousands of networked electronic libraries/databases.

There is an increasingly wide range of digital resources from formally published electronic journals and electronic books through databases and datasets in many formats, that is, bibliographic, full-text images, audio, video, statistical and numeric datasets. Further, a digital library is not a single entity although it may have digital contents created in-house or acquired in digital formats stored locally on servers.

A digital library may also act as a portal site providing access to digital collections hold elsewhere through the networks.

Digital Libraries Operated with the digital information and Internet Technology for Communication and transmission of information. Conventional or handwritten or printed text can-be converted into digital format easily. It is also easy to store the Digital Information in Compact form. Digital information can be transmitted and received anywhere in the world where infrastructure to send and receive is in place.

For the transmission and communication of Digital Information, form one place to another, Internet is the most popular technology. Internet today is not just another means of communication. Those who use Internet regularly, know that Internet is power. It gives a user not only all kinds of information, but also enables him/her to do things one could not even dream of till recently. Internet gives to its users so much competitive advantage that those without access will face significant disadvantages. In other words, access to the Internet can enable people in all kinds of ways including providing access to education, removing barriers of distance and remoteness, and enabling one to get all kinds of information and to close business deals. But at the same time, lack of Internet access would put a person at a tremendous disadvantage. Earlier the Internet was costlier. But increasing Research and developmental activities of Telecommunication and Networking Industry made the Internet easier, reliable as well as economical.

For the success of digital libraries the universal access to Internet is very important. In others words, universal access here refers to all kinds of people from all the parts of India must be able to access the Internet and digital libraries. India today has about 22 million telephones and less than 0.7 million Internet connections for its one billion people.

In India, there are more rural areas, where internet connection is not accessible. Even though accessed, it is not problematic and not economical. For providing solutions to such problems, Telecommunication and Networking Technology has developed and the present paper explains on such technology.

1. Internet : Its Problems

Today, Internet access is becoming increasingly important. Those who have Internet access have rapid access to all kinds of information, and this could create another divide between the haves and have-notes. A telecom network installed today must provide widespread access to internet.

Internet access on the existing telephone network appears to be very simple-just connect a modem to your telephone line, dial up a router of an Internet Service Provider (ISP), and get searching. Unfortunately, there are several problems in using Internet in this manner, accentuated by the specificities of the telecom network in India. The various problems are discussed as under:

The PSTN, in many parts of the world and especially in India, has been designed to handle 0.1 Erlang ????? traffic per subscriber. While this is largely sufficient for voice telephony, Internet access complicates the matter. While a voice call lasts only for a few minutes, an Internet call usually lasts much longer. Most studies have shown that an Internet user offers a load as high as 0.3 Erlang during peak hour. As the ratio of the Internet users to the total users grows, the PSTN will just not be able to handle the load. The network will get congested and fail to complete a large number of calls. The second problem in accessing the Internet by making a switched telephone call to the ISP, has to do with the analogue modem connection between the subscriber and the ISP. The analogue link, in India, is just not reliable, mainly due to the variable quality of the copper local loop. This is even more so when a subscriber is located in a small town, where the trunk could also be analogue. The quality of the dial-up link varies, and while it does provide 28.8 kbps connectivity occasionally, it often provides only 9.6 kbps or 4.8kbps, Sometimes the modem link also drops, requiring redialling and a new connection. Besides, this method of access works out to be very expensive in India. If one is situated in a metro and uses a local call to an ISP for Internet connection, the telephone charges alone works out to nearly Rs.28 per hour. The charges paid to the ISP are extra. For subscribers in small towns and rural areas making toll calls to an ISP, the amount become astronomically high. If one is located around 250 kms from the nearest ISP, the call charges for one hour of Internet access works out to Rs. 1,200. The third bottleneck occurs at the ISP end. The investment in telephone lines and modems increases rapidly and linearly with the number of customers an ISP serves.

2. Telecom and Networking Technology

To solve the problems of Internet, certain networking technology is operated as under:

2.1 Remote Access Switch

In India, the solution to this Internet tangle is emerging in the form of a low-cost *Remote Access Switch (RAS).* Here, one explicitly recognises that the telecom network is a circuit-switched network whereas the Internet is packet-switched. Circuit-switched voice connections occupy a full circuit, but only for a short duration. Internet connections, however, last for much longer durations, but utilisation is very bursty. When such a connection is made over the circuit-switched PSTN, the advantage of bursty traffic cannot be exploited. Yet, the circuit-switched telephone network is the only available access at homes and offices. The Internet cannot avoid this network, especially when millions of connections are to be made.

RAS equipment is co-located with the local exchange (or even RLUs or RTs of a Fibre Access Network, as we will show later) and connected to it using standard El interfaces. A subscriber desiring an Internet connection, dials up the RAS and sets up a circuit-switched local call, as shown in Fig. 1. The call uses only the local exchange port of the PSTN. These exchanges today have very little blocking and can therefore handle the much longer holding-times (and therefore, higher Erlang traffic) of an Internet call When several subscribers set up Internet calls to the RAS, the RAS multiplexes the bursty data from all these subscribers, and routes the data to the ISP using one or more 64 kbps channels. The 64 kbps connections between the RAS and ISP router could be leased or on dial-up basis, and as shown in Fig.3, the calls take the RAS-Exchange-PSTN-ISP route.

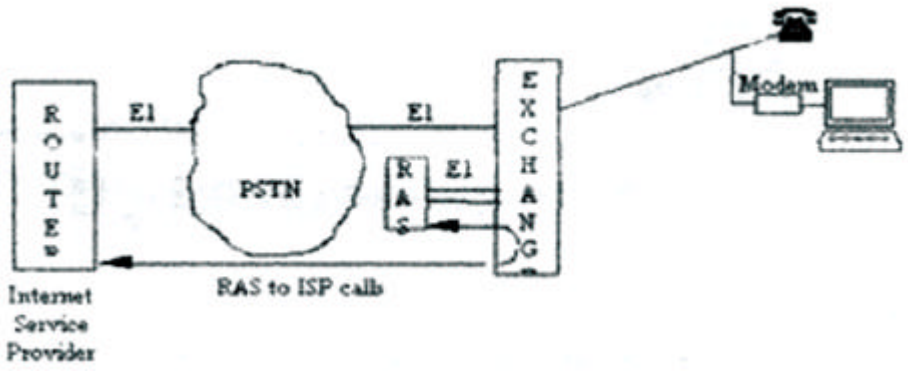


Fig. 1. Remote Access Switch

The call charges can now be low as only intra-exchange calls are being made for such access. The number of connections between the RAS and ISP are now small, and utilise only the reliable digital trunks. No modems are required at either end. Further, multiple subscribers are now being served on each 64 kbps link between a RAS and ISP. Assuming that upto 10 Internet connections use one 64 kbps slot, a single El link (consisting of thirty 64 kbps slots) to router could serve 300 Internet calls. The RAS, while providing an attractive solution to Internet tangle, contributes to a very low per-line cost.

2.2 corDECT Wireiess Access Network

In the mid-nineties, as the cost of the backbone network and switch core reduced substantially, the emphasis shifted to access technologies. Wireless access was and continues to be the most talked about. However, the key to the successful large-scale deployment of Wireless in Local Loop System in India is the right choice of technology. It is important that the wireless solution chosen has a final deployed cost comparable to and preferably even lower than that of the wired solution. Yet, wireline voice quality and data communications at upwards of 28.8 kbps are required. Further, the system must support subscriber density as high as 5,000 per sq. km. A study of available international wireless standards reveals that the choice narrows down to PCS standards such as DECT, PHS and PACS. These standards can be implemented at low cost, and provide wireline quality, high subscriber density, and high data rate, but have small radio range. While microcellular solutions based on these standards are suitable for dense urban areas, one needs to find *innovative deployment strategies* in other cases, so as to cover a wide area.

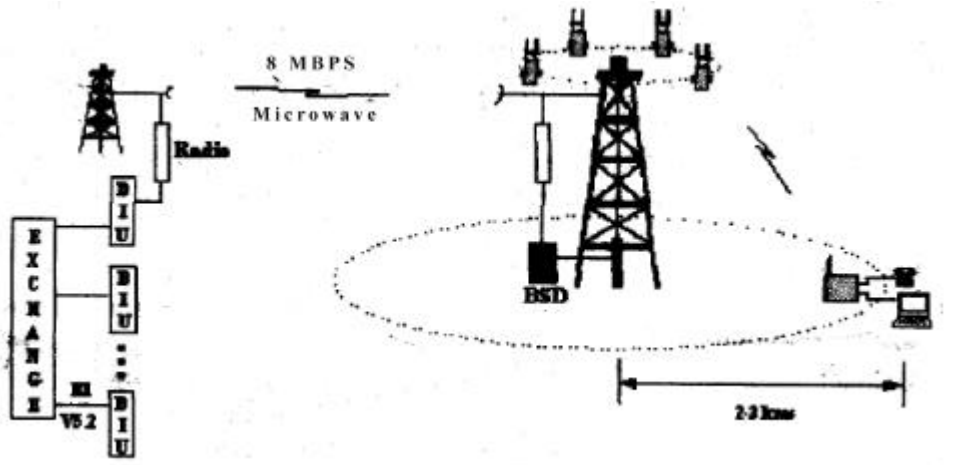


Fig. No.2. corDECT Wireless Access Network

The Telecommunications and Computer Networking (TeNeT) Group at the Indian Institute of Technology Madras (ITTM), located at Chennai, has been playing a key role in defining and developing access technologies suitable for India. Along with Midas Communication Technologies (Pvt) Ltd., Chennai, and in partnership with Analog Devices, USA, for IC development, HIM took up the development of a DECT-based Wireless in Local Loop system.

The system, referred to as corDECT, has an interesting architecture, especially for its fixed part. The fixed part consists of a DECT Interface Unit (DIU) acting as a 1000-line wireless switching unit providing a V5.2 interface towards the main exchange. It also consists of weather-proof Compact Base Stations (CBS) connected to the DIU either on three pairs of copper wire carrying signal as well as power, or onfibre/radio using El links through a Base Station Distributor (BSD). The DIU, CBS and BSD are built primarily using Digital SignalProcessors('(DSP)* with the DIU having nearly 100 DSP ICs. This soft solution, while cutting down the development time, also ensures that the cost of the fixed part is no more than 15% of the total per-line

cost in a fully loaded corDECT system. This, in turn, allows deployment flexibility, and cost-effective solutions can be found for dense urban areas as well as sparse rural areas For example, a new operator who wishes to initially deploy 5000 lines in a mid-sized town/city in the very first year, would use the deployment scenario as shown in Fig.2. Allthe DIUs are co-located with the main exchange and connected to it using the V5.2/E1 interface. Each DIU is connected to a BSD located on a roof-top at a suitable part of the town using a point-to-point 8 Mbps microwave link. At the BSD site, about 12-15 CBS (each serving 50-70 subscribers at 01Erlangs each), along with the micro-wave equipment, are mounted on a 15m roof4op tower to serve an area of 2-3 kms. The subscriber terminal is a wallset (WS), with either a built-in antenna, or a roof-top antenna providing a line-of-sight link to a CBS. The WS has an interface for a standard telephone (or fax machine, modem or payphone) and RS232/V.35 interface for a computer, enabling Internet connection at 28.8/64 kbps. No modem is required as both the wireless link from the WS to the CBS, and the link from the CBS to the DIU, are digital. Digital data is thus routed all the way from the WS to the ISP.

This deployment scenario of 5000 lines Uses no cables and can be made operational in 2 to 3 months. What is particularly attractive is that the total deployed cost of the corDECT Wireless Access Network works out to Rs. 14,000 per subscriber. Even if the system is not folly loaded to begin with, the cost per line does not increase significantly, since the cost of the fixed part is a small percentage of the total cost.

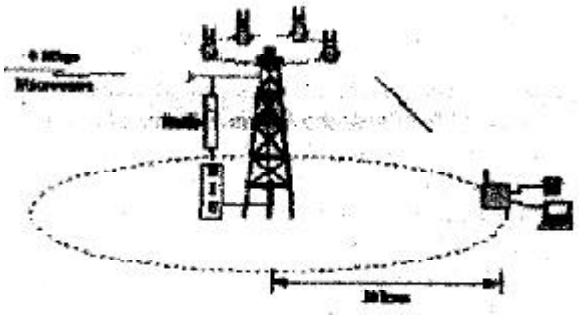


Fig. No.3 CorDECT Wireless Network (for 10 Ions)

In later years, the operator can choose to increase the number of lines to a much larger number, using an optical fibre grid for connecting DIUs to BSDs. The total deployed per-line cost does not alter significantly.

The corDEGT system also offers an excellent deployment opportunity for a small town and its surrounding rural areas. To serve about 1000 subscribers in a small town, an operator needs a tower (about 35m high), somewhere in the town centre. The DIU is located at the tower-base and the base-stations are on the tower. The DIU is connected to the main exchange located upto 30 Ions away using a 8 Mbps microwave link (typically in the 2 GHz frequency band). The base-stations now serve subscribers within a radius of 10 kms using wallsets with roof-top antenna providing line-of-sight links, as shown in Fig.3. The subscriber density served could be as low as 3 subscribers per sq. km, and once again the total deployment cost of the accedd solution works out to Rs 14,000 per line. Internet connectivity at 28.8/64 kbps can be provided to each subscriber at no additional cost.

Deployment in sparser rural areas is possible using the corDECT Relay Base Station (RBS). The solution provides deployment with subscriber density as low as 0.5 subscriber per sq. km at a total cost of Rs. 18,000 per line. A two-hop DECT link is used to provide connection to the subscriber. One link is between WS and

RBS, whereas the other link is between RBS and CBS. Both RBS and CBS use high-gain directional antennas, and are mounted on towers, making a 25 km link possible. The 5 km maximum link distance due to the guard-time limitation of DECT is overcome by use of auto-ranging and timing adjustment. This technique is used in the RBS to support a 25 km link, and to enhance the CBS range to 10 km.

Finally, efficient transmission of packet-switched data on a circuit-switched network is being ensured by the corDECT system, by codeploying a RAS with DIU. Data calls over corDECT are handled differently at DIU from voice calls. The DIU directs an Internet data call from a wallset to the RAS on one 64 kbps slot of the El interconnection. The RAS concentrates Interact data from different subscribers and sends them on one or more shared 64 kbps channels set up between RAS and ISP, via DIU and PSTN.

Here, an Internet call from WS does not enter the PSTN at all. Only themultiplexed data on the few shared 64 kbps channels traverses through the PSTN. The data "calls" from WS to RAS terminate in the Access Network itself.

2.3 Fibre Access Network

The TeNeT group of ITEM along with Vembu Systems (Pvt.) Ltd , Chennai, and Midas Communication Technologies {Pvt.) Ltd., Chennai, has also taken up the development of a cost-effective Fibre Access Network Designed in accordance with the scheme discussed above, the Fibre Access Network again uses a new[7] approach with an aim to provide, apart from the conventional POTS sen ice, large-scale Internet connectivity at a cost affordable in India. As shown in Fig.4, the N-ISDN and HDSL physical layers are exploited in the short copper loop between the RT and subscribers These relatively high-speed digital links carry both voice and data. The digitised voice signals are directed by the Access Server (AS) towards the RT and then to the Main Exchange. However, the Internet data is separated and passed on to a built-in RAS. After concentrating the Internet data from multiple subscribers, the RAS feeds it to the ISP, via the FAN and PSTN, either on leased lines or on dial-up circuit-switched lines. The subscriber terminal provides multiple telephone sockets and an etliernet interface The result is one of themost cost-effective means of providing medium and high-speed permanent Internet connections on a wide scale.

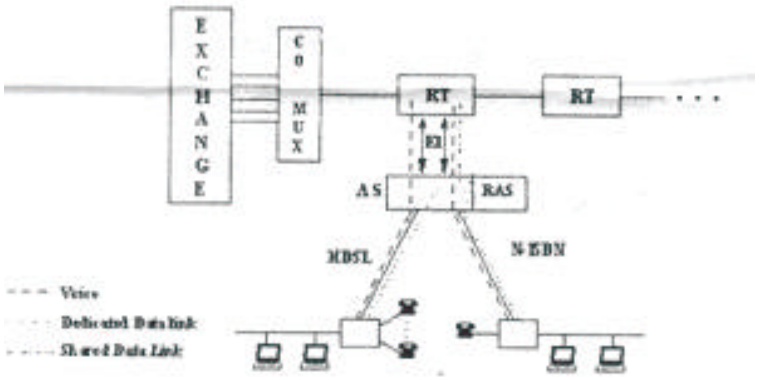


Fig. No.5. Fibre Access Network

Today, the cost of providing Plain Old Telephone Service (POTS) using this FAN is around Rs.9,000 per line. Further, the high-speed permanent Internet Connection costs an additional Rs 8,000.

3. Internet and Management products

The TeNeT Group of 1ITM along with Banyan Networks (Pvt) Ltd., Chennai, is in the process of developing a whole range of Remote Access Switches and Access Servers, including those tailor-made for the corDECT Wireless Access System, and the Fibre Access Network as described above. It is also developing a RAS with built-in digital modems to provide Internet connectivity to existing POTS (Plain Old Telephone Service) subscribers. In all the products, emphasis is on low cost while maintaining high functionality. The additional cost of the RAS amounts to no more than Rs.600 per Internet subscriber.

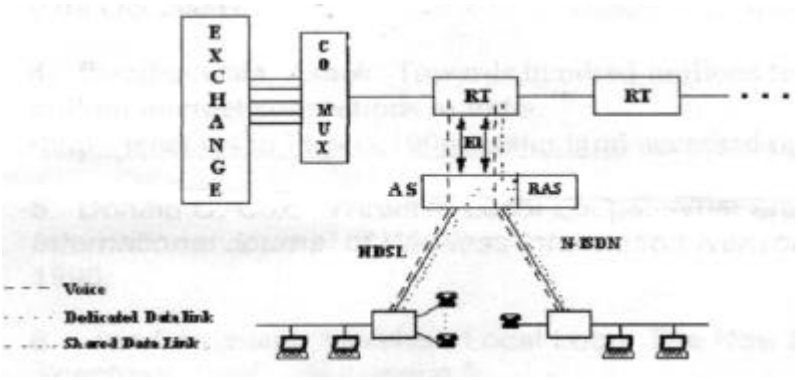


Fig. No.6. Remote Access Switch and Access Servers for Management

Network Management System (NMS) software is today being developed in India by a large number of telecom and computer networking software companies. The capability exists for developing a complex NMS for a large integrated network. Similarly, a number of Indian companies are now developing customer-care and billing systems for clients world-wide.

4. Conclusion

In India, digital Libraries are working only urban areas. In other words, due to networking barriers, till now digital libraries are searched and used by only urban people. This is so because there poor and uneconomical networking facilities in rural areas. The networking facilities in rural area is also costly.

Digital libraries will be successful, only if they will be accessed by all kinds of people both in rural and urban India. For this purpose, the networking technology specified above, are a major revolutionary change in accessing Internet and Digital libraries. These technologies also proved to be economical.

5. References

1. Kumbargoudar, P and Mestri, Mamata: Ideals and illusions of Digital Libraries: Indian scenario. University News. Vol. 40 No. 18 6th -12th May 2002. P.5-8.

2. Jhunjhunwala, Ashok: Unleashing Telecom and Internet in India. htrp://www.tenet.res.in Papers unleash,html accessed on 05th Oct. 2004

3. Jhunjhunwala, Ashok: Towards enabling India through Telecom and Internet connections. http://tenet.res.in/Papers/Intcon.html accessed on 5th Oct 2004)

4. Jhunjhunwala, Ashok: Towards hundred millions telephones and 25 million internet connections in India. (http://tenet.res.in/Papers/100m/100m.html accessed on 05th Oct 2004)

5. Donald C. Cox, "Wireless Local Loops: What are they?" International Journal of Wireless Information Networks, Vol.3, No.3, 1996.

6. Ian Channing, "Wireless Local Loop: The New Frontier", CDMA Spectrum, Sept. 1997, Issue 2.

7. Sain Morgan, "The Internet and the Local Telephone Network: Conflicts and Opportunities", IEEE Communications Magazine, Vol. 36, No. 1, pp. 42-48, January 1998.

8. corDECT Wireless in Local Loop, Analog Devices, Boston, USA, NT, Madras, and Midas Communication Technologies (Pvt.) Ltd., Madras, July 1997.

9. Jhunjhunwala, Ashok and Ramamurthi, Bhaskar: Indian Telecom and Internet Tangle: What is the way out? (http://tenet.res.m/Papers/Telecomtangle/telecomtangle.html accessed on 05th Oct 2004)

10. S. Morgan, "The Internet and the Local Telephone Network: Conflicts and Opportunities," IEEE Commun. Mag., vol. 36, no 1, Jan. 1998, pp. 42-48.

About Authors



Ms. Mamata P K is working as Assistant Librarian in P.G. Center, Bellary of Gulbarga University, Gulbarga since 1999. She did her MLIS(1998); MCom(2002) from Karnatak University, Dharwad. She has contributed 5 research articles in journals.
Email : kumbargoudar@rediffmail.com



Dr. G. Gopal Reddy is working as Librarian at Rashtriya Sanskrit Vidyapeetha, Tirupati, Andhra Pradesh. He has presented number of papers in seminar, conferences and journals. He is also a member of many professional bodies.



Mr. Praveen Kumar Kumbargoudar, is working as Assistant Librarian in P.G. Center, Nandihalli(Sandur) of Gulbarga University, Gulbarga since 1999. He did his MLIS(1994) from Karnatak University, Dharwad. He is doing research in Department of Library & Information Science from Kuvempu University, Shimoga. He has contributed 6 research articles in journals.
Email : kumbargoudar@yahoo.com

Digital Libraries : A Boon for Information Seekers

Mridulata Shrivastava Chitra Ingle

Abstract

Now a days, people responsible for organization and dissemination of informations are steadily switching over from traditional means to electronic tools of documentation system. In this connection the developments in electronic technology, digital libraries have come in vogue to help and support the librarians and users. Digital libraries collect, store, organize and disseminate the information in digital format. This paper explains that how the digitization can help in providing accurate information timely and save time of users and space for data storage, i.e. the data and information specially required by the scientists and research scholars.

Keywords : Digital Libraries

0. Introduction

Rapid growth and emergence of new subjects in the research field especially in scientific fields have made it very difficult or rather impossible to provide the complete and latest information in least time. To satisfy 2nd and 5th laws of library science i.e. "Every reader his/her book" and "Save the time of reader"[1] is becoming a challenge to the librarians/Information scientists. Providing desired information to the clientele in least possible time is the main object of any library or information centre. But now a days it is a Herculean task to access the relevant information from the mountain of information. According to a survey "The total out put of the world information crosses 4 trillion pages in one year which is growing at a rate of 6 to 11% per year over the past decade" [2] Even in India about 60,000 books in various languages are published annually.

Such huge information sources not only create the problems of storage and maintenance but also become inaccessible. Remedy to overcome this problem is the electronic technology or the digitization of documents to store, maintain and quick retrieval of information.

Digitization is the conversion of an item from one format (usually print or analog) into digital and in this process electronic photograph is made of a physical object. An image of the physical object is captured using a scanner or digital camera and converted to digital format, that can be stored electrically and can be accessed via a computer.

Revolution in information and communication technologies has introduced new format i.e. digital format. Now a days digital database alongwith bibliographical data, even full text are easily accessible over networks and on CD ROMs. It provides a new and advanced way of generating, publishing and distributing digital information to a wider community within the shortest possible time. Digital libraries are logical extension and augmentation of physical libraries. Digital library is an electronic library, which can be accessed from widely distributed places in globe with large and diverse repository of electronic objects using computer networks[3].

Donald Waters wrote that "Digital Libraries are organizations that provide the resources including the specialized staff, to select, structure, offer intellectual access to interpret, distribute, preserve the integrity

of and ensure the persistence overtime of collections of digital works so that they are readily and economically available for use by a defined community or set of communities."[4]

Digital library reduces the barrier of distance, timeliness, shared resources and content delivery. It can disseminate data to users without visit to library physically.

Digital libraries should have following elements:

1. Digital library is not a single entity.
2. Digital library requires technology to link the resources.
3. Linkage between Digital libraries and information services are transported to users.
4. Universal access to digital libraries must be a goal.
5. Digital library collection are not restricted to document surrogates but include digital artifacts that have not printed equivalent.

Digital library collection development is not at ease as acquiring and organizing print and non-print materials. The digital libraries came in changing large deposit of analogue data into searchable electronic document. Computer is capable of computing and having large disk storage space. The indexing of images for efficient searching, storage of images and browsing needs electronic engineers and computer scientists.

Computer's ability to store and process vast amount of information and communication technology with its ability to transmit the information from one location to another covers to form "information technology" or "informatics"

The digital revolution employing network based technologies places access to remote resources into the hands of scientist and corporate users. Presently libraries can exploit the use of new digital text, imaging, video, audio, automatic indexing and knowledge based search and retrieval products innovations.

Application of IT enables libraries to participate in library networks and facilitate wider accessibility of information through internet, intranet, CD ROMs.

1. Selection Criteria for Digitization

Following criteria can be adopted for digitization of documents :

- ✍ Value : Priority is given to high value documents. Its serves preventive preservation as well as security by reducing the handling of documents.
- ✍ Condition : Items which are not serviceable due to damage or fragility.
- ✍ Use : Original materials which have high frequency of demand or high retrieval cost.

2. Present Scenario in DRDE Library

1. In DRDE a LAN computer of 75 nodes has been setup. LAN has following applications:

- ✍ Online information access through internet on each node through proxy server.
- ✍ E-mail facility inside and outside campus.

- ✍ Library holding search facility (OPAC).
- ✍ Books – Author wise, Title wise, Keyword wise alongwith its circulation status.
- ✍ Periodical holdings - Year wise, Title wise, Subject wise, Current subscription list, Discontinue / continue status.
- ✍ Micro documents holdings.

2. Procurement in CD format

   Chemical Abstracts, Current Content Life Sciences, Current Contents Physical/Chemical & Earth Sciences, Medline.

3. Online Software - SciFinder

   A research tool that assists scientists and researchers in locating and processing information on wide variety of chemical and science related topics. It retrieves information contained in databases produced by Chemical Abstracts Services as well as the MEDLINE database from the National Library of Medicine. All records are in English.

   CA plus databases containing over 23 million documents from more than 9000 journals covering literature from 1907 to the present.

   The MEDLINE database covers biomedical literature from more than 3900 journals contains more than 13 million biomedical citations from 1958 to the present.

4. Web based full text search.
5. CD of some books & journals.
6. Computerized circulation.
7. Online Journals
   - ✍ Chemical Communication
   - ✍ Chemical Society Review
   - ✍ Current Science
   - ✍ Digit
   - ✍ FASEB Journal
   - ✍ Journal of Medical Microbiology
   - ✍ Microbiology Abstracts Sec.B
   - ✍ Organic & Biomolecular Chemistry
   - ✍ Pharmacological Reviews
   - ✍ Science
   - ✍ Toxicological Sciences
   - ✍ Virological & AIDS Abstracts
8. All in house operations are computerized.

3. On-line Search : A Case Study

Scientist's interest was to search an article published in "Journal of Clinical Microbiology" March 1991. Particular issue of the journal was not available in DRDE library. The issue was searched on the Website http://jcm.asm.org/ Online archival issues were available from Jan 1992 onwards only. Since the requirement was for the 1991 which was not available on the particular site of the journal. Hence it was difficult to access the information.

On searching the PubMed site i.e. www.pubmedcentral.gov/ , we found that the full text was available since its inception year i.e. 1975 Vol.1 and they provide its free access after 6 months of publication. On selecting the "Journal of Clinical Microbiology" Vol. 29 we could get from January to December 1991 and the article in March issue was obtained as full text.

In the above search we found that in the journal site was having full text from January 1998 onward while the earlier issues from January 1992 to December 1994 continued abstracts only and from January 1995 to December 1997 PDF and abstracts were available but the full texts have been covered by the other websites.

4. Future Plan

We are planning to develop digital library in following phases :

Ist Phase	:	Acquiring materials in digital format.
IInd Phase	:	Digitization of reports, special collection related to protection against chemical, biological warfare.
IIIrd Phase	:	Digitization of journals, theses and books available in DRDE library.

The place for exploitation of Internet by the Library and Information Centers are unlimited and endless. Internet provides a wealth of information to Library and Information Centers. It also provides free access to variety of information sources such as online e-books, e-journals both full text, abstracts and contents depending on the publisher's policy, e-news letters and so on.

5. Free Online Information Sources

The freely available online information sources which are identified by the author and compiled at the above mentioned URL, following are the important ones.

1. Electronic Books (http://www.geocities.com/ghosh_tbd/inf3.html)

   This provides link of some organizations and universities to providing free access to thousands of online e-books free of charges.

2. Gutenberg Project on electronic books online (http://www.promo.net/pg)

   This site is having browsing facility by Author and by Title. Author wise list of e-books is available at http://www.promo.net/pg/authors.zip and Title wise list of e-books is available at http://www.promo.net/pg/titles.zip and current list of e-books is available as GUTINDEXALL at the FTP site at the University of North Carolina ftp://ibiblio.org/pub/docs/books/gutenberg/GUTINDEXALL

3. The Online Books Page (http://digital.library.upenn.edu/books/)

   The Online Books Page is a website that facilitates access to books that are freely readable over the Internet. This site include the different sections like Books Online, News, Features, Archives.

4. Banned Books Online (http://digital.upenn.edu/books/banned-books.html)

5. This special exhibit of books have been the objects of censorship or censorship attempts.

5 University of Virginia's E-Book Library (http://etext.lib.virginia.edu/ebooks/ebooklist.html)

In this site 1,600 e-books are publicly available online.

6. Online Electronic Free.Journals (http://www.geocities.com/ghosh_tbd/jnl.html)

Near about 624 online electronic journals are identified and linked in the above mentioned URL. Out of 624 journals 351 journals are available in full text.

7. Databases (http://www.geocities.com/ghosh_tbd/dbase.html)

Near about 29 dabatases are available free of charge. Out of which following are the most important ones.

8. AGRICOLA (Agriculture Online Access) (http://www.nal.usda.gov/ag98)

It is maintained by the National Agricultural Library (NAL). It contains two parts i.e. (a) Online Public Access Catalogue (OPAC) for Books and (2) Journal articles, book chapters.

9. UnCover Database (http://uncweb.carl.org)

The UnCover service has now been integrated to Ingenta's http://www.ingenta.co full text delivery service. The contents and abstracts are available free of charges. It contains 12,087,668 articles and 26,529 publications.

10. US Patent Full-Text and Full-Page Image Database

11. National Library of Medicine Gateway (http://gatway.nlm.gov/gw/cmd)

The current Gateway searches MEDLINE/are pubmed, OLDMEDLINE, LOCATOR plus, MEDLINEplus, DIRLINE, AIDS Meetings, Health Service research Meetings, Space Life sciences Meetings, and HSR Proj.

12. Libraries-Virtual Libraries-Digital Libraries (http://www.geocities.com/ghosh_tbd/lib.html)

In this section reputed libraries all over the world, different digital libraries and virtual libraries are linked.

13. Digital Library of Virginia Tech (Virginia Polytechnic Institute and State University) (http://scholar.lib.vt.edu)

It provides (a) Scholarly Publishing Services including E-journals, Electronic theses & Dissertations, news, reports, survey database and Virginia Tech Publications and (b) Library Services and Archives including E-reserves, special collections, manuscripts, rare books, University archives and searching digital library and archives.

14. Digital Library: University of California (http://elib.cs.berkeley.edu)

It provides quick access of the collection, overview of the collection, image retrieval by image content, document image analysis and

15. Virtual Libraries (http://elib.cs.berkeley.edu)

This site contains the link of the virtual libraries on Agriculture, Business and Economics, computing, Communication and Media, Education, Engineering, Humanities, Information & Libraries, International Affairs, Law, Recreation, Regional Studies, Science and Society. [5]

6. Problems of IT based information system

Some shortcomings or cons of digital library are as follows :

1. Paper based publications are more comfortable in reading in comfortable postures.
2. Digitization of printed matter is expensive and difficult to maintain. It will be a formidable task with the associated software, human skills and copyright problems.
3. Internet based publications have two main problems :
   - ✍ Lack of organization of information, so one has to spend several hours for exhaustive information searches. Internet may be compared to a large library without a catalogue having bulk unclassified random shelved documents.
   - ✍ A survey was published in the journal "Computers in library" in 1995 found that librarians usually take less time to provide information from library collection than internet. A paper published in "Communication of ACM" 1995 also advised "If you are in a hurry, go to library not to the internet" [6]

7. Digitization scenario in India

At present following major libraries have adopted or are in process of digitization:

1. All India Institute of Medical Science, New Delhi.
2. Archaeological Survey of India, New Delhi
3. Banaras Hindu University Library, Varanasi.
4. Central Secretariat Library, New Delhi.
5. Department of Ocean Development, Govt. of India.
6. Department of Space.
7. Geographical Survey of India.
8. Indian Agriculture Research Institute Library, New Delhi.
9. Indian Institute of Management Library, Kolkata.
10. Indian Institute of Science, Bangalore.
11. Indian Veterinary Research Institute Library, Izzatnagar.
12. Indira Gandhi Memorial Library, Hyderabad.
13. Ministry of Environment & Forest, Govt. of India.
14. MS University Library, Baroda.
15. National Archives, New Delhi.
16. National Dairy Research Institute Library, Karnal
17. National Gallery of Modern Art, New Delhi
18. National Informatics Centre
19. National Instt. Of Advanced Studies on Rare Manuscript Preservation Project, Survey of India.
20. National Library, Kolkata.
21. Publication Division (Min. of Information & Broadcasting),New Delhi
22. Rajasthan Tourism Development Corporation Ltd.,
23. Sahitya Academy, New Delhi

8. Conclusion

Application of IT enables libraries to participate in library networks and facilitate wider accessibility of information through digitization. The key to the success of digital library lies in proper utilization and accessibility to usable and stable systems.

9. Acknowledgement

The authors are grateful to Er. K. Sekhar, Director, Defence Research & Development Establishment, Gwalior for encouragement and keen interest in bringing out of this paper. Authors are thankful to Dr. B.K. Bhattacharya, Scientist 'F' & Head TIRC for valuable suggestions during the preparation of this article. Thanks are also due to our colleague Mr. Narayan Singh for his help and co-operation .

10. References

1. Ranganathan SR(1957), Five Laws of Library Science. Madras: Madras Library Association Publication Series 23 .
2. Dhake RPS, Arora K(1995), Electronic library – A myth or a Reality. Ann.Lib.Sci.Doc. 42(4): pp 152-59.
3. Goswami SK, Ghosh BK, Digital Library Environment- Indian context. XX IASLIC National Seminar, Patiala, India, 2002,* pp 229-234.
4. Drake MA Ed., (2003), Encyclopaedia of Library and Information Science 2nd Edition, New York: Marcel Dekker Inc. pp 884.
5. Ghosh TB(2002). Freely Available Online Information Sources and their Impact 2002, on Libraries and Impact on Libraries and Information Centres. "Internet Engineering for Library and Information Centres" H.Anil Kumar et al Ed. Caliber 2002, Ahmedabad, INFLIBNET Centre,, pp 376-383.
6. Pandey AC et al (1999). Role of Computer Networks and Information Management. IASLIC xxvii All India Conference, AGRA , *pp28-31 .
7. Chandra R et al . Electronic References Sources of Biomedical Literature.
8. "Modern Technologies for Biomedical Information Handling." Lazar Mathew, T. et al Ed., Delhi, INMAS, *pp 85-90.

About Authors



Mrs. Mridulata Shrivastava obtained her M.Sc. in Chemistry and B.Lib.Sc. from the Jiwaji University, Gwalior. She joined DRDO at the Defence Research & Development Establishment (DRDE), Gwalior in 1974 and presently working as Officer-in-Charge at Technical Information & Resource Centre. She is the member of Computer Society of India.
Email : geetam_s@yahoo.co.uk



Mrs. Chitra Ingle obtained her M.A in Psychology and MLISc from the Jiwaji University, Gwalior. She joined DRDO at the Defence Research & Development Establishment (DRDE), Gwalior in 1981 and presently working as Senior Technical Asstt. 'C' at Technical Information & Resource Centre.
Email : chitraingle@rediffmail.com

Towards The Design and Development of E-Books : An Experience

P. Rajendran B Ramesh Babu
S Gopalakrishnan

Abstract

This paper advocates a strategy to select PG level thesis materials based on their intellectual value, and to define technical requirements for retrospective conversion to digital form based on their informational content. Defining the conversion requirements the document attributes may be the guarantee of building digital collections with sufficient richness to be useful for the long-term. There are factors that are compelling towards conversion of digitized form.

Keywords : E-Books, Digitization

0. Introduction

E-publishing has developed rapidly over the past couple years. Electronic publishing of books is a major development that is quickly causing changes in the industry. In recent years, the traditional publishing houses have also climbed on board and are converting of new releases and backlists into the available electronic delivery formats, including both e-books and print-on-demand technology. However, Adobe PDF is also a widely used format for e-books and it competes with the OEB standard.

The University of California's California Digital Library formed an E-Book Task Force in August 2000 to investigate the e-book market and develop guidelines, principles and strategies for making e-books a viable part of the University's digital collections. It also highlights elements and strategies we determined to be important for future academic use of e-books. The commercial production, sale and distribution of e-books that has changed how libraries need to deal with e-books, and what prompted our investigation.

1. Scope of this Paper

This papers presents the experience gained in the on going project of converting the Masters Degree level project's reports in final year Engineering Science. Further it explains the methodology for the creation of e-books such as converting MS Word document to PDF files; Creation of content page in HTML format; linking the PDF files to content page according to branch of Engineering Science and finally adding to the Adobe E-book reader for the ultimate and user purpose. At the initial stages of the projects a sample of M.Tech projects submitted to the Electrical and Electronics Engineering Department of SRM Deemed University has been considered. The experiences gained in this process are presented in this paper.

2. Concept of E-Book

Electronic books offer creative possibilities for expanding access as well as changing learning behavior and academic research. Content can always be accessible, regardless of time or place, to be read on PCs or on portable book readers. Books need never go out of print, and new editions can be easily created. One can carry several titles at once on a portable reader and, over time, build a personal library. Features such as full text searching, changeable font size, mark-up, citation creation, and note taking will enhance usability. Print text can be integrated with multi-dimensional objects, sound, and film to create a whole new kind of monographic work.

3. Information Materials Considered For The Project

The following are the ideal materials that can be created as e-books in libraries

- Student Projects / Theses / Dissertation
- Staff publications
- Seminar/Conference Proceeding published by the institution
- Institutional Publications (like Magazines, News Letters etc.,)
- Research Reports

4. Methodology Adopted for Creation of E-Books

The methodologies adopted for creating digital format are:

- Soft copies of the student's projects are obtained from departments in MS word format.
- Project works received from the members are different formats
- This can be complied all the files in one file.
- After compiling the files, it can be saved MS word format.
- The different steps involved in the creation of e-book format are:
- MS word to PDF
- Creation of Content page in HTML format
- Linking all the PDF files to content page by branch wise
- Adding to the Adobe E-book reader for use of the members

4.1 STEP -1 From MS Word To PDF Conversion

After compiling the documents in one file it can be converted in to PDF files. The process of conversions of the word document to PDF, using Acrobat writer 4.0 is shown in Figure. 1

- Select File option,
- Click Print,
- Find a pull down menu
- >From the menu select an Acrobat Distiller press OK button.

The Acrobat Distiller started to convert the files from MS word to PDF files, after completion of the process PDF files open in the computer and it can be saved to the location.

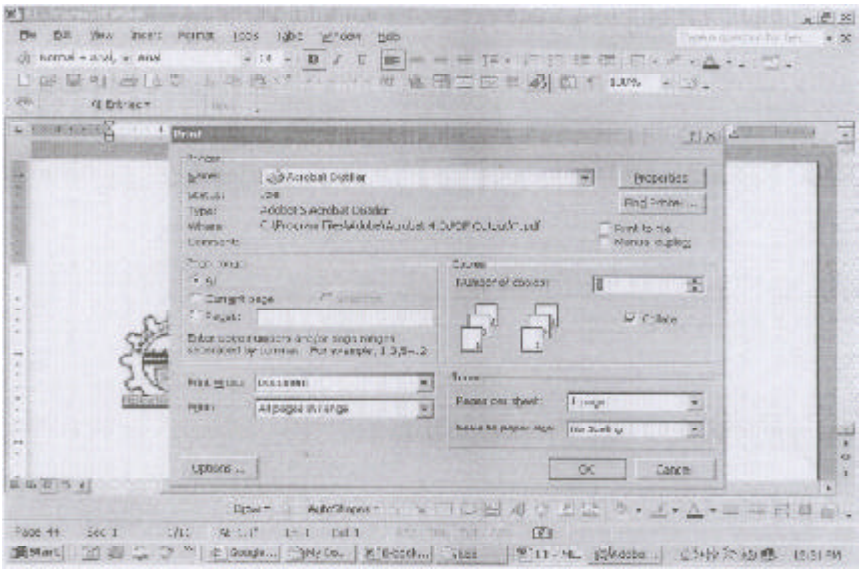


Figure -1

The converted document of MS Word to PDF will open in the same windowshown in the Figure -2 and this PDF document in the folder department wise.

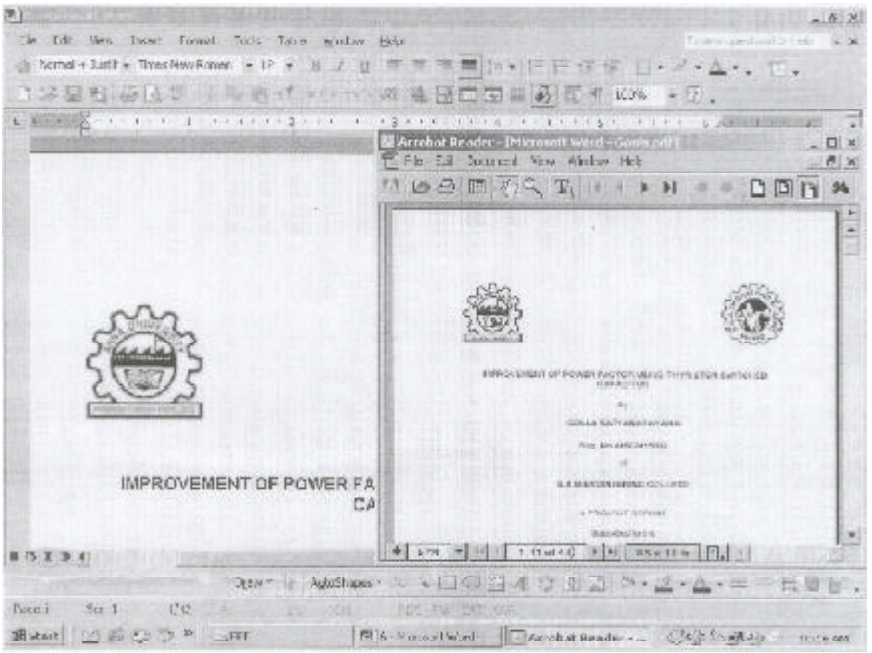


Figure - 2

4.2 STEP - 2 Creation Of Content Page Html Format

Content page is to be created in HTML file using Microsoft Front Page Editor. Content page is contains the Name of the Department, Year, Title of the Project and Name of the student who has done the project, under the guidance of the staff member. After completion of the above process the entire project is digitized to PDF Files and saved in one of the folder. This files is to

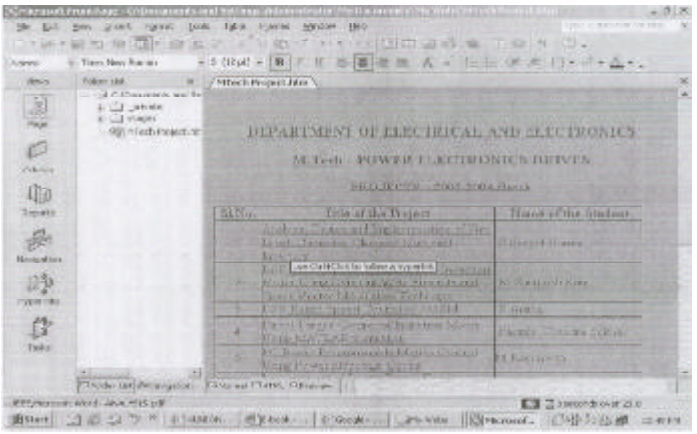


Figure - 3

4.3 STEP - 3 Creating Links to PDF files to Content Page by Department Wise

After creating the Content Page of the project titles it can be opened in Internet Explore window as show in the Figure - 4 and this Internet Explorer file to be added in to Acrobat e-Book reader for use members as show in the Figure - 5

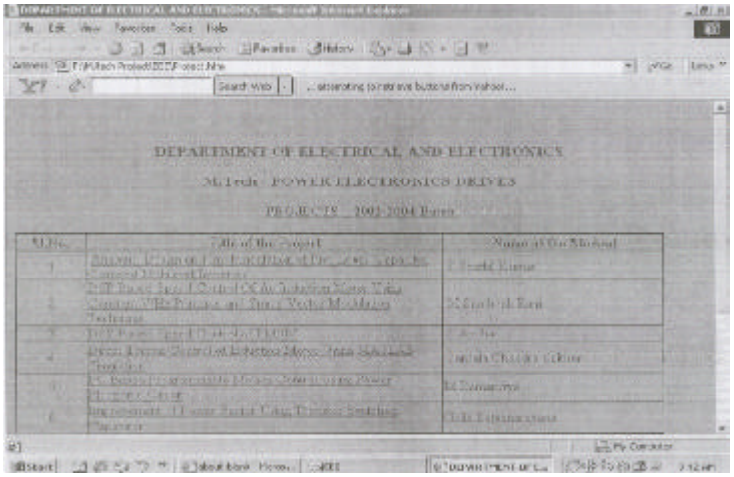


Figure -4

From the above screen left side of the window there is pull down menu can used to view the entire projects listed in the content page along with Title of the project and name of the person who have done that project work. Clicking the required projects, the selected project will open in the same window as a PDF file as shown in the figure - 6

The Acrobat e-Book reader is user friendly to use the multiple pages of e-books. The main directory file of the projects will be added in the e-books reader library and it will appear on the screen as shown in the Figure - 6. Clicking the M.Tech Project file, which appears on the content pages of the Department wise project of eBooks window, is shown in the figure - 5

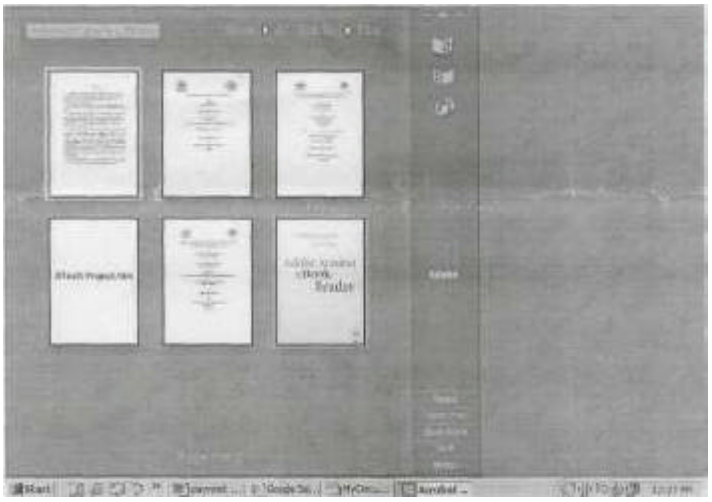


Figure - 5



Figure – 6

5. Problems In E-Book Creation

While developing e-book the author confronted the following problems.

- Font Issues
- Lack of a Standard format
- Digital Rights Management (DRM)
- Reproduction of Graphics
- Reader Hardware

5.1 Font Issues

Fonts are both an advantage and a disadvantage for e-Books. The ability to resize fonts to fit the needs of the vision-impaired reader is an advantage. However, fonts on a computer screen at sizes equal to those used in printed materials are not easy on the eyes. After scanning the clear images of the letters are not captured and therefore many times, an editing work has to be made.

5.2 Lack of a Standard format

Imagine trying to read a book if there was no agreement on how to put the words on a printed page. Should they be printed as black on white or white on black.

Should they run left to right or right to left. May be they should be printed from top to bottom. Should a printed book be a bound together or just loose page. Should it open from the top, the left or the right? Without agreement on these simple standards, reading would be a definite adventure. The problems is many tiems worse in the e-book publishing industry. Lack of a single overriding standard means that authors, publishers, and even readers must choose what format they will support.

We have seen that different combinations of hardware and software are better or worse for different types of content. That guarantees that there will be multiple formats supported by different vendors trying to take advantage of specific markets. Soft book Publishing's reader has a single 8" x 11" colour screen, which is more expensive but better suited to reproduction of more demanding reading materials like textbooks. Each reader has his or her own format. Rocket e-book is a binary format based on HTML and Soft Books Publishing uses a format based on Adobe. Pdf. An attempt has been made to crate a single universal standard using pdf.

5.3 Digital Rights Management (DRM)

Protecting an author's copyright is one of the prime concerns when distributing books via electronic format. Although copying an entire book is now possible, the cost and inconvenience of doing it manually has kept this type of piracy to a minimum. But when coping is as easy as duplicating a file, piracy becomes a major problem. Some method of securely distributing e-books and preserving the copyrights and royalties of authors is essential if e-book are to flourish. Adobe, Xerox, and Microsoft are just a few of the companies currently working on this problem.

5.4 Reproduction of Graphics

Full colour graphics, complex tables, and figures are not easily reproducible on small screens. Some e-Book formats don't even support the inclusion of images.

In order to lower the price and increase battery life some e-Book hardware uses only a black and white screen. All of these factors make reproduction of graphic elements on many e-book a challenge.

5.5 Reader Hardware

The problem associated with e-Books is Reader hardware and software itself major problem for using e books. e-Book readers range in size from a small handheld PDA to a desktop computer and dedicated readers are relatively expensive when compared to the price of a book. Of course, many people a already own personal computers and laptops, but the change in habits required by these devices has already been mentioned as a problem.

The variety of incompatible hardware, software, and formats also leads to a problem. Since many e-book format are not interchangeable, a consumer must choose carefully when purchasing a platform or they might not be able to read the books that they want.

6. Suggestions

Based on the experience, the authors propose the following suggestions:

- ✍ Needs uniformity in format of print version
- ✍ Soft copy is shall be in a standard format, with regard to layout; references citations; font etc.,
- ✍ Distinction shall be made between Bibliography and Reference in the thesis

7. Conclusion

Electronics publishing, or e- publishing, uses new technology to deliver books and other content to readers. Because the technology allows publishers to get information toreaders quickly and efficiently, it.is causing major changes to the publishing industry, as we know it. It will also impact they way we read, offering new hardware and softwaredevices. We are only beginning to see the ramifications of e-publishing. E publishing is a very broad term that includes a variety of different publishing models, including electronics books (e book), print-on-demand (POD), email publishing. . Today the trend is providing digital signature without using a pen. Wireless publishing, electronics ink and web publishing are some of the ongoing features of the era. More types of e-publishing are sure to be developed in the near future.

E-Publishing will have a great role in making the libraries and information centers towards digital environment. E-journals and e-books would become very practical media of dissemination of information in future. It may be a funny thing to see a student having a dozen of books in his packet along with e-book reader. A kind of 'mobile reading' could be realized in the coming generation

8. References

1. California Digital Library. Joint Steering Committee for Shared Collections. E-book Task Force. Report. March 15, 2001. Available at http://www.cdlib.orR/about/publications/
2. Gupta, KK and Gupta P K (2002). E-Book: a new media for libraries. In. Caliber 2002: Seminar proceedings. Ahmedabad. INFLIBNET Centre, pp.384-399.
3. Hawkings, Donald T (2002). Electronics book: A major publishing revolution. Part 1: General considerations and issues. Online. 24(3). July/August: 14-28

4. Ramesh Babu, B. and Gopalakrishnan S (2003). "E-Books as a learning tool for the rural youth". 5.In: Proceedings of the Workshop on the Production of Books for Rural Youth held during 8-9[1] November 2003 at Dravidan University, Kuppam (A.P.) , Chennai : Foundation for Information and Communication, pp. 19-21.

6. Ramesh Babu, B. and Gopalakrishnan S (2004). "Design and development of E-Books: An experience" IN: Digital Information Exchange: Pathways to build Global Information Society (SIS 2004) edited by Harish Chandra, P.Pichappan and Ramesh Kundra. Chennai: Indian Institute of Technology, pp. 17-23.

7. hltp://www. writers write.com/iournal/dec00/stork2.htm

About Authors

Mr. P Rajendran is working as a Librarian at SRM Institute of Science & Technology (Deemed University), SRM Nagar, Kattankulathur, Kancheepuram, Tamil Nadu. He has presented number of papers in seminar, conferences and journals. He is also a member of many professional bodies.

Prof. B. Ramesh Babu is a Professor in Department of Library & Information Science, University of Madras, Chennai. He has presented number of papers in seminar, conferences and journals. He is also a member of many professional bodies.

Dr. S. Gopalakrishnan is working as Assistant Librarian in Anna University, Chennai. He has presented number of papers in seminar, conferences and journals. He is also a member of many professional bodies.

Information Management in Indian University Libraries : A Survey

P M Naushad Ali

Abstract

The emergence of modern technology has created greater opportunities in the libraries in the 21st century. The technological innovations of the recent times have enabled the university libraries to provide right information to right users expeditiously, exhaustively and in minimum possible time. This technological advancement has also forced the library users to seek the alternative source of printed information with electronic characteristics. Automation of any library is a challenging task especially for university libraries having different type of users, collection and environment. Serious attempt has been made by a majority of university libraries to automate the library functions and services with few exceptions. A few libraries are in modernization phase others are in the initial stage of automation. The purposeful and systematic acquisition and application of information with the help of suitable information technology are the core ideas behinds the concept of Information Management. The paper discusses about the information management activities,, current status of IT application in select university libraries in the country and describes the modern infrastructure facilities available in these libraries. This paper also identified and discusses some of the issues associated with I.T. *based services in university libraries on users point of view.*

Dr. P M Naushad Ali, Reader, Dept.of Library & Information Science, Aligarh Muslim University, Aligarh.
E-mail : pmnali@hotmail.com

Mapping Metadata Standards

Ashwini A Vaishnav

Abstract

Information available on internet is vast & easy to access, at the same time it is difficult to retrieve right information because of lot of noise. One solution to this is cataloguing of web with some standard schemes. The paper discusses different metadata standards & attempts to map DC with TEI,EAD,MARC-21 & CCF.

Dr. Ashwini A Vaishnav, Professor & Head, Department of Library & Information Science, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad.
E-mail : dlisbamu@sancharnet.in

 © **INFLIBNET** Centre, Ahmedabad

Digital Preservation : Issues and Strategies

L Rajendran S Kanthimathi

Abstract

In this paper, we have discussed about the main issues that long-term preservation requirements rise for digital libraries, describe some of the current models for digital archiving, and discussed some of the outstanding issues for research and development. The remainder of the paper focuses a broad overview of the issues rather than an in-depth analysis of specific models. There are simply too many issues are too much activity to give all of the issues and strategies the attention that they deserve.

Mr. L Rajendran, Senior Librarian, Tagore Engineering College, Chennai.
E-mail : rajendranlak@yahoo.com
Dr. S Kanthimathi, Librarian (SG), Rani Anna Govt. College for Women, Tirunelveli, Tamil Nadu.

Web based Information Retrieval System through ODBC

Pratibha Verma Manoj Kumar Verma Kamalendu Majumdar

Abstract

Modern libraries are constituted within and by a tradition of techniques and practices that represent 100 years of codified professional knowledge. This article provides an overview of this tradition that created a complex environment of expectation and misunderstanding for introducing library automation. A generation of systems development was needed to assimilate and further develop this tradition through PWS (Personal Web Server).

Ms.Pratibha Verma, Professional Trainee, Central Library, Indian Institute of Technology, Kharagpur.
E-mail : pratibha_21sep@rediffmail.com
Mr. Manoj Kumar Verma, Visiting Lecturer, DLIS. Guru Ghasidas University, Bilaspur, Chattisgarh.
Mr. Kamalendu Majumdar, Assistant Librarian, Central Library, Indian Institute of Technology, Kharagpur.
E-mail : kamal@library.iitkgp.ernet.in

 © **INFLIBNET** Centre, Ahmedabad

INFLIBNET – The Right Network for University Libraries

G Dhanasegaran

Abstract

Defines the concept of consortia – Traces the development of Library Consortia – Explains the aims – States the Functions – Enumerates the Advantages and Limitations – Describes the Consortia initiatives in Indian Scenario – Milestones of the Library Consortia in India – Speaks about the various barriers to Library Consortia in the Indian context – INFLIBNET's march towards Standard Consortium – Points out the suggestions.

Mr. G Dhanasegaran, Deputy Librarian, Dr. T. P. M. Library Madurai Kamaraj University, Madurai, Tamil Nadu.

Scenario of Digital Library and its Services in Bangladesh and India : Views of Library and Information Scientists

Ch Ibohal Singh Th Madhuri Devi Th Shyam Singh

Abstract

The views of Library and Information scientists of Bangladesh and India collected during PLANNER – 2004 are analysed to understand the present scenario of Digital Library (DL) and its services in the two countries. Technological requirements developed and DL Services rendered by both the countries are studied. The problems and suggestions for the successful implementation of the DL are also highlighted.

Ch Ibohal Singh, Visiting Lecturer, DLIS, Manipur University, Imphal.
Th Madhuri Devi, Associate Professor & HOD, DLIS, Manipur University, Imphal.
Th Shyam Singh, MLISc Student, DLIS, Manipur University, Imphal.

 © **INFLIBNET** Centre, Ahmedabad

Traditional Library - Towards Digital Library in India

Bhupen Goswamee Nityananda Pathak

Abstract

In this paper an analytical discussion is made about the objectives and activities of digital library. Mentions the importance of Digital Library system in academic library and information centers. Explain the scenario of academic libraries and information centers. Highlights the impact of networking on digital library. Stress the role of Government and national agencies in bringing networked environment in academic information services in India.

Mr. Bhupen Goswamee, Libraian I/C, K.K.Handiqui Library, Gauhati University, Guwahati.
Mr. Nityananda Pathak, Library Prof. Asstt., K.K.Handiqui Library, Gauhati University, Guwahati. Assam, India

Subject Gateways : A Tool for Retrieving High Quality Information Over the Web

P Ganesan

Abstract

The main purpose of the web is to provide quality information. Most of the organizations, educational institutions, and individuals are flooding information on the web are currently looking for ways to help their users to discover high quality information on the Internet in a quick and effective way, since the resource available on the web is immense. Unlike, print resources, web is mostly unfiltered, since the publication through web is very easy. Some of the problems relating to information available via Internet are the ease of publishing, information overload, inaccuracy, and ephemeral nature materials disseminated through personal and commercial homepages. Though the web resources are immense, excellent information also resides along with them. Researchers and academics do not always have the time, inclination or skills to surf the Internet for resources that could support their work. Since Internet publishing and communication become more commonplace this could disadvantage some researchers as they will miss valuable information and communication resources. In the traditional information environment human intermediaries, such as publishers and librarians, filter and process information so that users can search catalogues and indexes of organized knowledge as opposed to raw data and disparate information. Subject gateways also work on the same principle - they employ subject experts and information professionals to select, classify and catalogue Internet resources to aid search and retrieval for the users. Users are offered access to a database of Internet resource descriptions which they can search by keyword or browse by subject area. They can do this in the knowledge that they are looking at a quality-controlled collection of resources. A description of each resource is provided to help users assess very quickly its origin, content and nature, enabling them to decide if it is worth investigating further. This paper discusses about the subject gateways, its relevance and characteristics.

Mr. P Ganesan, Documentation Officer, IGM Library, University of Hyderabad, Hyderabad.

 © **INFLIBNET** Centre, Ahmedabad

Today's Need of Digitization Libraries

Prasad Pandit Chaudhari Harsha Ravindra Bhole

Abstract

The Libraries in 21st century will be known as paperless or electronic i.e. digital library. The digital library system unlike traditional library system, located in a building or particular location, it is distributed all over the world through internet. Anyone can access it at anytime from anywhere. This article deals with different aspects of digital library such as need, benefits including how to automation of library. We hope that digital library will turn into reality in coming future and provide facility of round the clock availability of information anywhere.

Mr. Prasad Pandit Chaudhari, Faculty of Computer Engg., Government College, Jalgaon.
E-mail : pra_priya @rediffmail.com
Ms. Harsha Ravindra Bhole, Faculty of Computer Engg., Government College, Jalgaon.
E-mail : harsha_bhole@indiatimes.com

Digital Library Technology & Services

Dipak Krushnarao Bhalekar Prashant P Deshmuk

Abstract

Digital conversion of Library materials has advanced rapidly in past few years. Digitization of all categories of materials is now possible with the advances in digital technology. The paper discuss the concept of Digital library, Component, Technology, Functions and services of Digital Library.

Mr. Dipak Krushnarao Bhalekar, Librarian, B.N.College of Engineering, Pusad. Dist. Yavatmal, Maharashtra**.**
Dr. Prashant P. Deshmuk, Librarian, B.N.College of Engineering, Pusad. Dist. Yavatmal, Maharashtra.

 © **INFLIBNET** Centre, Ahmedabad

Virtual Reference Practices in Libraries of India

Mahendra Maheta

Abstract

As public access to the internet increases, libraries will receive more and more information online, predominantly through email. Object of this paper is to check the current developing, testing, and evaluating procedures and mechanisms that will enable libraries to work in providing reference assistance over the Web to support patrons' image information needs. The user-centered project is based upon a successful model for digital reference practice that has been widely embraced in the digital library community. This approach is expected to yield new insight into users' image seeking behavior that will help libraries to provide transparent access to visual resources across collections and institutions. This article presents an overview of the project and discusses the challenges involved in helping users find appropriate images on the web.

Mr. Mahendra Maheta, Librarian, Vivekananda Institute of Hotel and Tourism Management, Bodhigram, Dist. Rajkot, Gujarat.
E-mail : librarian007@rediffmail.com

Digital Libraries and Preservation of Digital Materials

Amrit Pal Kaur

Abstract

Digital Libraries are creating an environment in which the distances between libraries and patrons are shrinking bringing with their ambit environs hitherto considered, geographically hopelessly remote. The present article is divided into two parts. The first part gives details of digital libraries and issues related to it. This part includes introduction, history, definitions and different views of authors, factors effecting emergence of digital libraries. This part also explains various types of digital libraries that are divided on the basis of user group, geographical scale and coverage, document type and on the basis of digital contents. At the end of this part, some advantages and disadvantages of digital libraries, some problems and issues such as information accuracy, copyright and intellectual property right, technological obsolescence, life of digital materials and pricing issues related to digital libraries are also discussed. The second part explores various strategies and methodologies for preserving digital materials and also focuses on the pressing need for libraries to develop strategies and practical action plans to maintain the safety and accessibility of the world's historical and cultural heritage.

Dr. Amrit Pal Kaur, Reader and Head, Dept. of Library & Information Science, Guru Nanak Dev University, Amritsar, Punjab.
E-mail : amrit_lisc@yahoo.co.in

 © **INFLIBNET** Centre, Ahmedabad

Going Digital

Rajesh Kumar Tripati R K Singh Rajesh Kumar Singh Pravish Prakash

Abstract

This article provides an overview of recent developments in information technology and how it is evolved in to digital library concept. As information technologies have developed, the digital library is making the library undergo a changing paradigm of its role to create, organize, and distribute information resources. Digital libraries have created and promoted innovative information services with digitization of resources. The development of digital libraries has been attracting the attention of many countries and India is no exception. To build the digital library, various innovative projects are currently in progress involving a range of different libraries and institutions. This article also discusses about various digital mediums.

Mr. Rajesh Kumar Tripati, Lecturer, DLIS, Dr. Ram Manohar Lohia Avadh University, Faizabad, Uttar Pradesh
Mr. R K Singh, Librarian, Central Library, Dr. Ram Manohar Lohia Avadh University, Faizabad,
Mr. Rajesh Kumar Singh, Research Scholar, Babasaheb Bhimrao Ambedkar University, Lucknow, Uttar Pradesh
Mr. Pravish Prakash, Research Scholar, Babasaheb Bhimrao Ambedkar University, Lucknow.

Library Portals

Sachin Vishwanath Vaidya

Abstract

Various portals are mushrooming in every discipline; library and information science is no exception to it. This paper begins with defining the various terms coming under the umbrella of portals. It then details various bibliographical tools and attributes of portals. It begins the discussion of library portals with need for the same. All other aspects of library portals, like meaning, definition, coverage, software, are covered. Then the paper details classification of library portals. The paper ends with describing the strategy for making the best use of library portals.

Mr. Sachin Vishwanath Vaidya, Librarian, Sir Vithaldas Thackersey College of Home Science, S. N. D. T. women's University, Juhu Road, Mumbai.
E-mail : sachin_svt@sify.com

 © **INFLIBNET** Centre, Ahmedabad

Role of Digital Libraries in E-Learning

Sankar P Manikandan S

Abstract

This article describes areas and mechanisms aimed at providing educational organizations strategies for implementing electronic learning (e-Learning). As library collections are steadily following their catalogues in becoming electronic and machine-readable, so e-Learning is tracing a similar path beyond distance learning to enterprise-wide services that support and extend the entire curricula and related institutional services. New computer based storage and communications technologies are making possible many progressive methods for the creation and delivery of educational resources. Because of the impact that digital libraries and computer networks are having, the role and importance of 'conventional' approaches to educational provision need to be reevaluated. This is particularly so in the case of post compulsory education and the growing need for facilities to support 'lifelong' learning activities within both academic and non academic organizations. Discusses the important role of digital libraries as a resource for the creation of an infrastructure to support e-learning.

Mr. Sankar P, Student of DLIS, Bishop Heber College, Trichy, Tamil Nadu.
Mr. Manikandan S, Librarian, D J Academy for Managerial Excellence, Coimbatore

Models and Methods of Library Consortia

M Suriya G Sangeetha

Abstract

In the age of knowledge revolution the libraries face enormous challenges and opportunities. As campuses move into the information age, the mission and role of the library is being redefined. While the amount of information, libraries need to acquire continues to increase, the resources available to do so are insufficient. The recent economic recession, money for education and libraries in India became very tight, requiring cuts in serial subscriptions and book purchases for academic libraries. At the same time, subscription prices were soaring, as were the costs and number of databases and journals available. Library planning is now essential in order to maximize available resources and take advantage of emerging technologies. Collaboration is a strategy to extend library access, share the costs of library collections and services, and develop an academically and economically sustainable model of scholarly communication. The two areas identified for partnership viz., a) developing the collection on shared basis and b) developing the services for exploiting such collection. This paper aims at gaining some insight into the maturing process of Library consortia as success for partnership. The focus of the paper lies on (i) what is a consortium? (ii) The salient features for library consortium, (iv) Advantages and Disadvantages in purchase of Consortia, (v) Purchasing Model of Library Consortia, (vi) Few examples of Library Consortia.

Dr. M Suriya, Prof. & Head, Dept of Library and Information Science, Annamalai University.
E-mail : au_suriya@rediffmail.com
Ms. G Sangeetha, Ph.D. Scholar, Dept of Library and Information Science, Annamalai University.
E-mail : salem_sangeetha@yahoo.co.in

 © **INFLIBNET** Centre, Ahmedabad

Digital Library Consortia : An Indian perspective

O N Chaubey Rajesh Kumar

Abstract

The paper discussed the historical background & concept of consortia. Highlights present scenario of digital library consortia in Indian prospective. Some well known consortia of India as well as abroad also discussed. Also thrown light in various types and steps involved in consortia.

Dr. O N Chaubey, Senior Documentation Assistant, Indira Gandhi National Centre for Arts, New Delhi.
E-mail : lohani_minoo@rediffmail.com
Mr. Rajesh Kumar, Librarian, Sri Sringeri Sharada Institute of Management, New Delhi.
E-mail : vkj@library.iitkgp.ernet.in

Consortia Among Trichy Libraries : A Proposal

B S Swaroop Rani N Prabakaran A Suresh Kumar

Abstract

Convergence of technologies one of the most significant features of the last decades of the twentieth century. The Internet, which is perhaps the most exciting development of today, is a classic example of convergence. It represents the coming together of computer and communications. The emergence of Internet and Multimedia technology has influenced considerably the handling, storage, retrieval and dissemination of information. The study is a virtual network of automated library systems within a specified area. The main purpose for the system is for a unified virtual library system that would enable dissemination of knowledge not only to a particular cader of people but also to the entire society as a whole. The study is a proposal to form a consortium among college libraries in Trichy city. The source code is given in HTML.

Dr. B. S. Swaroop Rani, Reader, DLIS, Bishop Heber College, Trichy, Tamil Nadu.
E-mail : sirishree_b@yahoo.co.in
Mr. N. Prabakaran, Student, DLIS, Bishop Heber College, Trichy , Tamil Nadu.
E-mail : prabakarnrk_m@yahoo.co.in
Mr. A. Suresh Kumar, Student, DLIS, Bishop Heber College, Trichy, Tamil Nadu.
E-mail : mervyn_serine@yahoo.co.in

 © **INFLIBNET** Centre, Ahmedabad

Digital Libraries in Modern Era

Prashant Bopapurkar Anamika Shukla Manju Shukla

Abstract

Digital Library is emerging as a new field of research for already existing fields like Library Science, Computer Science, Database management, Network Management, Web Technology, System analysis, System Design etc. It is actually an outcome of combined efforts of all these fields and so, is having a wide domain of discipline and research issues as its background. This paper presents a general system architecture for Digital Library, Design and research issues associated with it and demonstrates impact of these issues on Digital Libraries. An attempt has been made to provide general guidelines for planning a digital library. It also provides a framework for thinking about the field of Digital Library and analyses Digital Library by speculating that the Digital Library can be modeled as modern library.

Dr. Prashant Bopapurkar, Librarian, Kamla Nehru College, Korba,Chattisgarh.
Ms. Anamika Shukla, HOD, Asst. Professor (CS) Kamla Nehru College, Korba, Chattisgarh.
Ms. Manju Shukla, Visiting Lecturer, School of Pure and Applied Physics, GGDU, Bilaspur, Chattisgarh,

Design and Development of Consortia for Special Libraries in Kerala

M Bavakutty Mohamed Haneefa K

Abstract

Libraries are now facing severe economic constraints. They are encountering budget-cut coupled with the escalating costs of resources and diverse information needs and growing expectations of users. Special libraries in Kerala are no exception to this phenomenon. Joining or forming consortia can solve many of the problems faced by these libraries. This paper evaluates the existing resources and infrastructure of special libraries in Kerala, affirms the perceived need for consortia, discusses the requirements and benefits of consortia and propose plan for the establishment of consortia for special libraries in Kerala.

Prof. M Bavakutty, Former University Librarian, Professor and Head, Department of Library and Information Science, University of Calicut, Kerala
E-mail : bavakuttym@yahoo.co.in
Mr. Mohamed Haneefa K, Senior Research Fellow, Department of Library and Information Science, University of Calicut, Kerala
E-mail : haneefcalicut@yahoo.com

 © **INFLIBNET** Centre, Ahmedabad

Exploring the Necessities of an "E-journal Consortia" for the Users' of Academic Institution – A Project Plan

Bidhan Ch. Biswas Swapan K Dasgupta Mokbul Rahaman

Scholarly and research journals have long suffered and journal intake in academic libraries have declined to a level far lower than the normal needs of academic users, due to shrinking library budgets. This is a common global phenomena. A recent article in Library Journal reports that journal prices have raised by 215%. Research libraries have failed to procure the minimum needs for their academic users. The situation in India is compounded by continuous decline in the value of Rupee in foreign trade. The resultant effect is a steep decline in the library resources of Indian libraries causing a very wide gap between what is available and what is needed and affordable. So that in the academic community is forced to cancel substantially or completely their subscriptions for foreign journal. This causes information gap, which leads to knowledge-gap in the developing worlds like India, which is not desirable, but e-journal consortium can counter this crisis countered by inflationary trends. This article aims to develop a project plan and also to initiate an e-journal-consortia and to support the safeguarding of existing academic crisis by sharing resources for academic community and also to attempt to provide some solutions in this context.

Mr. Bidhan Ch. Biswas, Lecturer (Sr.), Dept. of Lib. & Inf. Sc., University of Kalyani
E-mail : bidhan_kly@yahoo.com
Mr. Swapan K. Dasgupta, In-charge, Internet Centre, University of Kalyani
E-mail : dasgupta_swapan@yahoo.com
Mr. Mokbul Rahaman, MLISc (student), Dept. of Lib. & Inf. Sc., University of Kalyani

Utilization of UGC-Infonet E-Journal Consortium An appraisal in the context of NAAC Accreditation of Universities in India

Nasirudheen T P O M Bavakutty

Abstract

The E-Journals Consortium is a corner stone of UGC-Infonet programme, which aims at addressing the teaching, learning, research, connectivity and governance requirements of the universities in India. The paper examines the utilisation of this ambitious project by the universities. Also check up whether there is any correlation between the use of the consortium and the level of accreditation of the universities by the NAAC.

Dr. Nasirudheen T P O, Senior Lecturer, Dept. of Library & Inf. Science, Farook College, Calicut, Kerala
Prof. M Bavakutty, Former University Librarian, Professor & Head, University of Calicut, Kerala
E-mail : bavakuttym @yahoo.co.in

 © **INFLIBNET** Centre, Ahmedabad

Environment of Digital Libraries

D Prabhavathi
K Kanthimathi

D Prabhavathi
K Muthu Chidambaram

Abstract

The term digital library has been used to characterize a large storehouse of digital information accessible through computer. Like a traditional library, a digital library serves as an archive of knowledge that spans many topics. With digital technology, images are used to reproduce rare items, allowing for virtually universal copying, distribution, and access. The technology also makes it possible to bring collections of disparate holdings together in digital form, making resource sharing more feasible.

Mrs. D. Prabhavathi, Assistant Professor, University Library, Sri Padmavathi Mahila Viswavidyalayam, Tirupati, Andhra Pradesh.
E-mail : prabha_dg@yahoo.com
Mrs. D. Prabhavathi, T.T.D. Plot No. 8, D. No. 2 - 182, Near Veterinary Hospital, M. R. Palle, Tirupati, Andhra Pradesh
Dr. (Mrs.) K. Kanthimathi, Senior Grade Librarian, Rani Anna Govt. Degree College for Women, Tirunelveli, Tamil Nadu
Dr. (Mrs.) K. Muthu Chidambaram, Reader in Sociology, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu.
E-mail : muchid9@yahoo.com

Intelligent Software Agents for Library Applications

Minoo Lohani
V K J Jeevan

Abstract

This paper presents a brief outline of the potential applications of intelligent agent technology in libraries with a list of illustrative examples.

Ms. Minoo Lohani, Professional Trainee, Central Library, Indian Institute of Technology, Kharagpur.
Mr. V K J Jeevan, Assistant Librarian, Central Library, Indian Institute of Technology, Kharagpur.

 © **INFLIBNET** Centre, Ahmedabad

Establishing Institutional Digital Repositories in An Academic Environment

R M Vatnal

Abstract

Digital formats of Institutional Publications can very well constitute a repository. Creation of Digital repository is an information activity. Creation of Institutional Digital Repository (IDR) is an important development in the Open Archive Initiative (OAI). Library Professionals can comfortably manage the repositories. The success of these repositories lies in the cooperation of faculty, support from the university and active role of library. It will emerge as a powerful driving force in promoting virtual learning process. This article deals with concept of Institutional Digital Repositories, Qualities of Digital Documents, benefits of repositories, OAI and issues involved in setting up of an IDR.

Dr. R M Vatnal , Assistant Librarian, University Library, Karnatak University, Dharwad
E-mail : vatnal@yahoo.com

 © **INFLIBNET** Centre, Ahmedabad

AUTHOR INDEX

Accanoor, Kalyani 688
Ahuja, J P S 457
Anuvasanthi, M 630
Aradhya, Manjunath V N 94
Asemi, Asefeh 648
Bachalapur, M M 531
Balasubramanian, P 470, 526
Basu, Dipak Kumar 32
Bavakutty, M 744, 745
Bhalekar, Dipak Krushnarao 739
Bhardwaj, Raj Kumar 441
Bhaskaran, R 392
Bhole, Harsha Ravindra 739
Bista, Sanat Kumar 70
Biswas, Bidhan Ch. 745
Bopapurkar, Prashant 744
Chandrashekara, M 396
Chandwani, M 83
Chaubey, O. N. 743
Chaudhari, Anirban Ray 32
Chaudhari, Prasad Pandit 739
Chauhan, Suresh K 658
Chelatayakkot, Veerankutty 348
Chidambaram, K Muthu 746
Chitrajkumar R 146
Cholin, V S 658, 668
Chopra, H S 549
Choudhury, B K 580
Choukhande, Vaishali G. 428
Dange, Jitendra 428
Das Gupta, Indranil 132
Dasgupta, Swapan K. 745
Deepa, T 630
Deepak, P. 331
Deshmukh, Prashant P. 739
Dhanasegaran, G 737
Dube, Sonia 128
Duraiswamy, K 50, 77, 342
Durrani, Omar Khan 22
Fatima, S. Sameen 117
Francis, A. T. 497
Gaddagimath, R B 539
Ganesan, P 738
Gangadharan, N 146
Gonsai, Atul M 484
Gopal Reddy, G 712
Gopalakrishnan, S 727
Goswamee, Bhupen 738
Guru, D S 377
Haneefa K., Mohamed 679, 744
Hemalatha, R 290, 300
Hemamalini, R 290
Hemamathi, R 300
Hemantha Kumar, G 94, 377
Hussain, K H 146
Idicula, Sumam Mary 8
Ingle, Chitra 720
Ingle, Maya 83
Jalaja, V 348
Jange, Suresh 539
Jani, N N 484
Jayabalan, E 236, 244, 253
Jayaprakash, H 531
Jeevan, V K J 221, 746
Jotwani, Daulat 612
Kamalakkannan, P 50, 621
Kannappanavar, B. U. 409
Kanthimathi, K 746
Kanthimathi, S 470, 526, 736
Karthikeyani, V 50, 621,
Kaur, Amrit Pal 740
Kaza, Padmini 553
Kazi, Mostak Gausul Hoq 214
Keshari, Birendra 70
Khan, M T M 697
Krishnamurthy, M 404
Krishnan, A 236, 244, 253, 290, 300, 621
Krishnan, R 117
Kuffalikar, Chitra Rekha 158
Kumar, S 358
Kumbar, Mallinatha 517
Kumbargoudar, P K 712
Lalitha, P 420
Lapp, Erda 589
Lohani, Minoo 746
Lohar, Manjunath S 517
Madhu, K. N. 409
Madhuri Devi, Th. 737
Maharana, B 580
Maheta, Mahendra 740
Majumdar, Kamalendu 197, 736
Mamata, P. K. 712
Manikanadan, S. 742
Manoharan, A. 630
Manoj Kumar, K 679
Mini Devi, B 325
Modi, Nileshkumar K. 43, 109

Mulla, K R 396
Munshi, M. Nasiruddin 214
Murthy, T A V 128, 230, 420, 635, 658, 668, 697
Nagabhushan, P 377
Naik, Ramesh R. 249
Naik, Umesha 594
Nair, G Hemachandran 575
Nair, Shivashankar B 1
Nasipuri, Mita 32
Nasirudheen. T. P. O. 745
Naushad Ali, P M 735
Nazi, Mohd. 209
Nessa, Najmun 132
Nisha, Faizul 209
Noushath, S 94
Om Vikas 271
Parameswaran, Sandeep 331
Patel, Yatrik 128
Pathak, Nityananda 738
Pathy, S K 580
Patil, Raghavendra J. 658
Paulraj, K 470, 526
Peter S, David 8
Pianos, Tamara 448
Policegoudar, S B 539
Prabakaran, N 743
Prabhavathi, D 746
Pradhan, D K 580
Prakash, K 668
Prakash, Pravish 741
Prakashe, Veena A. 62
Prasad, Sandhya 259
Pugazendi, R 236, 244, 253
Rahaman, Mokbul 745
Rajeev, J S 146
Rajendran, L 736
Rajendran, P 727
Rajesh Kumar 743
Rajeswari, D 705
Rajyalakshmi, D 158
Rama Devi, T 606
Ramana, Y. V. 370
Ramesh Babu, B 727
Ramshirish, M. 318
Rathinasabapathy, G 414
Rathod, V R 43, 109
Rawtani, M. R. 457
Reshmy, K R 259
Sahoo, Kshyanaprava 221
Sahu, Hemant Kumar 475
Samyuktha, R 565
Sangeetha, G 742
Sankar, P 742
Satyabati Devi, Th. 230
Senthamarai, C 290
Shah, Leena 358
Shah, Mukesh Kumar 358
Shah, S M 43, 109
Shantharajah, S P 342
Shanthi, N 77
Sharma, G K 432
Sharma, J C 192
Sharma, S K 432
Sharma, Sumati 178
Sherikar, Amruth 539
Shet, K C 22
Shivakumara, A S 396
Shivakumara, P 94, 377
Shivalingaiah, D 594
Shrivastava, Mridulata 720
Shrivastava, P N 432
Shukla, Anamika 744
Shukla, Manju 744
Singh, Ch. Ibohal 737
Singh, Debnath 32
Singh, Mohinder 178
Singh, Prachi 559
Singh, R K 741
Singh, Rajesh Kumar 741
Singh, Th. Shyam 737
Singh, U. N. 197
Singh, Yogendra 635
Soni, Nilesh B. 484
Srivatsa, S K 259
Sumam, Aparajita 186
Suresh Kumar, A 743
Suriya, M 742
Swaroop Rani, B S 743
Tripathi, Rajesh Kumar 741
Vaidya, Sachin Vishwanath 741
Vaishnav, Ashwini A. 735
Varma, Vasudev 167
Vatnal, R M 747
Verma, Keshri 309
Verma, Manoj Kumar 736
Verma, Pratibha 736
Vijayakumar, J K 697
Vijayakumar, M 409
Vyas, O. P. 309
Yadav, R T 505

KEYWORD INDEX

Academic Library 517, 648
Access Control 679
Access Point 484
Aggregators 668
Agribusiness 457
Agricultural Resources 457
Archives 158
Area Study 158
Authentication 342
Automated Language Processing 117
Ayurveda System 575
Braille Script 22
Braille Translation 22
CC Mine 290
Character Recognition 83, 94
Column Block matching 377
Common Message Platform 167
Computer Security 342
Consortia 497, 531, 635, 658
Consortia Model 635
Content Management 167, 192, 209, 230, 325,
Content Management System 167, 192
Contour Detection 94
Copyrights 253, 697
Cryptographic modules 342
Cryptography 342
Data Communication 621
Data Mining 259, 290, 300, 309
Data Structure 300
Database Systems 290
Desktop Publishing 32, 146
Digital Divide 22
Digital Information Services 414, 688
Digital Library 197, 209,271, 325 396, 414, 448, 526, 539, 549, 553, 559, 565, 575, 589, 594, 606, 612, 630, 688, 720
Digital Mapping 158
Digital Preservation 358, 370, 420
Digital Reference Service 553
Digital Resources 517, 648
Digital Signature 342
Digitization 158,325, 358, 370, 414, 428, 727
Disambiguation 83
Dissertations 414
Distance Measure 94
Document Image Mosaicing 377
Document Reconstruction 32
D-space 197, 594
E-learning 244, 549, 559, 606
Emulation 420
Encapsulation 420
Encoding standard 32
E-Print Archives 580
E-books 727
E-resources 221,404, 531, 635, 648, 658, 668, 705
E-theses 697
Evaluation of E-resources 221
Faceted Metadata Language 186
Frequent Pattern Approach 309
Friend Agents 392
Friend Network 392
Ganesha 594
Globalization 128
Graphic Analysis 50
Greenstone 594
Handwriting Recognition 77
Image Processing 50
INDEST 221
Indian Languages 1, 8, 22, 32, 77
Indian Library Consortium 497
Indian Scripts 32
Indic Scripts 132
Information and Data Security 470
Information Management 249, 271
Information Retrieval 8, 178, 259, 348, 392
Information Services 441, 448, 505, 526, 688, 705
Internationalization 128
Internet 712
Internet Resources 178
Internet Security 679
Invariant features 94

IPR 253, 697
Iran 648
IT Based Services 539
Knowledge Management 214, 249, 271, 612
Legal Metadata 441
Legal Text Retrieval 441
Library Automation Software 132
Library Automation 539
Library Consortia Model 497
Library Home Page 475
Library Networking 470, 497
Library Security 470
Library Services 475
Localization 128, 132, 146, 318
Machine Translation 70
Manuscripts 230, 358, 370
MARTIF 318
Media Obsolescence 420
Medical Information System 575
Metadata 420
Metadata Information Architecture 186
Migration 420
Mobile Agents 392
Model Driven Architecture 43
Multilingual Computing 1, 8, 128, 146
Multimedia 236
Multimedia Encryption 253
Multimedia Integration Language 236, 244
Natural Language Processing 1, 8, 43, 62, 83, 109, 117
Nepali Language 70
Network 484, 621
Network Security 679
Networking 539, 712
OAIS 420
OCR 22, 32, 94
OLIF 318
Ontology 259
OPAC 348
Open Archive Initiative 580
Open Source Software 594
Overlapping Region 377
Page Layout Analysis 32
Patents 178
Pixel Value Matching 377
Portal 404, 448, 475, 505, 565, 589, 612, 668
Preservation 158, 396, 428
Real Time Multimedia 244
Re-engineering 497
Reference Service 553
Repositories 580
Research Library 409
Route Discovery 621
Routing Protocol 621
Scholarly Communication 580
Search Engines 331
Security Policy 679
Security Risks 679
Semantic Web 259
Statistical Parsing 83
Streaming Media 236, 244
Subject Gateways 457, 505, 565
Survey 517, 635
Telecommunication 712
Terminology Database 318
Terminology Interchange 318
Texture Analysis 50
Texture Synthesis 50
Theses 414
UGC – Infonet 658
Unicode 128, 132, 146
Universal Networking Language 70
Universal Virtual Computer 420
University Libraries 497, 589
Visually Disabled 22
Watermarking 253
Wavelet Decomposition 377
Web OPAC 348
Web search Interface 331
Web Services 432
Wireless Network 484, 621
Wireless Security 484
WS-Security 432
XFML 186
XML 420, 432
XML-Encryption 432
XML-Signature 432