DIGITAL LIBRARIES: THE ROAD AHEAD

SHALINI R URS

Abstract

The field of digital libraries (DL) has been a product of as well as an active participant and a catalyst of the changing times. Coinciding with the decade of internet, the decade old field of digital libraries has emerged as one of the fast developing and continually evolving and most transformative and consequential domains of study, research and education. The field has opened up a new frontier of application of technology for information management. The quantity and variety of information available is increasing, the ease and convenience of disseminating and accessing information is being redefined all the time, mainly due to the strides made by the field of digital libraries. The author highlights the trends and futuristic of digital libraries.

Keywords: Digital Libraries

1. Preamble

The past decade (1995-2005) represents one of the fascinating and exciting times. It has also been a period of paradigmatic shifts in the world of scholarship and learning, business and management, governments and society. The field of digital libraries (DL) has been a product of as well as an active participant and a catalyst of the changing times. Coinciding with the decade of INTERNET, the decade old field of DL has emerged as one of the fast developing and continually evolving and most transformative and consequential domains of study, research and education. The field caught the attention and imagination of policy makers, researchers and educators alike. Digital libraries have evolved into systems; tools and technologies to translate the Utopian dream of universal access to information and knowledge a distinct possibility and a reality. The field has opened up a new frontier of application of technology for information management. The quantity and variety of information available is increasing, the ease and convenience of disseminating and accessing information is being redefined all the time, mainly due to the strides made by the field of digital libraries

2. Digital libraries – early visions and conceptualizations

Though the field of digital libraries that we are all familiar with my be only a decade old, early conceptualizations are traced to the early visions of ‘universal access to knowledge’ such as the often cited HG Wells’ essay on the future of world education - World brain: the idea of a permanent encyclopedia’ published in 1937 where in he
outlined the core of such an institution as world synthesis of bibliography and documentation with the indexed archives of the world... A great number of workers would be engaged perpetually in perfecting this index of human knowledge and keeping it up to date. Today we are getting closer to those visions – with one important difference it is not ‘a great number of workers’ who are engaged in the tasks of indexing but the machines. Another most important early figure but less quoted is Paul Otlet who struggled tirelessly for decades with the central technical, theoretical, and organizational aspects of the problem central to society - making recorded knowledge available to those who need it. The classic ‘Libraries of the future’ by J.C.R. Licklider in 1965 may perhaps be considered as a watershed in the transition from physical libraries to digital libraries.

While these thinkers were envisioning the idea of ‘universal access to knowledge’ from the perspective of ‘access’ there were others who were dreaming of systems that would enable and enhance human memory through a personalized and interwoven store of information... Vannevar Bush's concept of Memex in the 1930s, a microfilm-based “device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility.”. This conceptualization is a significant milestone in the development of digital libraries, as it inspired Ted Nelson and Douglas Engelbart, in their independent formulations of the various ideas that became hypertext – the technology that is at the core of the Web.

3. Technical evolution

In terms of technical evolution, one can trace the roots of digital libraries to the Information Retrieval Systems outlined/developed by Perry, Kent, Taube and others. The automatic indexing and search system developed by Salton, has not only paved the way for the more sophisticated IRS of the seventies and early eighties, which in turn led to the online systems of the eighties. The intellectual lineage of digital libraries is firmly rooted in the field of information retrieval systems. However, digital libraries as we know them today were conceived and built only since the 1990s. Digital libraries, also called “electronic libraries” suddenly changed from the relatively obscure concern of a few people in computer science and the library profession to a very popular topic of many research groups. The emergence of INTERNET as a new medium of not only the scientists and scholarship but also of the society at large during the nineties engendered the birth and fuelled the growth of the field of digital libraries. The credit for coining and firmly entrenching the phrase digital libraries goes to Tom Kalil and Mike Nelson, President Clinton and Vice President Gore advisers who liked the idea of access to information. The emergence and growth of the Internet and high bandwidth connectivity, combined with low cost processors and memory, have encouraged the creation and use of digital content on a grand scale. The situation today is one of information-driven endeavors spanning broad areas of human activities. Commercial, academic and public interests were fueled by US government interest including that of Vice President Al Gore, under the National Information Infrastructure label. The last decade has been marked by an explosion of interest, research, and development in the
area of digital libraries. Much of the impetus for this surge of interest stems from the large scale funding made available by the Digital Library Initiative 1 and 2 (http://www.dli2.nsf.gov/). These US initiatives have spawned a global response and ripple effect across the academia industry, and Governments thus thrusting the developments further and afar. The field spans diverse areas including all manner of and new forms of information systems and institutions; all forms of digital media and content; new means of, tools and technologies for creating, retrieving, accessing, sharing information.

4. The Technologies behind digital libraries

Many technologies have been responsible for translating the dream of universal access to knowledge to reality. These different technologies may be categorized as – computing, networking and presentation technologies. The great leaps in the computational and storage capabilities of the digital computers are the key drivers behind digital libraries. Most components of digital computers seem to follow the Moore’s law of doubling every 18 months in terms of exponential growth in performance. While increasing processing power helped handle the large volumes, the multimedia capabilities not only enable us to store, manipulate, and display images, sound, and video, but also make digital library content more expressive and rich with regard to interactions. The secure languages have enabled transactions that maintain privacy and security. Advances in database management have resulted in multiple/alternative approaches to the design and development of DLs. Coupled with advances in computational technologies, the evolution of high-speed networking, the birth and development of the Internet, open protocols, and the ubiquity of TCP/IP - all have helped enormously in advancing global access to the information stored in disparate computers. With the ever expanding bandwidth and low cost of access, digital libraries have been democratized. The digital divide has perhaps is replaced by digital dividends. What made digital libraries usable were the presentation technologies. The flexibility and diversity of presenting information to the user in myriad ways is perhaps what makes digital libraries popular, usable, and user-friendly. Dramatic improvements in the world of encoding of ideas, information, and images as well as their presentation of the same to the user were responsible for the increasing acceptance of digital libraries. Developments in mark up languages, e.g., SGML, HTML, and XML; as well as their presentation methods, e.g., hypertext, VR, graphics, sonification, multimedia, document interchange, word/desktop processing & publishing, and scholarly publishing are all part of digital library evolution.

5. Organizational and Societal framework of digital libraries

Technological innovations are propelled by the social forces. Ideas/technologies will not survive without the emergence of appropriate institutional/organizational frameworks. Several institutions and organizations have steered the course of digital libraries. The W3C has been largely responsible for steering the course of the WWW and agencies like the Council on Library and Information Resources (CLIR), Corporation for National Research Initiatives (CNRI), Digital Library Federation (DLF), Coalition for the Networked Information (CNI – http://www.cni.org), and OCLC (http://www.oclc.org) have all shaped the course of development of DLs.
In the UK the Electronic Libraries Programme (eLib) of JISC (Joint Information Systems Committee) initiated in 1994 fostered the development of areas such as digitization, digital library preservation and electronic document delivery. The JISC has played a critical role in providing the framework for digital library movement in the US. Instrumental to the promotion of the digital library research and development in Europe is the creation of the DELOS Network of Excellence on Digital Libraries. DELOS is an initiative funded by the European Commission’s Information Society Technologies 5th Framework Programme (IST-FPS). And currently the EU IST programme is the major framework for funding and funneling the development of the field of digital libraries.

6. **Digital Library Journals**

- D- Lib magazine published by the Corporation for National Research Initiatives (CNRI) and supported by Defense Advanced Research Project Agency (DARPA) and NSF
- Journal of Digital Information – a peer reviewed journal publishing papers on the management, presentation and uses of information in digital environments, supported by Texas A & M university libraries
- International Journal on Digital Libraries, a quarterly published by Springer
- ARADNE a quarterly journal published by the UKOLN based at University of Bath and supported by the Joint Information Systems Committee (JISC) and Museums, Libraries and Archives Council, UK

7. **The first decennial and the success stories**

Some examples of success stories/outcomes of first decade of DL research:

- The Google search engine (www.google.com), is a spin off of the research projects at the Stanford University database group, one of the six DLI-I projects
- The Alexandria Digital Library (ADL) (http://www.alexandria.ucsb.edu/) which is includes a distributed digital library with collections of location indexed/georeferenced materials.
- Image searching and retrieval systems at the University of California at Berkeley (http://elib.cs.berkeley.edu/vision.html).
- Funding and support from foundations such as Mellon Foundation and others have also resulted in exemplary digital collections, and innovative tools and technologies. Jstor (www.jstor.org) and ArtSor (www.artstor.org) are examples of redefining digitally enabled scholarship.
New genres of documents and collaborative authoring systems are another outcome of the DL research. Mention may be made of Wikipedia (http://en.wikipedia.org/wiki/) a multilingual Web-based free-content encyclopedia that is written collaboratively by volunteers, allowing articles to be changed by anyone with access to a web browser.

Innovative systems and models for preservation and archiving of digital content and collections. Some global exemplars of Web Archiving Projects include the Internet Archive (http://www.archive.org/), the PANDORA (http://pandora.nla.gov.au/) and others are some of the successful endeavours. Collaborative models for Journal archiving such as LOCKSS (http://lockss.stanford.edu/)

Creating ripples in the world of scholarly communication system by way of engendering a new scholarly communication model, popularly known as Open Access Publishing including Open Access Initiative, Institutional repositories.

Development of DL software such as Greenstone (www.greenstone.org), DSpace (www.dspace.org), Fedora (http://www.fedora.info/) and others

Standards, Tools and frameworks such as Handle System (http://www.handle.net/); Digital Object Identifier System (www.doi.org); Open URL (http://www.niso.org/committees/committee_ax.html); Dublin Core (www.dublincore.org); OAI-PMH (Open Archives Initiative Protocol for Metadata harvesting) (www.openarchives.org/OAI/openarchivesprotocol.html)

Commercial organisations have also responded positively by developing new formats and models for digital content and collections. New partnerships and alliances between organizations have also emerged.

8. The Indian digital library scenario

The digital library developments in India may be categorized under five broad classes –digital library conferences and events; the digitization and digital collections building projects; the consortia based approach to enabling access to electronic journals and other digital resources; and the Open Access and institutional repository movements in India.

Indian DL Conferences: One of the early conferences in India that brought together individuals and agencies involved in the what may be called as the Pre Web digital content issues was the InfoTex 93 - An International Conference on Database Production and Distribution: Resources, Technology and Management, held at Bangalore, 28 Nov - 1 Dec 93 organized by Informatics India Limited. The first digital library conference of India was the SIS 96 – the 15th Annual Convention and Conference, 18-20 Jan 1996, Bangalore organized by the Society for Information Sciences. The major conferences that following the web and the Internet era and events which created the buzz around the field of digital libraries were the ICADL 2001 held in Bangalore(www.icadl2001.org) and the ICDL (International Conference on Digital Libraries) organized by TERI (the Energy Research Institute) during 2004 and 2006.
The digitization and digital library projects: The DL projects that have been able to build a noteworthy test bed of collections include the Digital Library of India (DLI) Initiative that is the Indian part of the Universal Digital Library and the Million Books to the Web projects. The Million Books to the Web project is a collaborative project between the Carnegie Mellon University, Pittsburgh and the Indian Institute of Science, Bangalore and many more organizations across the US, India and China. The Traditional Knowledge Digital Library (TKDL) is a knowledge repository of the traditional knowledge setup by the NISCAIR (National Institute of Science Communication and Information Resources, CSIR (Council of Scientific Research) Government of India, the objective of this library is to protect the ancient and traditional knowledge of the country from exploitation such as bio-piracy and unethical patents (http://www.niscair.res.in/). The Muktabodha Digital Library and Archiving Project was begun in 1995 as a manuscript microfilming project focusing mainly on photographing at-risk and rare palm-leaf Vedic Shruti Ritual manuscripts from both private collections and from libraries. (http://www.muktabodha.org/). Kalasampada: Digital Library - Resource for Indian Cultural Heritage (DL-RICH) project sponsored by Ministry of Communication and Information Technology (MCIT) (http://www.ignca.gov.in/drich/) aims to use multimedia computer technology to develop a software package that integrates variety of cultural information and helps the users to interact and explore the subject available in image, audio, text, graphics, animation and video on a computer in a non-linear mode. One of the enduring digital library projects from a University is the Vidyanidhi digital library and eScholarship portal project at the University of Mysore (www.vidyanidhi.org.in), which began as a pilot project in 2000 to demonstrate the feasibility of Electronic Theses and Dissertations (ETD) in India with funding from NISSAT (National Information System for Science and Technology), DSIR, Government of India. Today Vidyanidhi funded by the Ford Foundation is one of the largest repositories of Indian theses with more than 5000 full texts and 100,000 metadata records.

Access to E-Journals: The efforts and initiatives to enhance access to electronic journals and digital resources to the academic community in India has resulted in two major consortia based programmes – the Indian National Library in Engineering Science and Technology Consortium (INDEST), IIT-Delhi-http://www.indest.iitd.ac.in. The Ministry of Human Resource Development (MHRD) funded project provides for subscription to electronic resources for 37 institutions including IISc, IITs, NITs, IIMs and a few other centrally-funded Government institutions through the consortium headquarters set-up at the IIT Delhi. The other national level consortia is the UGC – Infonet Programme spearheaded by the INFLIBNET (Information and Library Network, a University Grants Commission’s inter university centre. Under this programme universities across India have been enabled to build information infrastructure for accessing the electronic resources subscribed by the consortia (http://www.inflibnet.ac.in/).

The Open Access movement in India: OA movement has been a very active one with many proponents, training programmes and initiatives. One of the Institutional repositories is the eprints@iisc (http://eprints.iisc.ernet.in/) at the Indian Institute of Science, Bangalore which currently has more than 6000 publications. The National Centre for Science Information (NCSI) at the IISc has been organizing many training
programmes as well. OpenMED (http://openmed.nic.in/) is an Open Access Archive in area of Medical and Allied Sciences including Bio-Medical, Medical Informatics, Dental, Nursing and Pharmaceutical Sciences. It is international in scope and includes both published (post-prints) and unpublished (pre-prints) documents having relevance to research in these disciplines. It is hosted by Bibliographic Informatics Division of National Informatics Centre (India). Academics of Science in India are also supporting the Open Access movement by making journals available online free. Indian Academy of Sciences and its journals are available open access and are available free and full text is available as PDF files on each journal's website. Indian National Science Academy is also makes its journals available online and is free to access. One significant contribution from the private sector to the Open access movement in India is the Open J-Gate from Informatics India Limited, Bangalore (www.openjgate.org). Open J-Gate is an electronic gateway to global open access journal literature. Launched in 2006, Open J-Gate provides seamless access to millions of journal articles available online. Open J-Gate is also a database of journal literature, indexed from 3000+ open access journals, with links to full text at Publisher sites.

10. Issues and Challenges of the first decennial

The euphoria and the hype of the digital library enthusiasts during the early nineties were replaced by the realization that building digital libraries is more than going online with a website. Creating effective digital libraries poses serious challenges. The integration of digital media into traditional collections is not straightforward, like previous new media (e.g., video and audio tapes), because of the unique nature of digital information—it is less fixed, easily copied, and remotely accessible by multiple users simultaneously. Some of the more serious issues facing the development of digital libraries the first decade of digital libraries maybe summarized under the following –

The Technical Issues: The technical issues centre on digital library infrastructure and architecture such as the following –

- high-speed local networks and fast connections to the Internet
- relational databases that support a variety of digital formats including audio and video
- full text search engines to index and provide access to resources
- a variety of servers, such as Web servers and FTP servers
- Electronic document management functions that will aid in the overall management of digital resources including access control and authentication services.

Content and collections issues: One of the largest issues in creating digital libraries is the building of digital collections. Obviously, for any digital library to be viable, it must eventually have a digital collection with the critical mass to make it truly useful. There are essentially three methods of building digital collections - Digitization - converting paper and other media in existing collections to digital form; Subscription
to original digital resources - created by publishers and scholars. Example items would be electronic books, journals, and datasets; Access to external materials by providing pointers to Web sites, other library collections, or publishers’ servers.

**Digitization** : The primary methods of digital collection building are digitization. Digitization is the process of conversion of any fixed or analogue media—such as books, journal articles, photos, paintings, microforms—into electronic form through scanning, sampling, other means. An obvious obstacle to digitization is that it is very time consuming and expensive. In fact some of the early digital library efforts began with digitations projects. For example the Mercury Electronic Library Project at the Carnegie Mellon University was an early attempt to establish a campus based digital library of journal articles in computer science. It was followed by the CORE (Chemistry Online Retrieval Experiment) involving Cornell University and OCLS, American Chemical Society and others. The JStor is one such example of building a huge archive of journals. The first challenge is how do we go about selecting materials for digitization?

- Retrospective conversion of all of the collections - essentially, starting at the beginning and ending up with the last item.
- Digitization of a particular special collection or a selected one
- Showcase collection - digitizing particularly good examples of some collection strength
- Needed collection – digitizing those materials that are in most demand more accessible.
- On demand approach - where one digitizes and stores materials as they are requested.

In addition to these strategies the basis of selection would include the criteria such as their potential for long-term use; their intellectual or cultural value; enabling or enhancing access than possible with original materials (e.g., fragile, rare materials); copyright restrictions or licensing issues.

The digitations issues such as the standards for digitization, tools for effective means of web enabling the digitized materials, Optical Character Recognition (OCR) issues are critical to the digitization. The common practice followed by most digital library initiatives is to outsource the digitization work to some reliable vendors. However, it is important to realize that outsourcing also demands significant demands in terms of ensuring quality though stringent specifications, proper workflow and quality audit processes. There are many standards and guidelines available for specifying the standards for digitization. National Library of Australia has developed very useful guidelines [http://www.nla.gov.au/digital/standards.html](http://www.nla.gov.au/digital/standards.html).

When it comes to Indian Language content digitization, the major obstacle is the non availability of OCR software for Indian Languages. OCR Systems are software that let us convert scanned images containing texts into computer processable, editable files in a variety of formats such as ASCII, ISCII or UNICODE. A survey of the scenario
reveals that, we still have a long way to go in this area and there is tremendous scope for R & D. At this point there is just one complete OCR package for one of the Indian scripts – Devanagari (Hindi) and one for Telugu. Bangla – there is a good research and development effort but still no product. There is couple of prototype level systems for other languages – Assamese, Oriya, Gurumukhi, and Malayalam. There is one system by one firm for Tamil and nothing beyond some R & D efforts in respect of Kannada
Another major gap is that none of these systems are extensively tested and evaluated.
Given this background, there have been efforts by the Technology Development for Indian Languages (TDIL) division of the Ministry of Information Technology to establish a consortium to develop OCR for Indian Languages. Another recent effort is the setting up of Centre for Indian Language Content Management at the International School of Information Management (ISIM), Mysore (www.isim.ac.in) supported by the Rediff.com.

**Content Management Issues**: The content management issues include such factors as metadata and digital assets management issues such as persistent indenters and others. Metadata is important in digital libraries because it is the key to resource discovery and use of any document. Anyone who has used Google or any of the other search engines on the Internet knows that simple full-text searches don’t scale in a large usable collection management. One can get thousands of hits, but most of them will be irrelevant. While there are formal library standards for metadata such efforts are very time-consuming to create and require specially trained personnel. Human cataloguing, though superior, is just too labour extensive for the already large and rapidly expanding information environment. Naming, identifiers, and persistence and others are also related to to metadata. It is the problem of naming in a digital library. Names are strings that uniquely identify digital objects and are part of any document’s metadata. Names are as important in a digital library and are needed to uniquely identify digital objects for purposes such as: citations; information retrieval; linking and for the purposes of managing copyright. Any system of naming that is developed must be permanent, lasting indefinitely. This means, among other things, that the name can’t be bound up with a specific location. This is very much unlike URLs, the current method for identifying objects on the Internet. A global scheme of unique identifiers is required, one that has persistence beyond the life of the originating organization and that is not tied to specific locations or processes. These names must remain valid whenever documents are moved from one location to another, or are migrated from one storage medium to another. Three examples of schemes proposed to get around the problem of persistent naming are PURLs, URNs, and Digital Object Identifiers (DOI).

**Rights Management Issues**: Rights management has been dubbed as the most perplexing challenges of digital library environment. As libraries move from the physical medium to the digital, library staff are increasingly confronted with the challenges of addressing the copyright and other intellectual property rights issues related to digital information. Copyright has become a hot topic and a vexing issue for all those who have a stake in scholarship and scholarly communication. In the digital world, the very premises and philosophy of copyright are being questioned and voices are being heard to review its very tenets. The issue of rights ownership transgresses into the realm of hairsplitting issues of creativity, work for hire and other equally contentious
matters. In the world of scholarship and intellectual heritage, libraries play a very important role. Libraries are the voices for the 'public good'. But in the digital millennium how do we balance the often conflicting interests? The role that libraries play in the scholarly communication process is shaped by the provisions of the copyright. Copyright laws are an instrument of balancing the interests of creators and the societal obligations to facilitate the free flow of information. Safeguarding the private and public interests has been reduced to a win or lose situation. The Digital Millennium Copyright Act (DMCA) of 1998 in the US is one such example. Retaining the balance between the public and private concerns is the key to addressing the challenge of achieving and equilibrium of intellectual property rights.

Preservation: All civilizations have been concerned with the preservation of its artifacts—whether they were stone inscriptions or parchment or manuscripts—whether they were scholarly materials or government or business transactions. This issue is compounded in the case of digital materials as they are not 'human eye' readable. Digital Preservation is not a new concern— it has been with us since the introduction of computers into our lives in the sixties. Digital Preservation encompasses a broad range of activities. Activities that are aimed at extending the life and after life of digital materials. It includes digitization as a means of preserving fragile materials. Preservation of digital objects including converted cultural artifacts—such as books, journals and other materials; born digital materials—both 'published' and 'grey materials'; and 'digital grey materials' covers such genres as 'web sites' 'Weblogs' and others. In the context of technological obsolescence preservation of digital objects has become a major concern for the society.

11. The grand challenges and the road ahead

At the end of first decade of the digital libraries, many attempts were made to review the developments of the past and identify the future challenges to chart out a roadmap for the future. Clifford Lynch in a recent paper identifies the four compelling areas (http://www.dlib.org/dlib/july05/lynch/07lynch.html). They are—Personal information management; Long term relationships between humans and information collections and systems; Role of digital libraries, digital collections and other information services in supporting teaching, learning, and human development; Active environments for computer supported collaborative work offer the starting point for another research program. In another paper Lorcan Dempsey identifies many different areas and directions such as flat applications and liquid content; new social and service affordances; new business and organizational patterns; the new information hubs, the long tail and attention and others (http://www.ariadne.ac.uk/issue46/dempsey/intro.html ). The Larsen report of the NSF Workshop on research directions for the digital libraries held in 2003 identifies the following – expand what can be searched; use context for information retrieval; integrate information spaces into everyday life; Reduce data to actionable information; and Improve productivity through information access.

Based on the above I would summarize that the simple challenges of digital libraries for India would be to –
Build more and more digital content and collections (expand the content base) through digitization as well as creating and managing born digital materials through a sustained national programme.

Develop, adopt and adapt standards at all phases of life cycle of digital materials, whether born digital or digitized.

Evolve effective digital content management strategies including long term preservation.

The Grand challenges are –

- Personalization: Making digital libraries more personal. Develop middleware software for making searching and accessing information more effective and customizable. Life and libraries are moving into ‘personal spaces’ from public spaces. The digital libraries of the future are to be ‘my libraries’ kind.

- Managing convergence: the digital technologies and the communication technologies are witnessing unprecedented convergence. Ubiquitous INTERNET, mobile phones and smart hand held devices are demanding that digital libraries and collections are also move into ubiquitous access.

- Effective resource discovery: Managing unstructured digital information by developing tools and technologies for text mining, summarization, and structuring and automatically creating metadata.

The first decade of digital libraries has seen fantastic accomplishments and the future presents exciting possibilities and challenges.

References


Biography of Author

Dr. Shalini R. Urs is professor, DLIS, University of Mysore, International School of Information Management, Mysore, India and directs the Vidyanidhi: Indian Digital Library of Electronic Theses and Dissertations. She is spearheading the digital library movement in India through several initiatives. She serves on the Executive Board and the Steering Committee of NDLTD a global initiative based at Virginia Tech, USA. She served as coordinator of the UNESCO International Guide for ETDs. As Fulbright Scholar she was a visiting faculty at the Virginia Tech, USA during 2000-2001.

Email: shalini@isim.ac.in