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.

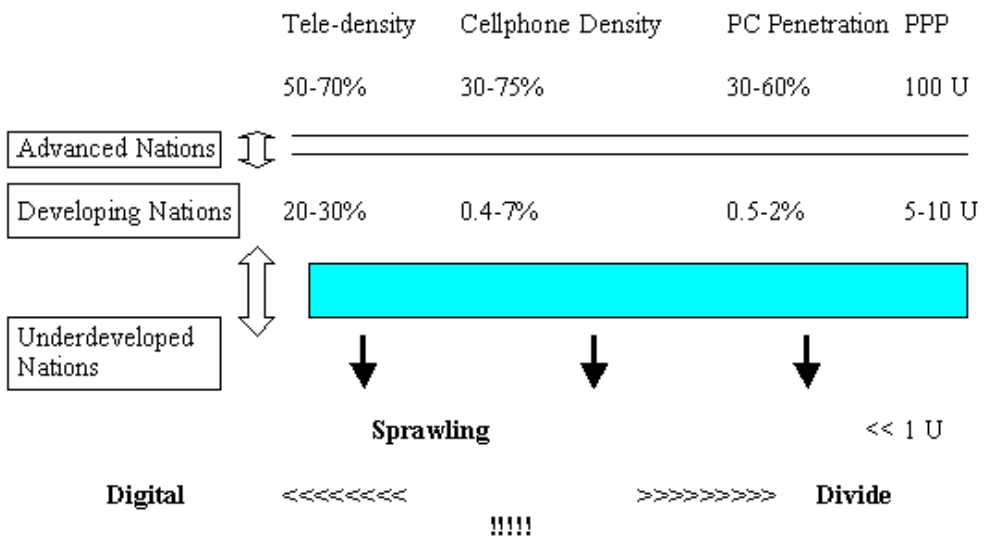
There is gradual collapse in authorship, quantity and quality of translation in other languages.

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