Digital Library

Meaning of a Digital Library is different for different people, organizations and communities. For a school student a Digital Library is collection of digital documents, database, video games and leaning materials accessible via computer network. For a space scientist collection may be available over Internet in the form of GIS and CAD data, satellite imagery, video gallery and so on so forth. For a farmer Digital library is a collection of information over a portal in the form of Government information, Land records, agriculture market information, etc. In a nutshell, Digital Library is a collection of information which is digitized, organized for a group of people or community, gives users power they had never with traditional libraries.

Pradip Kr. Upadhyay
Technical Director, NIC
pku@nic.in

A digital library is a library in which collections are stored in digital formats (as opposed to print, microform, or other media) and accessible by computers. The digital content may be stored locally, or accessed remotely via computer networks. A digital library is a type of information retrieval system

The DELOS Digital Library Reference Model *http://www.delos.info* defines a digital library as:

An organization, which might be virtual, that comprehensively collects, manages and preserves for the long term rich digital content, and offers to its user communities specialized functionality on that content, of measurable quality and according to codified policies.

Digital Library Standards

a. User Interface:
Common Web Browser

b. Data Handling and Interchange: Graphic Formats JPEG, TIFF, GIF, PNG, Group 4 Fax, CGM

Structured Documents HTML, XML, PDF

Moving Pictures/3-D MPEG, AVI, GIF89A, QuickTime, Real Video, ViviActive, VRML

c. Metadata: Resource Description Dublin Core, WHOIS++ Templates, METS, MODS, MARC, TEI Headers, Other Open Source and Domain Specific Standards, PREMIS (Preservation Metadata: Implementation Strategies)

Resource Identification URN, PURL, DOI, SICI

d. Search and retrieval: Federation and Harvesting: FTP-enabled, OAI-PMH for intermittently transfer data from one system to another

Federated search: Z39.50 protocol, SRW Protocol

e. Security, Authentication and payment services:

Emerging e-Commerce Standards

Major Characteristics of Digital Library

- Variety of digital information resources
- Digital Libraries Reduce the need for physical space
- Users at remote
- Users may build their own personal collections and space by the facilities provided by Digital Library
- Provide access to distributed information resources
- Same information resource can be shared by many at the same time
- Paradigm shift both in use and ownership
- Collection development be based on potential usefulness and appropriate filtering mechanisms be followed to negotiate the problem of plenty
- Ability to handle multilingual content
- Presupposes the absence of human intermediaries
- Should provide better searching and retrieval facilities
- Digital information can be used and viewed differently by different people
- Digital Library breaks the time, space and language barrier

Long-Term Retention

Once a document category is designated as a record, it must be assigned a retention period according to a retention schedule. A records retention schedule is a timetable that identifies the length of time a record must be retained in active or inactive status before final destruction. The schedule may be based on:

- Statutory and legal requirements, which must be researched and documented.
- Business continuity requirements that is, maintaining customer relationships or preserving intellectual property.
- The opinion of the company's chief legal officer, chief financial officer and chief executive officer.

Organizations maintaining an archive for very long periods of time (more than 20 years), have retention challenges that require special management processes and careful tracking of emerging trends to ease the cost of technological change. However, such concerns should not stop present initiatives and deployments. Digital preservation of content requires consideration of file formats, applications support and storage platforms, and the retention period. Adobe PDF/A is an example of an emerging standard that is being widely adopted by organizations as an alternative to traditional PDF and TIFF formats. Tracking industry groups and standards bodies that are working to provide consistent ways of storing content and its metadata will be important, but it will be equally important to understand the general market adoption and longevity of any standards. Storage technologies for archive and retention of content will also need to evolve to deliver more cost-effective storage, while providing a smooth transition to new technologies with minimal repository disruption.

Steps in Building Digital Library

Digital Content Management System is the most complex and advanced forms of information systems because it involves collaboration support, digital document preservation, distributed database management, hypertext, information filtering, information retrieval, instructional modules, intellectual property rights management, multimedia information services, question answering and reference services, resource discovery, and selective dissemination of information. Digital Library design should facilitate "one-stop-shop. There are various experiments, research and real deployment going on using Open Standards, Open Source Softwares and Open Technologies. But on the other hand, commercial players are also showing their presence using their tools and technologies. It is upto the organizations for choosing the technology platform and implementation strategies, because finally Total Cost of Ownership (TCO) matters.

a. **Digital Library System Design and Development:** Design of a digital library should facilitate variety of information resources residing on variety of computer systems in different parts of the world to a number of users of differing notions and needs. It should be One-Stop-Shop. Alternative to in-house development, there are a number of open source and commercial software solutions available. For example: Popular digital library open source softwares are Greenstone *(http://www.greenstone.org)*, Dspace *(www.dsapce.org)*, Fedora *(http://fedora.info/)*, Eprints *(http://software.eprints.org/)*. Popular commercial softwares may be Sharepoint from Microsoft, Websphere from IBM, ContentDM, LSPremia.

b. **Selection of materials :** Contents can include paper and electronic documents, audio, video, photos, images, e-mail and Web site content in other words, records are media independent. Enterprises cannot apply a general formula to decide what to keep and what not to keep.

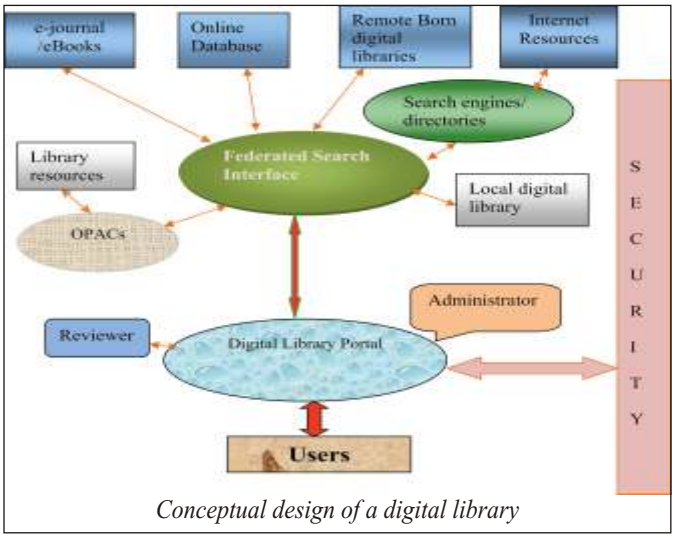
c. **Digitization:** Digital documents may be born-digital, created using digital publishing tools (e.g. Word, LaTex, DTP), or created by converting from an analogue format to digital format or converted from one digital format to another to suite the requirements of a particular Digital Library. The process of capturing and converting from analogue to digital format is often called as 'digitization' or 'digitalization'. The processes involve scanning, page layout analysis, image scanning.

d. **OCR (Optical Character Recognition):** This is another type of scanning for converting the document in computer readable and retrievable format. It is different from image scanning. Documents are converted into text and graphs and images. Alternative to OCR is manual typing, image files or combining images and OCR.

e. **Tagging and Metadata**

f. **Providing Access**

g. **Retention schedule and refreshing.**



Conceptual design of a digital library

Copyright

Many legal issues are associated with the development and use of digital libraries. Intellectual property rights are a major concern, others are authenticity of information resources, and privacy and security of users and institutions. Librarians, Publishers, lawyers and Governments over the years formulated regulations that control the intellectual property rights of owners of materials. Even then the materials may be free or very cheap; we should not take these content in developing our own Digital Library without the permission of the owner. Scanning the pages of copyrighted documents available in the libraries without permission and addressing legal issues are not permitted. Digitizing our own publications and contents and creating digital libraries is well within copyright. Sometimes many people download the licensed contents from Digital libraries and transmit them over Intranet and Internet without permission, that cause legal problems.

Organization of Information in Digital Libraries

The quantities of material stored in digital libraries pose the problem of finding what users need. Popular knowledge representation schemes in digital libraries are Library Classification, Indexing words and Thesauri, Metadata creation, faceted knowledge structures, Hypertext, Vector Models, XML and Semantic Web, Folkosonomies, Ontology and Simple Knowledge Organization Systems (SKOS).

The Manager is an Open Source Tool for creating and visualizing SKOS RDF vocabularies. The Manager facilitates the management of thesauri and other types of controlled vocabularies, such as taxonomies or classification schemes. The tool has been implemented in Java. *(http://thmanager.sourceforge.net/).*

Traditional cataloguing is an example of what is now called "Metadata". Metadata is a key component of the provision of online catalogues that are searchable across the Web. In order to use the Semantic Web to its best effect, metadata needs to be published in RDF formats. There are several initiatives involved with defining metadata standards in the library and publishing community, including: Dublin Core Metadata Initiative, MARC, ONIX, PRISM.

Future of Digital Libraries

The Future of Content Is Modular and Miniature. Semantic Web also termed as Web 3.0 might see its existence where computers will be capable of understanding information and performing tedious tasks such as finding, sharing, and combining on Web. There is lot of talk about Semantic Web, which is supposed to draw sense out of data in a meaningful and impactful manner. Also, semantic Web developers are releasing new XML formats, which supposedly would be the final step to the completion of Web 3.0 and hence Library 3.0.

Library 2.0 + Semantic Web + Security + Artificial Intelligence = Library 3.0

Particular kinds of documents in the future will be assembled actively by the consumer or via business process at the moment of need. Content producers will need to identify such document families, then develop related content as modular components that can be assembled in various ways. New content consumption models will drive content providers to produce granular, well-described content and to distribute it in unconventional ways.

For enabling knowledge connectivity in rural areas of India, we need to have a comprehensive plan for developing new infrastructure (viz., development of OCR Software in all the Indian languages, language independent operating system, database servers, search engines, web servers and messaging servers) for extending the digital library services in regional languages. This will enable the digital library initiative to percolate to the rural masses in the form of e-Governance, tele-education and tele-medicine. *i*