Digitization : Basic Concepts

B Mini Devi

Abstract

The introduction of digital libraries is changing not only the face but whole body of the libraries around the world. In a global village the concept of digital library is of great importance. Hence the process of digitization. The author discusses the steps involved in the process of digitization, types of materials to be digitized and the problems that the libraries are facing on this respect.

Keywords : Digitization, Digital Libraries, Content Management.

0. Introduction

The rumbling change that shook each and every field of library and information science is the emergence of digital libraries. In a country like India, rich in cultural, spiritual heritage indigenous research and development in science and technology, humanities, the preservation of her ancestral information sources becomes the duty of librarians and information scientists. The growth of internet, communication channels, information technology and digitization of documents are revolutionizing the traditional concept of library. Any user from any part of the world can access the information he wants in a digital environment.

1. Digital Library

According to the Berkeley Digital Library Project, University of California[1] , the digital library is be a collection of distributed information sources, producers of information make it available, and consumers find it perhaps through the help of automated agents”

The Information Infrastructure Technology and Applications (IITA) working Group considers “digital libraries as systems providing users with coherent access to a very large, organised repository of information and knowledge”[2]. According to Association of Research Libraries ‘digital libraries are not a single entity, require technology to link the resources of many, linkages are transparent to the users, permit universal access, not limited to document surrogates but extended to digital artifacts’. Some consider digital library as ‘library without books’ or ‘library without walls’. The synonyms to digital library are ‘virtual library’ or ‘electronic library’.

The purposes of digital library are[3]

- ✍ Collect, store and organise information and knowledge in digital form.
- ✍ Promote economic and efficient delivery of information.
- ✍ Put considerable investments in computing/communication infrastructure
- ✍ Strengthen communication and collaboration between research, business, government and educational communities.
- ✍ Contribute for lifelong learning opportunities

2. Materials to be digitized

The priority of materials to be digitized depends upon the type of library, category of users, finance, infrastructure, etc. However the following materials are to be given priority.

The scientists create a large quantity of information that are documented in the form of research reports. They are the authentic primary scientific data which is of great value to the coming generations. These reports form the basis of further research work.

2.1 Research Reports

As research publications are the result of original scientific investigations, the archiving of them needs utmost priority.

2.2 Journals

The articles published in journals in different subjects are to be traced and brought under one roof for archiving. In the field of Library and Information Science publishers from Asia work together as a group to publish their journals on major Website databases. As a result these journals get a wide publicity and usage. Recently Ulrich International Periodical Directory had selected 29 journals from India for digitalizing by taking into account of the standard and coverage of articles published by them.

2.2 Cultural heritage documents of the country

The preservation and archiving of the cultural heritage documents of our country are of great value to our future generations. Now-a-days this is one of the crucial issues which librarians of our country are facing. This may be due to the physical conditions of this type of documents, which are damaged or decayed due to various factors such as age, quality of paper, binding, insects, climate, handling etc.

2.3 Theses and dissertation

Theses and dissertations form a rare and valuable collection of many academic libraries. The digitization of them results in wide publicity and duplication of research can be avoided.

3. Digitizing

Digital libraries offer access to contents over computer and communication networks. In present days due to the escalating price, journals are out of reach to ordinary people. The paper used for printing these journals becomes brittle as time passes. In the case of books also the above situation exists. Moreover, the torn out, brittle and dusty documents create problems in maintenance and the health of librarians as well as the users. Here comes the importance of digitization which not only enhances the life of these documents but also provides easy access to wide audience with exhaustive search engines, and effective bibliographic control.

4. Methodology

- Content Searching and Selection
- Scanning
- PDF Creation & OCRing
- Content Indexing and Metadata
- Information Retrieval Procedure

4.1 Content Searching and Selection

The first step is to identify which materials are to be digitized and which are not to. The utmost priority should be given to the following factors

- ✍ policy of the library
- ✍ needs of the user
- ✍ type of document

4.1.1 Policy of the library

Now-a-days libraries give priority in digitization of their documents. In the case of a Science and Technology library, the digitization of research reports and journals should be done first because they are the result of original scientific investigation. On the other side in the case of a Social Science library, documents of historical importance should be given first preference. But in an academic library, research reports, theses, journals are to be selected first.

4.1.2 Needs of users

As users come from different cross section of the society, their needs are also varied. In a S & T library, scientists form major user community and their priority should be on research reports and journals. In a Social Science library, the user community constitutes mostly historians, literary writers and their importance will be on historical documents and literature books. But in an academic library, text books, journals, theses and dissertations are searched first by students, teachers and research scholars.

4.1.3 Type of document

Costly and rare books are to be protected from damage, multiliation and loss. Documents of old, cultural and historical materials are to be given prime importance- as they form part of history of our country.

4.2 Scanning

The fundamental conversion technique is scanning of the document. This can be done by sampling the image of the document on a grid of points. Each point is represented by a brightness code ie in black/ white colour. As in photographic work, a very high resolution is not needed in this case. Very good images can be created with a resolution of 300 dots per inch. The scanners take care of quality of paper of the document and spot-marks.

4.3 PDF Creation and OCRing

Portable Document Format (PDF) created by Adobe is a better format for storing page images in a portable format. PDF is the most popular page description language used today. A PDF document consists of pages, made of text, graphics and images, and supporting data. PDF can supports hyperlinks, searching etc. PDF can store bit-mapped images and Adobe provides optical – character-recognition software for the creation of PDF files.

The way to generate a PDF file is to divert a stream of data, then the file can be converted from post script or from another format, stored, transmitted over a network, displayed on a screen and then printed. The PDF thus created are batch optimised so that maximum files are occupied in minimum space and hence searching become quicker.

Optical Character Recognition is the technique of converting scanned images of characters to their equivalent characters. Here a computer program separates out the individual characters and then compares each character to mathematical templates.

4.4 Content Indexing and Metadata

Indexing should be done for easy retrieval of the information. The Dublin core is usually used for the purpose. From 1995, an International Group led by Stuart Weibel of OCLC has been working to device a set of simple metadata elements that can be applied to a wide variety of digital Library materials. The set of elements developed by the group is known as the Dublin core.

The fifteen elements constitute the metadata set of Dublin Core are title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage and rights. There are two options in Dublin core, (i) minimalist option to meet the original criterion of being usable by people and (ii) structured option which is more complex requiring special guidelines. Storing the metadata and data together is convenient for long term achievement since computer programs have access to the data and to the metadata at the same time. For an HTML page, the attaching of the metadata is done in the page by using special HTML tag<meta>, which comes from an HTML description of the Dublin Core element set. In file types other than HTML, Resource Description Framework (RDF) are developed by World Wide Web consortium.

4.5 Information Retrieval Procedure

Information can be retrieved from the file by asking a query, a search term. The query may be a single search term, a string of terms, a phrase in natural language etc. Full text-search – facility is also there. A Boolean query consisting of two or more search terms related by AND, OR, NOT can also be used.

5. Problems Facing Digitization

Even though libraries and librarians all over the world are marching towards digitization, there exist some constraints in the process and their maintenance. The problems facing digitization are

5.1 Longevity of Storage media

Many of the storage media praised by people all over the world may become less useful only long after they become unreadable. Thus documents digitized and stored in such media become useless and their maintenance will be more difficult than print media. The digital archival media today used are magnetic tapes, CD-ROM discs and DVDs. From the scene magnetic tapes disappeared because of their short life due to demagnetization, material decay and oxidation.

During 1980's CD-ROMS emerged into the field and boasted of a longer life span of 30-100 years. Now a days most of the CD's go to the way of 51/4 diskettes. DVD having several standards pushed CD's behind the screen. The changes and improvements of storage medium put serious questions about the future of digitized materials and their alteration.

5.2. Technology obsolescence

The technology behind digitization is undergoing drastic changes continuously. The computer hardware, software, storage media etc are undergoing great revolution. The digitized materials become unreadable if the background devices become obsolete as time passes by which ultimately results in the loss of data. Like print media, digital media is also affected by light, heat, moisture, insects, acid content and air pollution. Digital storage media are always under the threat of above factors. While selecting the storage medium, technological obsolescence should be taken into consideration.

5.3 Migration

The periodic change of digital systems from one configuration to another to overcome the problems caused by technological obsolescence is termed as migration. Migration to a new storage system is more expensive and this will ultimately result in the loss of data.

5.4 Selection of Documents

In an age of information explosion and information pollution, librarians are in a dilemma about 'what type of records are to be digitized' and 'what type of records not to be digitized'. The documents in high demand today may become obsolete even tomorrow because of the vast developments in the subject and printing and publishing industry. A digitized document deselected from the collection is lost for ever. To overcome the problem, librarians should seek the advice of subject experts in each field and users of the library about the importance of each and every record and from this list selection of records for digitization can be done.

5.5 Copyrights

The issues regarding copyright raise serious matters before librarians in digitization. Research scholars usually include graphs, data from books and journals without prior permission of the author. In a digital library users are always demanding back issues of journals and rare historical archives for which the library has no copyright. This may lead to serious dissatisfaction about digitization among users. As a final solution to this matter, librarians must be given permission to digitize copyright works in connection with digitization.

6. Use of Digitized Materials

The typographic standards ie titles, headings, subheadings, typefaces, paragraphing and folios, followed by printed sources are lacking in digital information. The users give utmost importance to the above factors. The reading of a book from online is time consuming, laborious task, causes several problems to the health of the user, requires more money for downloading and printing it. Eventhough writers and reading community are charging the society that 'reading is dying', in reality this can only be taken as an allegation in the digital environment. Actually, serious readers always prefer printed version and digital information only as a supplement and not as a substitute. The concept of 'digital divide' is of special mention at this juncture.

7. Conclusion

The digitization of collection of a library opens its doors to the world so that local collections get a wider exposition. In the field of Science and Technology, there is emergence of interdisciplinary and multidisciplinary subjects and research reports. Articles are being published in science journals in a huge amount than in the parts. The escalating price of the journals are not affordable to each library. The emergence of E-journals and digitization of journals abstracts and indices reduce the burden of their procurement and save storage of space. Although there are drastic changes in digital technology, finance, staff training, manpower, infrastructure etc are serious problems to be tackled before libraries attempt for digitization. In a country like India having great history in traditional medicine, ancient art, culture, architecture, etc, the information that our great ancestors gave us through inscriptions, archives, and through rare books is to be digitized for our future generation.

8. References

1. Communications of the ACM, 38 (H), 1995, P. 59 – 60.
2. Lynch, Clifford and Gracia – Molina, Hector. (Eds.) Inter-operability scaling and the digital libraries research agenda, 12 August 1995. A report on the May 18-19, 1995 IITA Digital Libraries Workshop. (http://ww-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html).
3. http://sunsite.berkeley.edu/ARL/definition.html (accessed on 8.4.2003).

About Author

Ms. B Mini Devi is currently working as Technical Assistant in the University Library, University of Kerala. She holds MSc (Botany) MLIS and has over 10 years professional experience. She has also qualified the UGC NET for JRF & Lectureship and is doing research in the field of Informetrics.