National Conference on Information Management in e-Libraries, 26 -27, February 2002 IIT, Kharagpur. Ed by Parthan s. Allied Publishers Limited, New Delhi, 275-285.

Nitya Archives: Innovative blending of techniques for Selective Access to Information from Digitally Organized Text (SAIDOT)

K Ravindran Asari, K H Hussain, R Raman Nair

Abstract

The invention of digital technology for information storage and retrieval is a milestone in the process of making formal communication flexible and transparent. It is only by harnessing digital technology the wealth of information, which is growing at an exponential rate, can be effectively managed. By developing capacious secondary storage devices like Compact Discs and hard disks, it is made possible to have digital libraries and digital publications. The present trend is to convert printed records into digital records – a situation comparable to the Middle Ages when manuscripts were converted to printed documents with the invention of printing. Nitya is a full text retrieval utility developed by CIRD (Centre for Informatics Research and Development) in Kerala for the creation and maintenance of digital libraries. The utility can be used for bringing out digital publications too. Nitya is created by interfacing UNESCO's database management system CDS/ISIS and Adobe Acrobat Reader. Searching of the digital library and opening of the full text is made possible by this front end.

Introduction

Ever since man started recorded communication, he has been perfecting or inventing techniques to make the communication process more flexible and transparent. The milestones in this direction are the invention of alphabets, paper, printing press and digital recording technique. For more than five

centuries, the print medium has been almost entirely dealing with information generation, storage and dissemination. The print medium will be with us for the foreseeable future. But as a medium to retrieve required information from a large store and to transmit information regardless of space and time, the print medium has limitations. The major drawbacks of the print media are the following:

- the print media occupies more space;
- retrieval of specific information from a large store of information is very difficult and time consuming;
- related information, associated graphics and sound cannot be integrated at one point as the print media present information in a linear format;
- the print media cannot provide interactivity;
- the quality of printed or written documents deteriorates when preserved for longer period;
- in the case of printed documents, transfer from place to place is only by physical means.

The emergence of microform publications could lessen the demand for storage space to some extend. A number of publications were converted into microform to save storage space. Such publications include, back volumes of journals, library catalogues, bibliographies, reports, newspapers and old publications. Although it reduced storage space, other problems like selective dissemination, transfer of information regardless of space and time, integration of related information including graphics and sound and interactivity persisted.

The Emergence of Digital Technology

These problems could be solved only when it was discovered in the 1930s that the ‘on’ and ‘off’ states of an electric circuit could be interpreted as ‘1’ and ‘0’ and they could be used as the alphabet of a flexible communication media. Information recorded using this binary language is more appropriately known as digital information. Although, the binary system or digital

communication could be invented in the 1930s, its large-scale application had to wait until the 1960s when computers acquired enough memory to store large quantities of information.

The progress achieved in the digital technology and secondary storage media like high capacity hard disks and compact discs facilitated compact and durable storage of information of all kinds – text, graphics and sound. The hypertext and hypermedia features of the digital media can integrate related text and associated graphics and sound, which can be used in an interactive manner. In a networked environment the digitized information can be transferred regardless of space and time. Also, the needed information can be retrieved from any part of the world. More than anything else, the digitized information is fresh even after a longer period. Thus digitization has brought in flexibility and transparency in information storage and retrieval. Any information stored using the binary language permit fast retrieval of relevant information, integration of various components of communication like text, numbers, graphics, sound etc and interactivity.

The subsets of digital libraries include computerized catalogue and OPAC; Full Text Digital Archiving and Retrieval Systems; Electronic and Digital Publishing in the Net as well as on Compact Discs; and Virtual Libraries. But more precisely digital library is an electronic substitute of the conventional library. Let us look at some of the definitions.

Definition

According to Gail McMillan, Director, Digital Library and Archives University Libraries, Virginia Polytechnic Institute and State University, digital library is actually a library and not a mere mechanical store of digitized information. It needs to be evolved beyond mere storage and access to digitized information. A library is a fusion of resources in a variety of forms, including services and people supporting the entire life cycle of information beginning with creation to dissemination and use. A digital library works best when it is an integral part of a library that provides its users with access to information that has been evaluated, organized, and preserved in the most useful

formats. Digital libraries and traditional libraries share common goals and should interact as if they share a common soul.

Nevertheless, many of the existing definitions consider that digital libraries are just the digital information available through the Internet and mere stores of digitized information. For example, the International Cooperation on Digital Libraries" defines it as "a collection of digital objects along with methods of access and retrieval, and for selection, organization, and maintenance of the collection."

Actually a digital library should be a series of activities that brings together collections, services, and people in support of the full life cycle, from creation, dissemination, use, and preservation of data, information, and knowledge. The challenges and opportunities that motivate an advanced digital research initiative should be committed to such a broad view of the digital library environment. A digital library should be a seamless extension of the library that provides scholars with access to information in any format that has been evaluated, organized, and preserved. Access to this evolving collection of digital information should be provided through personalized systems as well as through the services of information professionals. The digital library adds value and saves time while extending the hours of access. It reduces the need for proximity to information resources, but still emphasizes the quality of those resources. It is a library that can be individually customized and, ultimately, will be easy to use.

When Johannes Gutenberg of Germany invented the art of printing during the 1440s, the trend was to convert every manuscript into printed form. Microfilming technique helped to convert at least archival materials like old books, library catalogues, journals and newspapers into microform. What is prevalent at present is conversion of printed documents into digital form. Now we have books, journals, newspapers, reference books and even libraries in the digital format.

International and National Efforts Towards Digital Libraries

Thousands of examples can be cited for digital libraries. Only three examples of international magnitude are cited here:

1. UNESCO memory of the World Programme

(http://www.unesco.org/webworld/mdm/index) This programme seeks to safeguard documentary heritage of the world. Naturally, it reflects the diversity of languages, peoples and cultures. 'It is the mirror of the **world** and its **memory'**

2. American Memory

(http://www.loc.gov/) American history in words, sounds and pictures and extensive online exhibitions of historic photographs and documents.

3. The NSF/DARPA/NASA Digital Libraries Initiative (DLI)

The National Science Foundation, Department of Defense Advanced Projects Agency and the National Aeronautics and Space Administration funded six research projects in 1994 aiming at developing new techniques for creating digital libraries

In India, several institutions and national level organizations have taken steps to develop digital libraries. Projects aimed at digitizing rare and out of print publications and manuscripts have been taken up in institutions like Indira Gandhi National Museum of Art, The National Library, Calcutta and Tamil University. The INFLIBNET has also initiated programmes to digitize rare collections of several institutions and older libraries. However, all projects aimed at creating digital libraries concentrate only on special collections. A number of utilities have been developed in India also to facilitate creation of digital libraries. Digital Publishing Solutions, Pune is an example of commercial agencies offering creation of digital libraries.

Depending upon commercial agencies always for the creation of digital libraries is neither practicable nor cost effective. Therefore, it is necessary to

make professional approaches towards the development of appropriate utilities and to train a new generation of library professionals to create and maintain digital libraries. Keeping these professional goals in mind Center for Informatics Research and Development (CIRD) has developed NITYA to achieve Selective Access to Information from Digitally Organized Text (SAIDOT)

NITYA

Nitya combines high-level text compaction technique and highly sophisticated free text search and retrieval procedure. The most outstanding aspect of Nitya is that any piece of information can be searched out from a huge store of information within seconds. Searches can be by title, by author, by date of publication, by a keyword or combination of keywords using the Boolean logic. It even permits proximity searches. Although, several systems are in vogue for the digital storage of information, they lack efficient search mechanism. Devoid of a search mechanism, digitally stored information in hard disks and compact discs remain as dump places of information, just the same way as documents converted into microform. The full potential of digital technology can be exploited only if selective access is made possible. Then only it can perform the traditional functions of a library. Nitya combines the capabilities of two best-known programs to achieve efficient information storage and selective retrieval. These programs are Adobe Acrobat and UNESCO's CDS/ISIS. The later stands for Computerized Documentation Service/Integrated Set for Information Service. CDS/ISIS is a high profile database management system, which is ideal for bibliographic database management. Under the auspices of UNESCO and several national governments, thousands of information specialists all over the world are trained on the use of CDS/ISIS. In India, the National Information System for Science and Technology (NISSAT) provides training through various institutions on the use of CDS/ISIS, who also distribute the package. Initially CDS/ISIS was available only in the DOS version. Now a Windows version is also available and it can be web enabled.

Adobe Acrobat Reader is a text reading utility which provides simple techniques for viewing text and graphics in Portable Document Format (PDF).

Method

Documents are first scanned and converted to Portable Document Format (PDF) using Adobe Acrobat. Images of the pages are converted to text ('Captured') and book marked/annotated. A database is created for the documents selected for the digital library and indexed using CDS/ISIS. The efficiency of the search depends on Indexing Techniques applied to build up Inverted file (Dictionary of indexing terms). Nitya provides facilities for searching archived documents using Inverted File and the search results are displayed in full bibliographic details. Keywords for query building can be selected from the Inverted file under different categories like author, title, subject, etc. Number of categories depends on the type and characteristics of the archive. (For example, if it is an archive of theses, the researcher, the guide, departments, university, subject, etc. can be categories.) The extensive search mechanism offered by CDS/ISIS permits searching of the archived documents by single keyword or combination of keywords. Proximity and truncated searches can also be accomplished in addition.

With the bibliographic details of the searched documents FULL TEXT button is appeared, pressing upon full text of the retrieved documents are displayed in Acrobat Reader. Acrobat Reader provides many view options to display pages. Viewing with Book Marks enables the user to navigate through the documents. Any part of the text can be zoomed up to 1600 percent. Also, any word can be located from any part of the displayed text with the help of Find option of Adobe Acrobat.

Paradigm used to develop Nitya is adopted from the traditional library information storage and retrieval practices. When a user enters the library, he first goes to the catalogue and searches the catalogue under appropriate access points such as author, title, subject, etc. This enables the user to get most relevant documents he wants. Once he gets the document, with the help of the content page he locates the relevant chapters or pages or

portions of the book. In the same way, when one opens Nitya, a dictionary of keywords (Inverted File) similar to the library catalogue appears. Selected terms in the dictionary are used to form queries. Search yields a hit of relevant documents with full bibliographic details. Full text is opened and the entire text can be navigated through book marks (*Library - Nitya analogy is explained in the chart provided at the end*.)

Scope and Applications

Nitya is so versatile that both reference retrieval system and document retrieval system are combined into one. Older books, archival materials like manuscripts in palm leaves, research reports, conference proceedings, parliament or legislative proceedings, hospital records, theses and dissertations, government orders, journal, newspapers, etc. can be digitized and stored either in hard disks or compact discs. The system can go even to the extend of converting part of a library collection into digital form.

Future Activities

Nitya is conceived as a flexible archival system. As the parameters to archive and retrieve Theses, Conference proceedings, Manuscripts in palm leaves, etc. change, CIRD plans to program different Nityas according to the need and vision of participating librarians and archivists. Like PDF files Nitya can be made to open files in other formats like DOC, RTF, HTML, Multimedia, etc. CIRD plans to extend Nitya application to other Indian languages too. Already techniques have been perfected to search in Malayalam - the regional language of Kerala. It may not be difficult to extend the facility to other Indian languages, if interested group come forward.

Consultancy by CIRD

CIRD identifies five types of institutions in need of digital library creation:

i. Institutions, which do not want to setup scanning and digitization unit at present.

ii. Institutions, particularly larger ones, which will outsource the work now, but has the potential to develop its own infrastructure and capability.
iii. Institutions having all the necessary equipment, software, etc., but do not have the digital library creation know-how.
iv. Institutions already started creating digital library using other packages, but decided to convert to NITYA after seeing its capability.
v. Agents who want to publish books like encyclopedia, directory, etc. electronically using Nitya S/W.

Conclusion

The Nitya package developed to achieve SAIDOT forms one of the efficient, simple and cost effective method for the creation and maintenance of digital libraries. Consultancy and training by CIRD aims transfer of technology to professional librarians and archivists. NITYA is an extension of the traditional library service in the information age, highlighting the importance of wisdom of librarianship in organizing knowledge.

References

1. Witten, Ian H. Visions of the digital library. International Conference on Asian Digital Libraries, 4th, Bangalore, 2001. Proceedings: Digital Libraries, ICDAL 2001, pp 3-15
2. Rowley, Jennifer. The Electronic Library. 4th ed. London, Library Association, 1998, 390p
3. Raitt, David. Ed. Libraries for the New Millennium: Implications for managers, London, Library Association, 1997, 288p.
4. Griffin, Stephen M. Digital Libraries and the NSF/DARPA/NASA Digital Libraries Initiative. In: Raitt, David. Ed. Libraries for the New Millennium. London, Library Association, pp 115-147.
5. http://www.unesco.org/webworld/mdm/index
6. http://www.loc.gov/