What's New in Computers?

Digital Libraries

T B Rajashekar

T B Rajashekar is with the National Centre for Science Information (NCSI), Indian Institute of Science, Bangalore. His interests include text retrieval, networked information systems and services, digital collection development and management and electronic publishing.

Digital libraries aim to provide access to information 'on demand' regardless of the location of the computer where it is stored. In this article we examine key features of digital libraries.

Introduction

The term 'Library' usually invokes a storehouse of information in the form of print publications like books, journals, reports, etc. and newer media such as films, filmstrips, video and audio cassettes. Most of us view the library as a place where such information 'containers' are acquired, organised, shelved and retrieved. More recently, many libraries (and information centres) have begun to use online database search systems like DIALOG and connections to the Online Public Access Catalogues (OPACs) of other libraries, over telecommunication networks. Many libraries have also taken advantage of CD-ROM technology to provide users network access to large information bases. With such remote information access, the walls of the library began to be less solid.

The way we view information sources is changing due to the emergence of global computer networks. Personal computers are now usually part of a network of computers and often part of a network of networks, spanning the globe. Until recently, computers and computer networks were accessible easily only to a privileged few - computer scientists, researchers and hobbyists. For most of the others they were out of reach, both economically and practically. Now, computers are less expensive and inter connectivity is much more widespread. People are becoming more computer literate, and systems are becoming simpler and

more accessible. This is happening in India also and is changing the way we communicate and think about information and libraries.

Digital libraries may be defined as electronic information collections containing large and diverse repositories of digital objects, which can be accessed by a large number of geographically distributed users.

The Internet and the World Wide Web (WWW) technologies are providing the technological environment and intellectual impetus for the development of 'digital libraries' libraries without walls, with containerless data and ideas. The Internet has enabled global connectivity of computers and the development of various tools and techniques for networked information provision and access. Starting with basic tools like e-mail (messaging), ftp (file transfer) and telnet (remote login), the Internet has progressed to provide user friendly tools like Gopher, WAIS and the WWW for information publishing and access. The WWW, which is integrating all other access tools, also provides a very convenient means for publishing and accessing multi media, hypertext linked documents stored in computers spread across the world, using text coding standards such as HTML and SGML (see *Box 1*). These tools have been used for providing access to a variety of information sources and services, including electronic journals and news letters, tables of contents, pre-prints, technical reports, software and data archives, library catalogues, discussion forums, reference sources, courseware, directories, etc. Given the situation where most of the researchers have desk top computers that are connected to the Internet (via institutional LANs), they can very quickly start accessing and publishing information on the Internet.

What are Digital Libraries?

Digital libraries is an evolving area of research, development and application and multiple definitions have been offered by workers in this area. Based on common aspects among these definitions, digital libraries may be defined as electronic information collections containing large and diverse repositories of digital objects, which can be accessed by a large number of geographically distributed users. Such repositories would exist

Box 1. HTML and SGML

The lingua franca of the World Wide Web (WWW) is HTML - Hypertext Markup Language. It is the standard language the Web uses for creating and recognising hypermedia documents. These documents are stored on Web servers. A Web client program (called a Web browser) sends request for documents to a Web server. A Web server program, upon receipt of a request, sends the document requested (or an error message) back to the client. Web documents are typically written in HTML and named with the suffix '.html'. HTML documents are nothing more than standard ASCII text files with formatting codes that contain information about layout (text styles, document titles, paragraphs, lists) and hyperlinks.

The following is an example of HTML marked up text :

```
<H1>An introduction to Digital Libraries</H1>
<H2>Topics covered :</H2>
<OL>
<LI>Introduction</LI>
<LI>What are digital libraries?</LI>
<LI>Digital library projects</LI>
</OL>
```

The following is a rendering of this HTML document by a Web browser :

An introduction to Digital Libraries

Topics covered :
1. Introduction
2. What are digital libraries?
3. Digital library projects

In addition to marking up of documents for formatting purposes, HTML supports a very powerful feature - embedding hypermedia links using Uniform Resource Locators (URLs). These URLs may point to Web documents and multi-media objects like audio, image and video, stored on the local machine or a Web server somewhere on the Internet. They may also link to network services like telnet (for remote login) , e-mail and ftp (file transfer).

For example, to make the sentence "Click here to go to IISc Web site", an actual link to IISc Web site whose URL is http://www.iisc.ernet.in, one would add the following tags to the text:

```
<A HREF="http://www.iisc.ernet.in>Click here to go to IISc Web site</A>
```

Box 1 Continued...

Any text editing program (e.g. vi, Notepad) can be used to create an HTML document. Many special purpose HTML authoring programs are also available (e.g. Hot Metal, HTML Assistant) which simplify the process of marking up. Free conversion software is also available for translating from many other formats into HTML (e.g. text-to-html, latex-to-html).

HTML is in fact an implementation of an ISO standard - the Standard Generalised Markup language (SGML) (ISO-8879, 1986). It is an international standard for the definition of device-independent, system-independent methods of representing texts in electronic form, using descriptive markups. More exactly, SGML is a metalanguage, that is, a means of describing a markup language. HTML is a specific implementation of SGML, used for marking up Web documents. SGML is increasingly being used for implementing markup languages, through what is called the Document Type Definition (DTD), for a wide variety of document types like manuals, books, journals, etc.

in locations physically near or remote from the users. Digital objects include text, images, maps, sounds, videos, catalogues and indices, and scientific, business, and government data sets as well as hypertextual multimedia compositions of such elements.

Operationally therefore, digital libraries will be network based distributed systems with individual servers responsible for maintaining local collections of digital documents ranging from sets of electronic texts to video-on-demand services. From the user's point of view, however, there should appear to be a single digital library system, integrating personal information, work group and organisational information collections and public digital libraries. Any given digital library system should therefore provide a coherent, consistent view of as many of these repositories as possible and allow users to seamlessly connect and interact with information with no regard to geographic location or time.

There is also general acceptance of the fact that digital libraries would need to span both print and digital materials. For the foreseeable future print on paper publications are expected to be

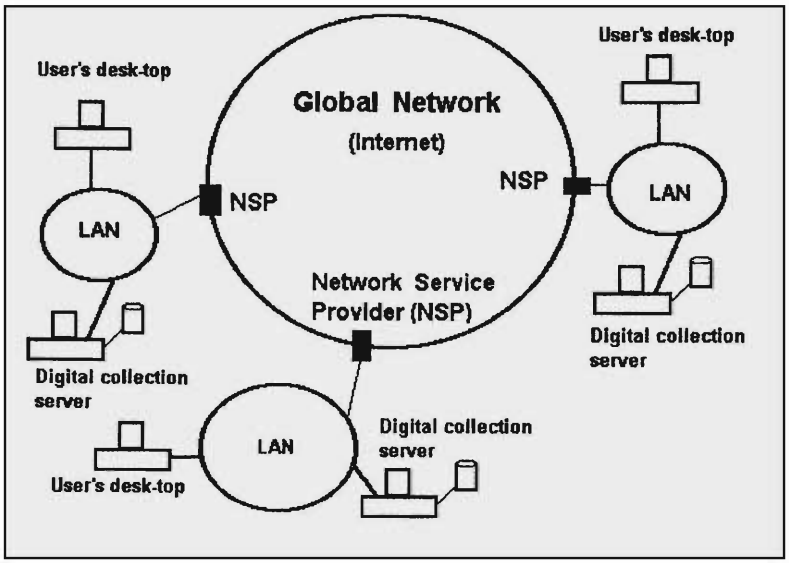

Figure 1 The global digital library environment.

around and thus digital libraries are expected to provide integrated, coherent access to both types of materials.

Key components of digital libraries, are therefore :

- Geographically distributed digital information collections
- Geographically distributed users
- Information represented by a variety of digital objects
- Large and diverse collections
- 'Seamless' access

The schematic of the emerging global digital library environment is shown in *Figure 1*. It shows user desk-top computers (e.g. PCs) connected to the global network via institutional local area networks and network service providers (e.g. Internet service providers). A variety of digital information collections will be mounted on powerful workstations in stand alone or networked configurations. Using powerful, but very user friendly browser programs (similar to the present day WWW browsers) on their desk-tops, users will be able to seamlessly connect to and extract information from geographically distributed digital libraries.

This process will be facilitated by intelligent information finding agents that will use variety of tools like meta databases and search engines to extract and deliver relevant information to the user's desk-top.

Using powerful, but very user friendly browser programs on their desk-tops, users will be able to seamlessly connect to and extract information from geographically distributed digital libraries.

Digital, Electronic and Virtual Libraries

A source of confusion in this area has been the use of terminologies like 'virtual', 'digital' and 'electronic' libraries. One person's digital library is often another's 'virtual library'. Some useful distinctions have recently been made :

Electronic library : A library that provides collections and/or services in electronic form, for example, optical video disk (which is not digital), CD-ROM, online, etc.

Digital library : A library that provides collections and/or services in digital form.

Virtual library : A library that does not physically exist, most often used to denote a library with distributed collections or services that appear and act as one. Typical example is a Web site with pointers and links to other sites.

From the above it may be seen that an electronic library is more inclusive than a digital library. However, digital library has come to be the preferred term, perhaps keeping in line with terms like digital audio and digital video. The current usage of the term 'digital library' appears to encompass both electronic and virtual libraries.

NCSTRL – An Example Digital Library

Let us take a look at one of the operational digital libraries. This is the Networked Computer Science Technical Report Library (NCSTRL). NCSTRL provides unified access to catalogue records and complete documents of computer science technical

NCSTRL provides unified access to catalogue records and complete documents of computer science technical reports stored in distributed servers around the world, through the WWW. NCSTRL can be accessed free of cost on the Internet using any WWW browser.

reports stored in distributed servers around the world, through the WWW. It is a direct descendent of two earlier systems, WATERS (Wide Area Technical Report Service), sponsored by the National Science Foundation (NSF) and Dienst CSTR (Computer Science Technical Reports) project sponsored by the Department of Defenses' ARPA. WATERS and Dienst have been operational since the middle of 1993. NCSTRL, put into operation in November, 1995, combines the best of both the projects. Participation in NCSTRL is open to all academic departments awarding Ph.Ds in computer science/ engineering and to research facilities of industry and government. Currently over 200 departments around the world are participating. NCSTRL can be accessed free of cost on the Internet using any WWW browser (http://www.ncstrl.org/).

NCSTRL is aimed to provide financial and scholarly advantages to all its users and participants. Researchers can gain easy and quick access to a large body of scholarly material. Authors gain a wider audience, especially those from less well-known institutions, as everyone has an equal chance of having their reports noticed. Departments get a clean, effective management system for their technical reports and eliminate much of their current copying and mailing charges. Several computer science departments are estimated to have already saved thousands of dollars.

NCSTRL provides a very simple search interface and allows search by author, title or abstracts. A retrieved document may be viewed as an HTML document or in PostScript. The technology underlying NCSTRL is a network of inter-operating digital library servers. These servers provide three services : *repository* services that store and provide access to documents, *indexing* services that allow searches over bibliographic records, and *user interface* services that provide the human front-end for the other services. These services inter-operate using an open protocol, enabling development of new kinds of services (such as contents alerting services).

At individual departments, NCSTRL can be installed at two levels - *Lite* and *Standard*. Lite version is intended for sites with few resources needing lower starting investment, while Standard version will offer greater functionality. Participating sites can choose to install either of them. Lite sites only need to put their reports online (via FTP or HTTP) and edit a catalogue file with the *techrep* tool (supplied with the software). The user interface and the index service for a Lite site executes on the central NCSTRL server, which acts as an NCSTRL gateway for all the Lite sites. NCSTRL Standard software will enable full local indexing and access, including thumbnail browsing and online reading.

NCSTRL can be installed at two levels - *Lite* and *Standard*. Lite version is intended for sites with few resources needing lower starting investment, while Standard version will offer greater functionality.

No matter which version is installed, the complete technical report collection will be available to all parties in a manner totally transparent to the users of NCSTRL.

Digital Library Projects

Since the past couple of years the idea of 'digital library' has moved to the forefront of discussion and research. Several digital library projects are currently underway in U.S.A., Europe, Australia, New Zealand and Singapore (see *Box* 2 for details of a few projects).

Advantages of Digital Libraries

Some of the key advantages digital libraries provide are

- **Ability to search**

The ability to search provides an enormous advantage to electronic materials when an ASCII version is available. Online searching has for some years been replacing printed abstract journals. Since most current material is now produced via computers, it can generally be provided in ASCII form and be searched. For those documents which are searched rather than

Box 2. Digital Library Projects

Impetus for digital libraries development has come from two major initiatives taken in the USA. First was the joint initiative of the National Science Foundation (NSF), Department of Defence Advanced Research Projects Agency (ARPA), and the National Aeronautics and Space Administration (NASA), in 1994, to fund several digital library development projects for a period of four years, mostly among academic institutions. The second was the signing of the National Digital Library Federation Agreement in May, 1995, led by the Library of Congress and 14 other research libraries. The purpose was to "bring together - from across the nation and beyond - digitized materials that will be made available to students, scholars and citizens everywhere". The focus of the NSF/ARPA/NASA initiative was to dramatically advance the means to collect, store, and organise information in digital forms, and make it available for searching, retrieval, and processing via communication networks. Some of the projects funded under this initiative include the following :

- Informedia digital video library (Carnegie Mellon Univ.) : Goal is to develop technologies for full-content search and retrieval from digital video libraries.

- Illinois DL project (Univ. of Illinois, Urbana Champaign) : Goal is to provide comprehensive search and display of complete contents of articles from science and engineering journals.

- Alexandria DL project (Univ. of California, Santa Barbara) : Goal is to provide access to variety of spatial information (digitised maps, images, air photos, and other graphical information relating to the counties of Santa Barbara, Ventura and Los Angeles).

- University of Michigan DL project (UMDL) : This has several projects including the Journal Storage Project which will digitise and provide access to all issues from the beginning through 1990 of 10 economics journals.

- University of Berkeley DL project : Goal is to develop technologies for intelligent access to multi terra byte databases (multi-media documents including photographs, satellite images, videos, and full text documents).

- Stanford DL project : Goal is to develop technologies for a single, integrated and 'universal' library, composed from a large number of heterogeneous repositories.

Since then several similar projects have been launched in Europe and other countries.

read (e.g. many reference books, compilations, etc.), electronics can be expected to take over shortly. Printed encyclopedias, for example, are giving way to CD-ROMs which are small, cheaper, and far more effective.

- ***Ubiquity***

Another key advantage is ubiquity - a single electronic copy can be accessed from a great many locations, by many simultaneous users. Copies can be delivered with electronic speed, and it would be possible to reformat the material as per the reader preference (e.g. character size). Since readers get a screen display of the object, rather than a physical object, loss rates by theft are eliminated.

- ***Support wider range of material***

Digital storage also permits libraries to expand the range of material they can provide to their users. Since audio cassette tapes and records cannot stand a large number of playings without deterioration, their digital representation (digital audio) can produce a format that is much safer and of better quality. Digital material can also permit access to video tapes and new kinds of multi-media materials that are created only on computers and have no equivalent in any traditional format.

- ***Preservation***

Another major advantage is preservation. Digital information can be copied without error. As a result, preservation in a digital world does not depend on having a permanent object and keeping it under guard, but on the ability to make multiple copies, assuming that at least one will survive.

- ***Access current information***

For researchers, digital libraries provide access to up to date current literature and thereby help them to be aware of current trends.

It is expected that several digital library projects currently underway will result in technologies that may be used readily to develop new digital libraries.

Technologies for Digital Libraries

Digital libraries are considered by many to be a very challenging research area, as it requires development and integration of several highly sophisticated hardware and software technologies, and pooling together of multi-disciplinary expertise. While many of the technologies required for the development and deployment of digital libraries and their access are available today, their performance needs to be scaled up to handle very large digital collections in networked environment. It is expected that several digital library projects currently underway will result in technologies that may be used readily to develop new digital libraries.

Some of the major areas of focus are :

- Multi-media object storage, retrieval and transmission
- Data Compression
- Digitisation (multimedia data capturing and conversion)
- Hypermedia navigation
- Authoring tools for creating electronic documents
- Multimedia object representation (e.g. HTML, SGML)
- Meta databases
- Display technologies
- User interfaces
- Search, retrieval and routing software

Issues in Digital Libraries

Work in the area of digital libraries has thrown up several issues and challenges. Key issues include the following :

- *Copyright*

It is very easy to copy, replicate, massage and distribute digital information. Enforcing copyright in digital environment is a major issue.

- *Technological obsolescence*

Hardware : The major risk to digital objects is not physical deterioration, but technological obsolescence of the devices to read them. While the lifetime of optical and magneto-optical cartridges is expected to be many decades, those of reading devices is only about one decade. (Solution? Reasonable refreshing schedule and use of widely marketed devices rather than special purpose devices.)

The major risk to digital objects is not physical deterioration, but technological obsolescence of the devices to read them.

Software : A more serious problem is software obsolescence. It has been pointed out that the variety of software formats far exceeds the number of hardware devices manufactured, and that these programs come and go more quickly than the hardware does. (Solution? Libraries should rely on standards like MARC and SGML, which are expected to exist in the foreseeable future.)

- *Cost of regular refreshing*

Digital preservation will be an ongoing operation, requiring considerable recurring expense. Librarians may worry that they are committing their successors to operations for which there will be no funds. Considering the rapid increase in the storage capacities and the decrease in cost per byte, this may not become a problem. If a library can understand how it will fund the first refresh cycle in five to ten years, it can expect that the next refresh cycle will be so cheap as to be insignificant.

- *Image V/S ASCII*

Some materials in digital libraries are available as images, and others as ASCII. In some cases these distinctions are inherent e.g. a photograph collection is always going to be images, and an OPAC is nearly always ASCII. In other cases, e.g. printed books and journals, both formats are possible. Image based products are provided usually by the publishers (e.g. ADONIS and IPO) on CD-ROM, whereas community based products are in ASCII

Suggested Reading

Not surprisingly, most of the publications related to digital libraries are accessible, usually free of cost, over the World Wide Web. We refer to a few key publications here.

- **Fox, Edward. Digital library source book, 1993 (URL:http://fox.cs.vt.edu/DLSB.html).**
- **Digital Libraries'94 : First Annual Conference on the Theory and Practice of Digital Libraries. June 19-21. College Station. Texas,1994.**
- **Digital Libraries'95. : The Second International Conference on the Theory and Practice of Digital Libraries. June 11-13. Austin. Texas,1995.**
- **Association for Computing Machinery (ACM). *Communications of the ACM*. April 1995. Theme issue on digital libraries.**
- **Association for Computing Machinery (ACM). First ACM International Conference on Digital Libraries. March 20-23. Bethesda. MD,1996.**

Continued...

and available online (e.g. the preprints service in mathematics). The widespread view is that online will generally win over CD-ROM and ASCII over image. Libraries, however, have to deal with both worlds.

Images occupy at least 10 times more space than ASCII. With the continuing fall in disc prices, the issue is not of storage cost, but that of network time and display capacity. For example, over a 64 Kbps ISDN line, a text page arrives in less than half a second while an image page takes 4 seconds which is a definite lag. In practice therefore, image systems are limited to local networks (and CD-ROMs), while the ASCII is easily sent over the Web. Another problem is demand placed by images on local display. Most users still use 640x480 screens, not capable of displaying a whole page at once.

Why are images preferred then by some providers? The reason is financial : the cost of keying a book with 300 pages and 500,000 characters is about 10 times the cost of scanning. Other reasons are the simplicity of handling the illustrations or tables on the pages. We can expect that as OCR gets better, the image formats will disappear.

- ***Interworking***

Another key issue is that of interworking of different digital libraries. Given the distribution of library resources around the world, no one expects that there would be a single digital resource. This means that we need methods for finding either individual items or collections in different places, and assembling virtual collections that users can search or browse.

- ***Who will run the digital libraries?***

All sorts of organisations are claiming the right to be the provider of information to the desk-top : online services, libraries, bookstores, publishers, telephone companies, telecommunica-

tion offices, university computing centres, and new startup companies. In the digital, networked environment, it appears that libraries have no unique claim as providers of scholarly information. In practical terms, it is likely that provision of current material will move back to the publishers, who can have sufficient control of who reads what. It is possible that the publishers too will be bypassed, as the authors self-publish on the Net or through some new venture. Witness for instance the variety of courseware available mostly free over the Internet.

- ***Pricing in the digital environment***

Pricing of information in the digital world is going to be very complex. Ownership is expected to give way to licensing, pay per use, etc.

- ***Preservation of electronic information***

Archiving and preservation of electronic information may be one of the most challenging of all tasks we have to solve over the coming two decades. Instead of having shops that sell rare books, we may have shops that sell rare PostScript interpreters. Libraries may have a role to play here considering their traditional responsibility to keep things for a long time and to be able to make them available for use.

Conclusion

Digital libraries are expected to bring about significant improvements over current modes of information publishing and access methods. Educators, researchers and students across the world will be among the first to benefit from digital libraries, particularly those in developing countries.

- IEEE. *Computer* magazine, May 1996. Theme issue on the US Digital Library Initiative.
- Berkeley Digital Library Sunsite. Information related to digital library projects and resources. (URL:htt://sunsite.berkeley.edu).
- *D-Lib* magazine. A monthly electronic journal devoted to reporting developments in digital library research. Maintained by the Coalition for National Research Initiative (CNRI), USA. (URL:http://www.dlib.org).

Address for Correspondence
T B Rajashekar
National Centre for Science Information,
Indian Institute of Science,
Bangalore 560 012, India
email: raja@ncsi.iisc.ernet.in