

CHAPTER 2

GLOBAL DIGITAL LIBRARIES: BUILDING THE INFRASTRUCTURE

Raj Reddy

INTRODUCTION

Building a digital library is a daunting challenge, involving a wide variety of technological, social and policy issues. One of the goals of this WTEC study was to compare the relative progress being made on these problems in the United States and Japan. To set the stage for the comparison and the chapters that follow, this chapter will put forward a futuristic vision of the digital library (and shared digital libraries) and lay out the obstacles to its realization. This will set the stage to compare the United States and Japan based on the panel's firsthand observations.

LIBRARY OF THE FUTURE

The library of the future will be digital and have the following features:

- contain all recorded knowledge online (billions of items)
- distributed, maintained globally
- accessible by:
 - any person
 - in any language
 - any time
 - anywhere on earth
 - via the Internet
- act as the information resource for the 21st Century

This notion of a digital library is radical; most current digital library efforts are aimed at building individual libraries to replace existing physical libraries that are burdened by financial and space constraints. However, the avowed digital library policy of Japan is that “each home will be able to access electronic libraries and electronic museums around the world via networks, allowing users to readily search and obtain worldwide information on books and art based on their own particular interests” (MITI 1994).

DIGITAL VS. TRADITIONAL LIBRARIES

The shift from traditional libraries to the digital is not merely a technological evolution, but requires a change in the paradigm by which people access and interact with information.

A traditional library is characterized by the following:

- emphasis on storage and preservation of physical items, particularly books and periodicals
- cataloging at a high level rather than one of detail, e.g., author and subject indexes as opposed to full text
- browsing based on physical proximity of related materials, e.g., books on sociology are near one another on the shelves
- passivity; information is physically assembled in one place; users must travel to the library to learn what is there and make use of it

By contrast, a digital library differs from the above in the following ways:

- emphasis on access to digitized materials wherever they may be located, with digitization eliminating the need to own or store a physical item
- cataloging down to individual words or glyphs
- browsing based on hyperlinks, keyword, or any defined measure of relatedness; materials on the same subject do not need to be near one another in any physical sense
- broadcast technology; users need not visit a digital library except electronically; for them the library exists at any place they can access it, e.g., home, school, office, or in a car

Everything Can Be Stored

The total number of different books produced since printing began does not exceed one billion. (The number of books now published annually is less than one million.) If an average book occupies 500 pages at 2,000 characters per page, then even without compression it can be stored comfortably in one megabyte. Therefore, one billion megabytes are sufficient to store all books. This is 10^{15} bytes, or one petabyte. At commercial prices of \$20 per gigabyte, this amount of disk storage capacity could be purchased for \$20 million. So it is certainly feasible to consider storing all books digitally.

Very Large Databases

A database of a billion objects, each of which occupies one megabyte, is large but not inconceivable. Once one is comfortable with sizes of this kind, it is feasible to imagine a thousand such databases, or to envision them all as portions of the same global collection. This amount of storage is sufficient to house not only all books, but all of the following:

- photographs
- legislative material, court decisions
- museum objects
- recorded music
- theatrical performances, including opera and ballet
- speeches
- movies and videotape

Distributed Holdings

When information is digitized and accessible over a network, it makes little sense to speak of its “location,” although it is technically resident on at least one storage device somewhere, and that device is connected to at least one computer. If the information is available at multiple mirror sites, it is even less meaningful to speak of it being in a “place.” While traditional libraries measure their size by number of books, periodicals and other items held, the relevant statistic for a digital library is the size of the corpus its users may access. This means that digital libraries will want to expand their “holdings” by sharing digital links with other libraries. Unfortunately, there seems to be very little sharing of this sort taking place at present.

How can we understand the unwillingness of libraries to share content? The question goes back to the old measure of the size of a traditional library—the number of books it holds. When a library expends funds to assemble digitized works, it loses a portion of its prestige (or thinks it does) by allowing other libraries to copy or access its data. Ultimately, however, *all* material should be accessible from *every* library.

Gore's Digital Earth

In 1998, Vice President Gore stated that “A new wave of technological innovation is allowing us to capture, store, process and display an unprecedented amount of information about our planet and a wide variety of environmental and cultural phenomena. ... I believe we need a ‘Digital Earth.’ A multi-resolution, three-dimensional representation of the planet, into which we can embed vast quantities of geo-referenced data” (Gore 1998). He then called on scientists to create a digital map of the earth at a resolution of one meter. Such a project will require technical innovation beyond that required even for a digital library containing every book ever written. The area of the earth in square meters is about 5×10^{14} . Storing two megabytes of data per square meter (which would include terrain data, imaging, environmental and other pertinent information) will require 10^{18} bytes, an amount roughly equal to the amount of digital storage currently present on earth.

DIGITIZATION

Ultimately, everything that people are interested in accessing will have to be digitized. The reason is that digital searching will become so easy, inexpensive, fast and ubiquitous that users will not tolerate, or will not access, traditional materials. Capture requires a concerted, shared, worldwide effort. The cost of digitizing is not trivial, so it makes little sense for any work to be digitized more than once. Yet without any registry of digitized works, many books are digitized multiple times, while others are ignored.

Converting text, images and objects to digital form requires much more than digital photography or even high-resolution scanning and requires, instead, the following:

- initial input, either scanning or keyboarding
- conversion to one of a set of standard formats
- optical character recognition (OCR) to capture text characters for searching
- OCR correction (since OCR is inherently error prone)
- creation and input of metadata and cataloging information
- special techniques for non-textual materials, such as music, images, videotape, etc.

Policy issues, discussed below, will determine the resources made available for digitization and how they will be allocated.

PARADIGM SHIFT

If digital libraries are to become truly useful, they must assist users in making the transition from paper books to digital hypermedia. Many people report that they derive a high degree of comfort from books for the following reasons:

- Portability. Books can readily be carried, are compact, light in weight and comfortable to read. Anything you can't read in bed will never displace a book.
- Reliability. Reading books would still be possible even if every computer on earth were down.
- Familiarity with the medium. The pages of a book are easy to turn, the book can be opened to any page, and the linear hierarchical organization of the material is easy to grasp.
- Low cost.
- Ability to annotate. Comments and corrections can be written in a book; passages can be marked for emphasis or studying, and a book can be resold to recover costs.

DIGITAL LIBRARIANS

A move to electronic libraries will alter the fundamental role of librarians. Far less attention will be paid to acquisitions, cataloging and circulation, and much more to systems, online assistance, navigation assistance and conversion issues. Unless graduate programs for digital librarians are established to create trained personnel, the real risk exists that users, particularly young adults who grew up in a digital generation, will outstrip the ability of librarians to assist them.

THE ELECTRONIC BOOK

It will be impossible to displace paper books if reading digital ones is inconvenient. In fact it is likely that an electronic book will have to offer significantly more capability than a paper one to gain wide acceptance. Toward this end, several manufacturers are developing E-books which attempt to mimic the essential features of traditional books while providing huge advantages, such as the ability to store 1,000 electronic books in the space of one paper book and keep the user's place in each.

The worldwide ecological burden imposed by paper books is immense. Tremendous energy and human resources are required to grow and harvest trees, convert them to paper, print and bind books, transport them to warehouses, maintain bookstores and ship the books from the warehouses to the bookstores. The processes of making paper and ink release noxious chemicals, to say nothing of the fuel required to transport boxes of books all over the world. All of this is done merely to disseminate information that could be disseminated electronically with far greater speed and efficiency.

The chief technical barrier to E-books at present is that they do not capture the familiar feel of paper volumes. When this problem is solved, we will then have to face the issue of how content is downloaded to them and paid for.

INTEROPERABILITY

Given that numerous libraries around the world will be developing digital collections, how will it be possible for a user of library A to access and view material housed in library B? The existence of numerous digital libraries will make it essential to share digitized items and ensure that cataloging, searching and retrieval tools at each one can be used readily with materials from others.

Formats and Standards

While standards may have the effect of inhibiting innovation, they are essential to interoperability. Agreement must be achieved on such fundamental issues as how text is to be stored. Is it straight ASCII, Microsoft Word, HTML, SGML, XML or something else? What kind of compression will be used? If text is compressed, how will searching be done? How are images, music and videotape to be represented? If agreement is not reached, at least the number of different ways in which works are digitized should be reduced to a number small enough to allow each library to support them.

Digital libraries must also have a second set of intake standards, going not to technology but to quality and reliability, which are discussed later on. Archivists question the permanence of digital materials since they note that electronic documents can be modified readily and the media on which they reside become obsolete at least once each decade. The question then is how an ever-expanding corpus of information will be converted to new media and formats as these evolve.

Metadata

This term is often used to mean information *about* an item, rather than the information in the item itself. Examples include the author, title, date of acquisition, price paid, donor, etc. It is particularly critical to capture metadata that is not present in or derivable from the item. For example, the author's date of birth is often not printed in a book but can be important in distinguishing among authors with similar names

(particularly parents and children). Libraries may “share” content by simply providing links, but uniform access to the content requires uniform metadata and a procedure for generating and storing it economically. It is of little avail to exchange documents at light speed if they must be held up for months until a human cataloger can prepare metadata.

Character Set Representations

This is not merely a question of different alphabets and writing systems, a major hurdle in itself, but also an issue of how characters are represented. For example, there are several widely differing mappings of Chinese characters into ASCII. There is some appeal to having a worldwide universal standard, such as Unicode, but the notion of attempting to list all of the world’s glyphs and freeze them in a standard reduces flexibility and tends to overlook obscure or variant writing systems and restrict the development of new ones. Possibly a standard should be developed that permits new character sets so long as the definition of the glyphs and the representation mapping is maintained in an accessible location.

SCALABILITY: THE BILLION-USER PROBLEM

A major problem encountered in digital library development is scalability—the expansion of system capabilities by many orders of magnitude. For example, a Web site, even one with huge capacity, may be choked if many people access it at the same time. Assuming that before long approximately a billion people will be able to connect to the Internet, if only one percent of them are interested in a topic (a number that is far too low for subjects of global concern such as the death of Princess Diana), that is a collection of 10 million people. If a server requires 100 milliseconds to grant access to a Web page, then the population would have to wait 12 days for everyone to see the same page. Therefore, technology that seems instantaneous when used on a small scale may become impossibly cumbersome when expanded.

One can imagine speeding up access to a page by adding more servers in response to anticipated demand, but even this numerical solution does not scale. If the problem is delivery of an HDTV movie (which takes 10 seconds to download at 10 gigabits per second), distributing the film to even one million people (a tenth of a percent of anticipated net users and fewer than the attendance at a major film during its first weekend of release), would require 120 days. Increasing the number of servers by an order of magnitude would not make the delay even remotely tolerable.

Bandwidth scalability is largely a hardware and networking problem. Keyword searching presents a problem of an entirely different sort. The commercial Web searchers now index approximately 50 million documents. A search can easily return 1,000 hits. This is a number small enough that a user could consider glancing at all of them to find what he wants. If the corpus being searched contained 50 billion pages (less than the number of pages in all books), a search might return a million hits, which would instead require a lifetime of effort to review. Therefore, building a digital library index, particularly one to be shared among many libraries, is not simply a matter of building a large one. Access methods, screening and navigation tools must also be provided.

Even if a library has a few million books, its staff members can be generally familiar with the nature and extent of its holdings. A library with a billion books and several billion other items would be qualitatively different and probably beyond the ability of any person to master. The sheer volume of transactions, catalog records, new acquisitions and help requests would be overwhelming. This is particularly true if the library permits access by computer programs as well as humans. It is apparent then that new organizational concepts on a grand scale will be required if digital information systems are to scale properly.

SYSTEMS AND ARCHITECTURE

A digital library “system” is composed of multiple hardware and software components, including the following:

- scanners
- computers and servers
- storage devices
- media
- catalogs
- converters
- networks
- displays
- multimedia interfaces
- usage measurement software
- human processing procedures
- reference assistants

All of the above are interconnected through networks and gateway software and must be designed for scalability, interoperability and reliability. This is a daunting challenge, particularly since the technology in many of the areas is changing rapidly. A system that uses one speech recognition system for input must be designed so that the recognizer can be replaced with another very quickly. Otherwise, the digital library, potentially one of the most responsive and useful systems in the world, can become fossilized.

SEARCH

The problem of locating items in libraries is frequently referred to as “search,” although that word tends to imply that one knows in advance what one is looking for, and possesses handles, indicators or index terms to serve as finding aids. This narrow view ignores the activity of browsing or even the higher-level function of becoming acquainted in general with a library’s holdings. Browsing in a traditional library is a physical activity—it involves scanning shelves on which related works have been placed in proximity, and occasionally withdrawing them from the shelves for examination. Browsing in a digital library is a logical activity mediated by a computer. It does not require physical proximity in any sense; indeed, two consecutive items examined may be stored on different continents. The question, then, is how can a library user (not to say the library staff) become familiar with the whole of recorded human information in a way that makes it accessible and useful?

We adopt the term “navigation” to mean moving about in a digital collection. Search is a directed form of navigation in which the goal is defined in advance with reasonable clarity. The result of a search may be an item, a collection of items, or any part of an item, even down to a single glyph. Tools must be provided that enable users to move about at varying levels of granularity within the corpus.

The usual requirement for a search is that the user is looking for a specific piece of information or a summary of what is available about a certain topic. A common case is that the user wants the answer to a specific question, such as when the postcard was invented. Only rarely does such a question translate naturally into a keyword query. Such retrieval is indirect in the sense that the user wants to learn A, but formulates a query B, to which he receives a set of retrieved documents that must be scanned to determine whether the answer to A is among them. It would be far better simply to allow the user to ask question A instead of requiring him to convert it to some query language.

Non-Textual Matter

The existence of Web searchers proves that text can be searched without being indexed or cataloged. At least on a microscopic level, documents can be located purely by their content. Many documents consist of text plus other information such as mathematical equations, tables and drawings that themselves cannot be searched directly but can often be located by the presence of related text. Purely non-textual matter is very different. Although substantial progress is being made on video searching (through the use of extensive

captioning cues, speech recognition and other aids), content searching of music and visual materials is non-existent or in its infancy. The problem is further complicated by the existence of work that combines media in various ways.

Translingual Issues

Most library items, particularly in non-English-speaking countries, are not in English. The central translingual library question is how users may navigate through materials in foreign languages and make effective use of them. Translingual search is currently a research problem for which obvious solutions do not work. A keyword search cannot be made multilingual merely by translating the keywords one at a time. The number of possible translations of each word may be very large, so an explosion in the number of hits may result. This approach also takes no account of idiomatic uses, untranslatable words such as particles, and numerous other language-related phenomena.

An interim solution is the use of translation assistants—programs that offer dictionary entries or partial or suggested translations of text portions. These show great promise for users who are at least partially familiar with the language of the retrieved document.

Synthetic Text

A user who is looking for general information on a particular topic is constrained in traditional libraries to go to an encyclopedia (which may have no entry or an outdated one on the topic of interest) or to refer to books that are generally about the subject under consideration. The time necessary for the user to obtain an overview at the appropriate level may be large because of the volume of repetitive material obtained. Programs are needed that are able to scan hits with the particular query in mind and produce abstracts, summaries, translations or analyses of the retrieved material.

INFORMATION RELIABILITY

It seems inevitable that the class of works available through digital libraries will include electronic-only publications, ephemeral and unreviewed materials and even fabricated or counterfeit matter. The ease of publishing on the Internet combined with the absence of traditional methods for evaluating reliability makes it likely that library users will be retrieving works of questionable authenticity and value. Issues concerning the Internet and digital materials include:

- Reliability. How can a user (or an automated agent) evaluate the reliability of digital materials? What information must be maintained about the source of the item and its creator to facilitate a decision?
- Version control. How can changes made to a document be tracked and the appropriate catalog entries updated?
- Archiving. What assurance can there be that the digital materials will be retained somehow in their original form for an indefinite period?
- Authenticity. How can the genuineness of materials be assured?
- Reviews. The system should allow the user to scan reviews of the retrieved work and then add his own reviews or comments to a database.
- Citations. How may a user readily learn which works have cited the retrieved work, either favorably or unfavorably?

ECONOMICS AND POLICY

While a huge amount of material is in the public domain and may be freely assimilated into a digital library, the most valuable items are recent and protected by copyright. In order to induce copyright owners to allow their content to be accessed or downloaded from digital libraries, mechanisms need to be developed to compensate them appropriately. In the most extreme case, an author might himself produce but a single electronic copy of a work. In order to justify his effort, he might have to sell it for \$100,000. Such a sale

would be impossible if the buyer were not able to charge for use of the material, and in fact charge enough to make a profit.

Fortunately, digital libraries theoretically permit precise measurement of the use made of content. A secure browser, for example, might prevent copying, printing or retransmission of material. Automated permission systems can be developed whereby users can pay directly for certain kinds of licenses. These in turn require metadata concerning the collection of rights the library has obtained for the item.

However, the implementation of charging requires another paradigm shift. The cost of building and maintaining traditional libraries is borne by governments, foundations and corporations, but hardly ever by individuals directly. Usage of materials is free, despite the high cost of maintenance. Note that authors receive substantial money on account of libraries, because currently each library that wants a book must purchase a copy of it, and the authors of popular books receive large royalties. In the digital world, the following are necessary to preserve this revenue stream:

1. centralized organizations finance
2. subscription fees
3. fees for individual use

POLICY

Digital library policy includes several areas:

- International cooperation. Many antiquities, national libraries and much television content are government-owned. Will governments share their materials with others? Can the world's nations cooperate to build a worldwide digital library?
- Government vs. private funding. How will digital libraries be funded?
- National priorities. Are digital libraries national priorities to be regarded as fundamental infrastructure such as roads, or must they compete with other projects for funds?
- Allocation of resources. Which works will be digitized first? Should priority be given to items that are decaying? How are budgets to be divided between software/hardware and research/development?
- Librarianship. How will educational programs for digital librarians develop?
- Copyright laws and conventions. Are the laws of various countries conducive to digital exchange of information? Can content holders prevent the use of materials in a digital age?

REFERENCES

- Gore, Albert. 1998. The digital earth: Understanding our planet in the 21st century. Speech delivered at the California Science Center, Los Angeles, California. January 31. (text available at www.opengis.org/info/pubaffairs/ALGORE.htm).
- MITI. 1994. Program for advanced information infrastructure. Japan Ministry for International Trade and Industry. May.