Digital Libraries: Social Issues and Technological Advances

Digital Libraries: Social Issues and Technological Advances

HSINCHUN CHEN

Artificial Intelligence Lab
Management Information Systems Department
University of Arizona
Tucson, Arizona 85721
USA
hchen@bpa.arizona.edu
http://ai.bpa.arizona.edu

ANDREA L. HOUSTON

ISDS Department
3194B4 CEBA
E. J. Ourso College of Business Administration
Louisiana State University
Baton Rouge, LA 70803
USA
ahoust2@lsu.edu

Abstract

The location and provision of information services has dramatically changed over the last ten years. There is no need to leave the home or office to locate and access information now readily available on-line via digital gateways furnished by a wide variety of information providers, (e.g. libraries, electronic publishers, businesses, organizations, individuals). Information access is no longer restricted to what is physically available in the nearest library. It is electronically accessible from a wide variety of globally distributed information repositories—"digital libraries".

In this chapter we will focus on digital libraries, starting with a discussion of the historical visionaries, definitions, driving forces and enabling technologies and some key research issues. We will discuss some of the US and international digital library projects and research initiatives. We will then describe some of the emerging techniques for building large-scale digital libraries, including a discussion of semantic interoperability, the "Grand Challenge" of digital library research. Finally, we offer our conclusions and a discussion of some future directions for digital libraries.

ADVANCES IN COMPUTERS, VOL. 48
ISBN 0-12-012148-4

1. Introduction

Over the last ten years, the way one finds or acquires information has dramatically changed. It is no longer necessary to leave the home or office to locate and access the vast amounts of information now readily available online via digital gateways, furnished by a wide variety of information providers (e.g. libraries, electronic publishers, businesses, organizations, individuals) [99]. Information access is no longer restricted to what is physically available in the nearest library but is electronically accessible from a wide variety of globally distributed information repositories.

Information is no longer simply text and pictures. It is electronically available in a wide variety of formats many of which are large, complex (i.e. video and audio) and often integrated (i.e. multimedia). This increased variety of information allows one to take virtual tours of museums, historical sites and natural wonders, attend virtual concerts and theater performances, watch a variety of movies, and read, view or listen to books, articles, lectures and music, all through digital libraries.

In this chapter we will focus on digital libraries, starting with a discussion of the historical visionaries, definitions, driving forces and enabling technologies and some key research issues. We will discuss some US and international digital library projects and research initiatives. We will then describe some of the emerging techniques for building large-scale digital libraries, including a discussion of semantic interoperability, the "Grand Challenge" of digital library research. Finally, we offer our conclusions and a discussion of some future directions for digital libraries.

2. Digital Libraries: Historical Overview

The ideas of several visionaries and their dreams for the future helped initiate the digital library concept. Three, whose ideas predate the computer, had visions based on entirely different technologies (microfilm and analog machines), so it is their ideas not their designs that predicted digital libraries.

Watson Davis founded the American Documentation Institute (ADI) in 1937 (renamed the American Society for Information Science (ASIS) in 1968). Davis was a pioneer in the field of subject indexing and was interested in documentation and in document classification and distribution. He interacted with the other visionaries of the time: Vannevar Bush (who discussed methods for indexing microfilm documents) and H. G. Wells.

In 1937 Davis published an article in *Science News Letter* (now *Science News*) predicting "a new way of duplicating records, manuscripts, books and illustrations" which would significantly change scholarly communication. The article also predicted greater information accessibility and inexpensive publication of manuscripts [72]. Davis foresaw the proliferation of journals and the difficulty of locating and integrating relevant information from several different journals. Other Davis ideas related to digital libraries include [23]: (1) *one big library*—from which users can order information from the nearest library with guaranteed prompt delivery and reasonable prices; (2) *auxiliary publications*—a repository of document copies available on-demand, replacing paper-oriented publications; (3) *one big journal*—one central location listing *all* publications in one index; and (4) *a world brain*—an idea presented in 1937 by H. G. Wells.

Herbert George Wells was an English novelist best known for science fiction (e.g. *The Time Machine*, *The Invisible Man*, and *The War of the Worlds*). Inspired by discussions with Watson Davis, Wells proposed a *World Encyclopedia* (presented at the 1937 World Congress of Documentation in Paris and described in his book, *World Brain*). He believed that the assembling and distribution of world knowledge was extremely ineffective, handicapping anyone relying on organized information for decision making. His solution, *Permanent World Encyclopedia*, was a *distributed* microfilm-based repository of all world knowledge divided into different topics, subtopics and levels consisting of "selections, extracts, quotations, very carefully assembled with the approval of outstanding authorities in each subject, carefully collated and edited and critically presented ... alive and growing and changing continually under revision, extension and replacement from the original thinkers" [72, 98].

His quote: "The time is close at hand when any student, in any part of the world, will be able to sit with his projector in his own study at his or her convenience to examine *any* book, *any* document, in an exact replica" [98] is now almost true, especially in the network form that he envisioned. The idea of a repository or encyclopedic intelligence is still being pursued (e.g. the CYC project initiated by Douglas Lenat in 1985).

Vannevar Bush was an American electrical engineer who oversaw US government support of scientific research during World War II. In 1945, Bush published an essay "As We May Think" in *Atlantic Monthly* describing a hypothetical information retrieval and annotation system called "Memex", a personal desksize machine that would store and retrieve abstracts of scientific articles [8]. Some interesting "Memex" features from a digital library perspective are: (1) using technology to organize and retrieve information, (2) creating "trails" (chains of associations) of information, (3) electronically annotating existing documents, (4) the idea that a desksized machine could store and use the contents of an entire university library, and (5) the idea of a community of scholars sharing and exchanging information. The descriptions of "Memex" trails sound remarkably like hypertext links. Although his technical implementation designs for "Memex" were not achievable, the description is similar to the gateway access digital libraries provide.

Prior to the Second World War, Bush described the "information overload" problem. In his words: "scientific literature is expanding faster than man's ability to understand, let alone control it ... publication has been extended far beyond our present ability to make real use of the record" [9]. The problem as he saw it was that "before knowledge can be used it has to be selected out of an undifferentiated mass ... knowledge that can not be selected is lost" [65]. Bush recognized that associative indexing could mimic human associative memory. He was concerned about the problems humans encounter when searching information indexed by an unfamiliar vocabulary or ontology—a phenomenon called the "vocabulary problem" or "semantic barrier" [63]. He blamed the problem on the "artificiality of systems of indexing" [9] rather than on the ambiguity of language.

There are many others who were influential in setting the stage for digital libraries. Some of them include:

- *Warren Weaver*—an MIT professor who in 1945 wrote an essay proposing machine translation [97], which initiated a stream of research on machine translation and on statistical approaches to language and text analysis (the latter made famous by Gerald Salton [81], generally considered the father of modern information retrieval).

- *J. C. R. Licklider*—a former Director of the Information Processing Techniques Office (IPTO) division of the Pentagon's Advanced Research Projects Agency (ARPA) who established the funding priorities leading to the development of the Internet, and the invention of the "mouse," "windows," and "hypertext". He is best known for three publications: (1) *Man-Computer Symbiosis* (1960) which describes his vision of computing which led to IPTO funding priorities mentioned above; (2) *The Computer as a Communications Device* (1968, co-authored with Robert Taylor), which discusses the future predicting that by the year 2000 millions of people would be on-line, connected by a global network; and (3) *Libraries of the Future* [51].
- *Ted Nelson*—influential in the development of hypertext, he coined the term "hypertext" in his paper "A file structure for the complex, the changing and the indeterminate" presented at the 1965 ACM national conference and is known for the Xanadu project.
- *William Wulf*—who wrote a white paper on national collaboratories in 1988 while working for NSF and has been influential in collaboratory research [18, 103].

3. What is a Digital Library?

What exactly is a digital library? A good general definition is that it is an entity concerned with "the creation of information sources and the movement of [that] information across global networks" [1] specifically identifying and delivering *relevant* information to interested parties. A digital library could be characterized as "a collection of distributed autonomous sites (servers) that work together to give the consumer the appearance of a single cohesive collection" [1]. In practice, each site will most likely store a large amount of information in a wide variety of formats on a wide variety of storage media [53]. Individuals accessing the information will have a wide range of expertise in key access-related areas such as computer literacy, collection navigation abilities, and domain knowledge [4].

Digital libraries are characterized by "collaborative support, digital document preservation, distributed database management, hypertext, information filtering, information retrieval, instructional modules, intellectual property rights, multimedia information services, question answering and reference services, resource discovery and selective dissemination of information" [30]. They allow information to be accessed globally, copied error-free, stored compactly, and searched rapidly. "A true digital library also provides the principles governing what is included and how the collection is organized" [49]. Advances in information technology have enabled digital libraries, and contributed to a blurring of traditional information-related roles. Anyone with access to the right equipment (a computer and access to storage on a server) can become an information provider and consumer.

Creating and maintaining a digital library typically involves the following phases:

- creating digital library content
- indexing and filtering information
- supporting universal access, and
- preservation.

3.1 Creating Digital Library Content

The first decision to be faced when creating a digital library is to determine what information to provide and what information to discard, either permanently (disposal) or through archiving. Unfortunately most of the information digital libraries wish to provide is not digitized, calling for additional decisions (i.e. determining digitization priorities and cost-effective conversion processes). Conversion challenges include both a technological component (e.g. advances in OCR technology and storage media) and an evaluation component (e.g. selection, prioritization, and choosing the appropriate level(s) of digitized quality).

Another major problem is the dynamic nature of digitized information [40]. The content can change over time, requiring the storage of either multiple copies or versions or some mechanism that allows version differentiation [50]. Other related challenges include identifying methods to capture and index continuous media in real time and techniques for processing, storing and managing vast volumes of extremely complex electronic information [1, 96].

Finally, since hypertext links allow digital libraries to provide pointers or links to information (as opposed to a digital copy), digital libraries must decide what form of access they will provide. Concerns about information ownership and archiving are becoming extremely important as a result. For example, if a digital library decides to provide a pointer or link to a certain piece of information, what happens when the owner(s) of the electronic copy decide it is no longer cost-effective to keep it? What are the responsibilities of the owner(s) to notify pointer or link owners of changes to or deletions of the original (a classic problem on the World Wide Web)?

3.2 Indexing and Filtering Information

After the acquisition and storage issues are resolved, the next set of challenges involve finding ways to make the *right* information available to the *right* individuals at the *right* time [41, 53]. Customers must be able to identify or locate potentially relevant information, filter it so that only the most relevant information is returned, and organize it (via ranking or categorization) into manageable units. Intelligent artificial agents will probably be heavily involved in future information location and filtering efforts [5, 44].

There are at least two different kinds of information location processes. The first kind is useful in a broad-based search where the information need has not yet been specifically defined. Relevant information will probably be widely dispersed among several distributed heterogeneous information sources. The key challenge is to present a seamless integration of information to the customer. Individuals interested in this kind of search will probably want the information to be summarized for quick perusal [90]. An alternative is to provide effective organization and categorization techniques that chunk information into manageable units which do not overwhelm human cognitive abilities.

The second kind of information identification process involves a very narrow, well-defined and focused search. This kind of search requires very detailed information. Because precision will be most important, effective filtering techniques will be required to return a small amount of the *most* relevant information.

In either case, the user interface will be critical. Even the most relevant information is worthless if the customer cannot understand the presentation [84]. The best digital libraries will have uniform but customizable, dynamic user interfaces that can smoothly integrate existing common data types (text, numeric, audio, video and image) from structured and unstructured sources with specialized types of data (maps, three-dimensional data, and continuous graphical data) and, potentially, new data types [1]. These systems will incorporate algorithms and techniques that enable semantic interoperability, so that humans can search in unfamiliar domains of knowledge (each with its own specialized vocabulary and ontology) using familiar vocabularies and ontologies [4, 53].

Another important aspect of information location, is finding key *relationships*, especially in distributed, heterogeneous information sources. Data mining, the extraction of patterns, associations and anomalies from large data sources [2], is a very promising research area that may produce significant payoffs for large-scale, complex multimedia digital library applications.

3.3 Supporting Universal Access

The ultimate goal of a digital library is universal access, which is consistent with the traditional library goal of providing public access to information. To accomplish universal access, digital libraries need to solve the problems of integrating distributed heterogeneous information and information sources by designing and implementing effective user interfaces that solve the "vocabulary problem" (via semantic interoperability, to be discussed in detail) [88].

One of the challenges to providing universal access is devising techniques that will support a wide variety of information display devices in handling voluminous, diverse and complex information. Not only is there a variety of operating systems in the computer domain, but there are a wide variety of display devices (e.g. palm tops, televisions, fax machines, video monitors, modems, and other information "appliances") to cope with. Accommodating legacy display devices and receivers is probably a more difficult problem than accommodating and integrating legacy information and information sources [93].

Another major challenge is the limited amount of bandwidth available for the transmission of electronic information to accommodate an increasing number of users and increasingly complex (and large) data sets. For equitable universal access to be achieved, intelligent use of the bandwidth, including the ability to guarantee bandwidth for a given period of time (in particular for law enforcement and emergency situations) must be identified and policies to support such uses enacted. These challenges, combined with the economic pressures faced by digital libraries, have led to a perception that information providers behave as "gate-keepers" of a service (or set of services), providing a "gateway" to knowledge [25, 26, 68, 78]. The current government and industry funded Internet-II project was designed to help alleviate some future problems in supporting universal access to digital libraries.

3.4 Preservation

Electronic media do not disintegrate as readily as other types. Paper is particularly vulnerable because it is susceptible to the problems of acid paper and binding disintegration as well as to destruction through innocent physical handling and vandalism. Other media (including tape, images, negatives, vinyl records, etc.) are susceptible to disintegration due to pollution, catastrophic events (floods, and other natural disasters), humidity, light, insect and other kinds of pests, mold and mildew, vandalism and human handling. However, constant changes and various enhancements in electronic document format (e.g. MARC, SGML, HTML, XML, etc.) and associated incompatibility problems will need to be addressed carefully in this digital age.

4. Drives Towards Digital Libraries

Over the past decade, several trends have steadily encouraged the transition to and expansion of digital libraries. The four major drivers are: economics, accessibility, new technologies, and standards.

4.1 Economics

It is cheaper to produce, store, distribute, and reproduce electronic information. Furthermore, digital libraries can cooperate with each other by providing a gateway (links) to information managed or provided by others, allowing specialization as well as conservation of acquisition and production budgets while still providing access to a wide range of information [49]. Other economic pressures driving libraries towards digitization include:

- *Inflation*: the rapid rise in library operating costs (especially in acquisition or collection expansion of scholarly journals). In the past 20 years, journal prices have soared by 400 percent while book and monograph prices have increased by 40% [31].
- *Volume*: the explosion in the amount, variety and complexity of information.
- *Maintenance*: the preservation crisis in existing collections, especially with regard to acidic paper (nationally the replacement cost of disintegrating print materials extrapolates to approximately $35–$45 billion [31]).
- *Multimedia*: the increasing amount of multimedia information that requires special viewing or listening facilities and different cataloging and storage requirements.
- *Collaboration*: the advantages from resource sharing among libraries and other information providers (both economically and in improved level of service).
- *Timeliness*: electronic information is easy to produce, distribute, and duplicate with few of the problems of multiple handling and redistribution, allowing a dramatic savings in costs [33].
- *Scholarly communication*: the severe cost problem associated with scholarly communication [26], in particular the excessive cost of

providing access to an *appropriate* number of scholarly journals [6, 31, 49] to maintain an *adequate* level of service. For example, according to Andrew Odlyzko of Bell Lab, "a good mathematics library spends $100,000 per year on journal subscriptions, plus twice more on staff and equipment. [The] US spends as much money buying mathematics journals as NSF [the National Science Foundation] spends on mathematical research" [66].

4.2 Improved Level of Service

Digital libraries have the ability to provide a previously unattainable level of service, i.e. "individual words and sentence search and delivery of information to the user's desk—information that does not decay with time, whether it is words, sounds or images" [49]. Information that was either previously unavailable or difficult to acquire is now readily available electronically (i.e. large government collections). Access to information can be improved in several different ways: access time (retrieval speed and/or timeliness), availability (recall), content (relevance), improved visualization (user interface) or some combination [6]. Historically, research has been focused on generic improvements to information access. The current trend is to *individually* customize or tailor user information access methods and interfaces.

Since the production and distribution of electronic information eliminates multiple handling and redistribution, there is a dramatic savings in time from production to use. Electronic information need only be created and stored *once* to be immediately available over a network *simultaneously* to multiple users as opposed to multiple copies being generated over time and provided via traditional (postal and/or manual) distribution channels [41, 49, 77]. Many Internet news Web sites offer information in real-time, for example, with no time delay in printing or delivery.

4.3 New Technologies

To effectively meet the information needs of their clients, digital libraries need to use a combination of technological advances and have the ability to design, construct, manage, and use global electronic networks [57]. They must be able to adapt rapidly to dynamic changes in technology and to cope with the size, scale, and complexity of both the networks themselves and the information available through them [4].

Many technological advances in information production, management and distribution are responsible for *enabling* digital libraries. They are too numerous to describe in detail but include such things as advances in:

(1) storage media;

(2) digitization or information capturing techniques (i.e. OCR technology);
(3) automatic indexing and organizing of large volumes of information [88];
(4) computing speed;
(5) network technology (including data compression);
(6) content-based search and retrieval [90];
(7) feature-based or texture-based search and retrieval [91];
(8) full-text indexing;
(9) resource or knowledge discovery;
(10) multimedia and hypertext;
(11) standards (i.e. Standardized General Mark-up Language (SGML) and Hypertext Mark-up Language (HTML), and Z39.50);
(12) object-oriented techniques; and
(13) improvements in user-interface design and data visualization [26, 42, 53, 76].

In the next section, we will review some of the recent technical advances in several large-scale, high-impact digital library projects. In particular, we will review new technologies that aim to improve semantic interoperability in digital libraries.

4.4 Standards

In order for digital libraries to be truly global gateways, it is important to have internationally accepted technical standards for electronic information representation, formatting, transmission, and protocols. This is the only way to ensure compatibility and therefore interoperability between equipment, data, practices and procedures necessary to achieve universal access [41, 42, 70, 93] and global electronic information exchange. Unfortunately, there are many social, cultural and political barriers to developing international standards, even when the benefit is clear to all.

Several international organizations are involved in standard development, including the International Organization for Standardization (ISO)—which was responsible for the SGML (Standardized General Mark-up Language). Another organization, IETF (Internet Engineering Task Force—see www.ietf.org) is specifically interested in Internet architecture and smooth Internet interaction and operation [42]. One of the most important standards from a digital library perspective is Z39.50 (the distributed information retrieval standard, see lcweb.loc.gov/z3950/agency/) which was adopted by ISO as the ISO 23950 standard [30, 61].

At a national level, while information and document standards such as SGML, HTML, TEI (Text Encoding Initiative), VRML (Virtual Reality

Modeling Language), and MARC (Machine-Readable Cataloging) exist, in practice most electronic information exchange occurs via e-mail, anonymous ftp, Gopher, and Web browser platforms with TeX, LaTeX, PostScript, PDF, ASCII text, and Word and WordPerfect formatted documents. Most of these formats do not have mechanisms to distinguish the contributions of multiple authors or versions, nor do they have the ability to include active links to other information. Many of the formats used in practice are commercial, proprietary, and therefore not platform independent which means they are not universally accessible. Will common practice dictate what standards become accepted or will some governing body take responsibility for thoughtful and independent (unbiased) standards development? If a set of standards is accepted and adopted, what kind of translation capability from these "legacy" formats will be provided? Careful research and policy related to standards will be needed to achieve a truly universal digital library community.

5. Digital Library Research Issues in the Social Context

Digital libraries allow people to interact with each other and information in novel ways. As is often the case with new technology, these abilities have become available ahead of the societal norms, conventions, and policies that help monitor, guide and evaluate their use. As a result, some economic, social, and legal issues have arisen in response to digital libraries and the technologies that have enabled them.

5.1 Economic Issues

While electronic information may be cheap to produce, store, modify and distribute, it is much harder to determine a fair or market price and cost for it than for a physical object. To date, there is no commonly accepted economic model that can accurately and fairly determine either costs or prices for digital library services.

Historically, information-related services were typically not broken down into individual transactions or unit priced [78]. As a result, most digital libraries and their clients have little idea what an information transaction is worth [49]. This is such a difficult challenge that Saracevic and Kantor [85] developed a derived taxonomy to address the problem of determining and measuring the *value* of a library's information and services. Customers know that electronic information has an almost zero incremental reproduction and viewing cost and therefore expect that access should be free or extremely cheap.

Digital library services are not free, however. Some method of compensation is necessary. Currently, there are at least two basic compensation models: (1) allowing free access but charging for content (i.e. freely accessing the index and table of contents, but charging for anything more) and (2) charging for access but allowing free perusal and consumption of the content [49]. These two models are not mutually exclusive and both co-exist on the Internet.

Several different digital library funding models have been proposed, but the basic models are either time-based (unlimited access for a given unit of time, e.g. a month), request-based (per request), or some combination of both [1, 26]. Some proposed models include [49]:

- institutional (public and private) subsidies—the current model for most digital libraries;
- "free" general services and charging by transaction for unusual services, especially those requiring human intervention;
- charging for everything—assuming that information services can be transactionalized and costed. Common suggestions include charging by: connect time, CPU usage, fee-per-search, fee-per-hit or retrieval, and download fees. A problem with all of these suggestions is that people generally do not understand how they work, creating a situation where charges appear to be unpredictable and resulting in unreasonable or unpredictable behavior;
- subsidizing services through advertising (typical of magazines, television, and the Web);
- other subsidizing mechanisms (for example, pledge breaks—public appeals for donations similar to public television and radio);
- taxes or other sources of public funds;
- subscriptions (pay for a given period, i.e. a year) or licenses (viable concept in the software market);
- memberships similar to "buying clubs" where individual consumers pool their resources to allow access to information (pricing issues could be resolved via price discrimination, non-linear pricing and service bundling [43]);
- charge authors a per-unit fee for the "privilege" of having their information and services accessible, then charge users for the nominal incremental cost of accessing the information (similar to a per-page author charge under consideration by some journals);
- opportunity cost—measuring the opportunity cost of providing information or a service as opposed to measuring the cost of expended resources. "Opportunity cost is determined by the relationship between

supply and demand for a given resource, so that the opportunity cost of an idle resource is close to zero but that of an over-utilized resource is so high that it is basically unaffordable" [1]; and

- using a detailed byte-by-byte charging algorithm—an interesting idea from Ted Nelson (CNRI) and CMU's NetBill project [49].

Current cost models and financial instruments used in traditional information production and consumption do not adequately address the needs of digital libraries. Fixed cost models are insensitive to changes in content and costs. Electronic information comes in a variety of formats with different associated production and distribution costs. Flexible and adaptable cost models are required to handle this diversity and complexity [1, 20]. Economic models for digital libraries require a series of specialized costing and pricing algorithms that can dynamically determine the cost and price of information or services and modify the model with a variety of environmental factors [20, 49, 70].

5.2 Legal Issues

Digital libraries exchange electronic information on a global level. Therefore, national governments will have to negotiate an international-level policy framework that can accommodate the exchange of information across international boundaries and differences in cultural values and laws (especially with respect to copyright, intellectual property, privacy, information ownership, fraud and other business crimes, taxation and currency exchange) [57, 74].

Now that almost anyone can be an electronic information provider, it is easier to perpetuate ethically questionable acts. For example, it is harder to prevent plagiarism in the digital age. The shear volume of electronic information makes it very difficult to enforce copyright laws or even to detect illegal copies. False representation and false information can easily be provided electronically, leading to concerns about information quality. These ethical considerations are challenging enough within a given nation or culture, but digital libraries are global, and different nations and cultures have very different perspectives, definitions, and social guidelines with respect to concepts such as plagiarism, copyright laws, "fair use" and "truth in advertising" [79]. How can internationally accepted ethical codes be developed and enforced in light of these issues?

In the United States, more localized legal issues surrounding digital libraries include [49]:

- *Ownership issues.* When a library owns a physical copy, decisions about acquisition and archiving are relatively straightforward. If a digital library only owns a link or gateway connection to the information, certain kinds of ownership problems arise [28]. For example, if a digital library decides to cancel its subscription to regularly published information (such as a journal), how will access be controlled? Obviously access to future issues should not be permitted but the right to access past issues has already been negotiated and paid for. Dynamically having to keep this kind of information results in complicated record-keeping and access control policies, procedures and processes for digital libraries. What should be done about an information provider that goes out of business or an information item that goes "out of print." In both cases the information provider can no longer afford (or wish) to support physical storage. How can the rights of owners of *links* to that information be protected?

- *Unauthorized access.* Electronic information appears to be more vulnerable to unauthorized access, theft and fraud than physical copies as such incursions are harder to detect. A variety of techniques are being investigated to help protect electronic information, including "firewalls," electronic signatures, encryption, special "rendering" or viewing software or hardware, and electronic watermarks.
- *Liability.* Traditionally US law distinguishes between authors and publishers who *are* held responsible (liable) for information that they have produced and distributors (the post office, libraries and bookstores) who *are not.* Digital libraries can distribute as well as produce information, raising difficult legal questions regarding their responsibility for information published, displayed or distributed from their sites. In situations where electronic information has multiple authors and multiple versions, how can expertise be determined and liability assigned?
- *Trademark infringement.* Trade-marked images (for example a state or university seal or a commercial caricature) can be copied or scanned and used as wall-paper or images in electronic information. Many organizations require notification and/or payment for using trademarks (trademark use is often interpreted as an organizational endorsement). How will these rights be protected?
- *Copyright and intellectual property rights.* Copyright issues, in particular copyright violation and the related intellectual property rights issues, are major digital library legal issues [49]. Pamela Samuelson [82, 83] from Berkeley is a well-known authority on the topic.

Virtually anything that can be copyrighted can also be digitized. Once digitized, anyone with a computer can copy it, modify it and distribute it to anyone else who has access to a network. Electronic information is easy to copy and redistribute, but it is difficult to distinguish a valid copy from an illegal one. The regulations that exist today (e.g. no downloading at all; no electronic storage—view only; no copies or distribution, even internally; no copies or distribution to third parties; and specific limitations on various types of use) are largely ignored [41]. Information providers (i.e. publishers) implicitly endorse this behavior by "looking the other way" in many instances. What are the responsibilities of digital libraries in enforcing copyright laws applying to the information and services they provide?

New copyright laws and practices, at least with respect to electronic information, are going to have to be created because the speed of technological advancements have left legal systems far behind [41, 95]. In the US, several proposals are being considered [49, 52]. Since new laws and practices must be enforceable, they must rely on new technology to help protect copyrighted material from unauthorized access [19, 32], reproduction, manipulation, distribution, and performance or display. New technology will probably also assist in the detection of copyright violations through new methods of authentication, management of copyright protected material (such as the clearing house model used predominantly by the music industry), and licensing techniques [41].

Several methods to protect intellectual property rights and copyright of electronic information are currently under investigation. They include [49]:

(1) *Fractional access*: this only applies to very large information sources (e.g. LEXIS/NEXIS) whose value lies in the *volume* of information and the knowledge that can be gleaned from analyzing the *entire* collection. There is no economic advantage to copying small portions of the data, and illegally copying the entire data source should be relatively easy to detect.

(2) *Interface control*: this requires a proprietary interface to access information, implying that universal access is no longer possible.

(3) *Hardware locks or "dongles"*: these are the hardware equivalent of interface control (a software solution). Access to information is restricted by proprietary access hardware (video games such as Sega or Nintendo are good examples).

(4) *Information repositories*: legitimate copies are *only* available from one large repository or source; any other copy is *not* legitimate. Some organizations exploring this approach include: in the US InterTrust (previously EPR—Electronic Publishing Resources) and CNRI (Corporation for National Research Initiatives, currently working with the US Copyright Office) and in Europe, Imprimatur

(Intellectual Multimedia Property Rights Model and Terminology for Universal Reference).

(5) *Steganography*: the embedding of hidden messages in information. Each legal copy is labeled with a different identification number allowing illegal copies to be tracked back to the original purchaser (i.e. "digital water-marks"—see [34] for a good example). The major problems are that the "hidden" codes or messages are easy to remove, hard to insert and while the method appears to work with complex images it does not work with simple text and may not even apply to audio data.

(6) *Encryption*: information is encrypted (sometimes in cryptolopes or secret envelopes), and cannot be interpreted without an encryption key (software or hardware dependent).

(7) *Economic approaches*: identifying ways to make it uneconomical to pirate or illegally copy electronic information. Ideas include: provider page charges to reduce the per-copy price, site licenses to reduce on-site cheating, and advertiser supported publications.

(8) *Flickering or "Wobbling"*: information technology that allows a customer to view but not capture information [48].

5.3 Quality and Security Issues

Publication (especially by reputable publishers and editors) lends credibility to information *content*. There is less concern about fraud, plagiarism, and unreliable or invalid information with non-electronic formats. Most non-electronic information (especially scholarly information) is subject to some kind of peer review or validation process, further augmenting perceptions of quality. Unfortunately, quality is harder to determine in on-line information. None of the existing searching engines has a way of evaluating electronic information quality (traditional information quality cues are typically missing) and therefore no way of sorting or filtering information by quality.

Lack of information about quality tends to limit searching to known experts (authors, Web sites, organizations, publishers or journals, etc.) or information recommended by known experts. Some search engines do use either an individual expert's profile or a group profile to request information (i.e. "give me information identical to what Joe Einstein requests"). Information integrity is still a problem, however, because electronic information is so easily modified. Electronic information is rarely guaranteed to be truly generated or endorsed by the expert. Currently, electronic scholarly publishing is a hotly debated topic. Much of the concern centers around questions of information and intellectual quality [26, 33, 71].

Issues of security and control are related to quality issues. Digital libraries need to address security in at least four areas [1, 41]:

(1) *Confidentiality*: protecting access to the information content (especially sensitive information, such as personal, financial or health information, and strategic business or national information) from unauthorized access and distribution.

(2) *Authenticity*: attributing information to the correct author(s) and validating it as original, accurate, and correctly attributed. This can be especially difficult in a multi-authored and multi-versioned environment [99].

(3) *Integrity*: protecting information content from unauthorized modification. This type of security involves a balance between easily enabling authorized updates and preventing unauthorized ones. The authenticity of modifications must be verifiable (challenging in a multi-author, multi-versioned environment).

(4) *Privacy*: protecting information access and usage patterns from unauthorized access and resale.

An important challenge for the implementation of any security technique is to balance the need for security with the need for performance (access and timeliness). Authorized access and modification must not be so difficult that it is never attempted, or abandoned before completion. Likewise, while validation techniques must be as accurate as technically feasible, they cannot be so time and resource intensive that the accessibility and timeliness of the information is compromised. Information is not valuable if it is not accessible, timely or useful.

5.4 Social Issues

There are several kinds of social issues faced by digital libraries. The major ones include:

- *Literacy*: in order to use a digital library a certain basic level of education or training is required, (i.e. a basic competence in the operation of a computer). Who will be responsible for providing basic computer skills and training? Should training be freely available through public education systems or should it be part of the services provided by digital libraries? Will access to training as well as access to the appropriate equipment (computer) and facilities (an account and storage space on a server) separate society into information "haves" and "have-nots"? If so, what are the implications of this division?

- *Cultural biases*: filtering and organizing electronic information to assist in coping with information overload is a useful service. Unfortunately, the result, deliberate or unintentional, is that the cultural biases and social values of the service provider are imposed [6]. The simplest example of this is language bias. Should information be accessed in the language in which it was generated, or should part of a digital library's information service be the ability to translate information into the customer's language of preference? One interesting approach in Japan is the MHTML project which provides multilingual browsing capability to a collection of Japanese fairy tales. The server can handle Japanese, Korean, Chinese, Thai, and several European languages including French and English [55].

  Translation of words (written or spoken) is relatively straightforward, but what about translation of non-language information (i.e. images or music)? Furthermore, information considered publicly appropriate for one group of people may be offensive or even illegal for another [42]. One solution may be to develop highly sensitive and individually customizable user-interfaces that could accommodate a given individual or group's cultural and linguistic preferences [29].

- *Ethical considerations*: the traditional librarian's perspective is that libraries have a responsibility to ensure socially and economically equitable public access to information [6]. However, not all governments, organizations or social groups support universal access and may, indeed, actively attempt to restrict access to certain kinds of information deemed inappropriate.

  Universal access also involves ethical issues related to censorship and cultural bias. Not all information is appropriate for all groups. Different individuals and cultures have different opinions about the accessibility and even the definition of material that could be considered inappropriate due to background (racial, religious, cultural), sex (including sexual preference), age and health. Included here is such information as: pornography, material generated by hate groups and other racial or religious persecutors, sexual predators (particularly child predators), drug dealers, terrorists and other criminals [49]. Should there be a limit on what information a digital library can provide access to? How could such a limitation be imposed and by whom? Or should there be limits on what kinds of information an individual can receive (who has the authority to determine and impose such limitations, and how might it be accomplished)?

- *Equality*: this issue addresses questions such as, is there equal access to information and do individuals have an equally likely chance of providing it. Experiences from some forms of electronic scholarly publication (i.e. e-print archive for high-energy physics) is very positive. Access is more equal in the electronic version than the printed version as the information is posted, and anyone can access it [33]. There are other instances, in the vast biomedical collections for example, where the volume of information is so huge that information is typically requested by a very small set of well-known and respected authors, journals, research centers, or some combination of those, making it extremely difficult for newcomers to get recognized and accepted. Fears about the lack of quality control in electronic information drive this tendency even more dramatically. There may be no difficulty in equitably providing electronic information, but how can consumers be encouraged to access it equitably?

6. Digital Library Research Activities: An Overview

In the last five years, there has been an explosion of digital library research, digital library initiatives and digital library programs both in the United States and around the world. In briefly describing some of these efforts, we have tried to suggest their breadth and variety. As these programs are constantly growing, expanding and changing, the most current source of information about them can be found through their Web pages, from which we have heavily drawn in the following sections. We apologize if the sites that we list have changed or are no longer active. There are many more digital library efforts than can possibly be included in this chapter. We have tried to offer descriptions of programs with which we are most familiar or consider interesting. Obviously, there are many interesting efforts that we may have missed.

The focus of digital library initiatives can either be *research* or technically oriented, primarily coming out of major research-oriented computer science departments, and emphasizing new methodologies, new technologies, new tools and new methods of storing and accessing information, or it can be *library* or socially oriented, examining collections and social issues and concerns related to digital libraries. In the United States, the federally funded digital library initiative projects are at the research end of the spectrum, while most of the digital library projects in other countries, like Japan, are more library and collection oriented [49]. However, many large-scale digital library projects incorporate both components, but in different proportions.

From an international perspective, the first US federally funded Digital Library Initiative (DLI), described in more detail below, has made at least two fundamental contributions. First, the research effort is substantial and is beyond the scope and definition of other digital library projects. All DLI projects are heavily research oriented and focus on exploring and identifying new scalable technologies for large-scale digital library content and users. Second, each project has been given a great deal of autonomy in defining the scope and focus of its research. While any large digital library effort requires resources commonly found only at the national level, and indeed most of the international digital library efforts are national government level efforts, this US effort essentially provides research "seed money," with minimal governmental intervention [42]. The response and support of this four-year research effort from the digital library community, both in the US and internationally, have been overwhelming.

6.1 Digital Library Research Activities in US

Many digital library research activities and programs are in progress in the United States and the number is rapidly increasing every month. Most of the research and work has occurred in the public sector and in institutions of higher learning, although the private sector has often generously contributed funding and technology exchange. As electronic commerce increases in its influence and its impact on the US economy, we believe that the private sector will become more involved in digital library research programs.

Because the US federal government has been responsible for the majority of the funding for US digital library efforts, this section is divided into two parts: US digital library initiatives that are federally funded and others. It is clear that the US government intends to continue funding this important research. Two new federally funded US digital library initiatives are the DLI-2 (Digital Library Initiative—Phase 2) announced in the spring of 1998, and the International Digital Library Initiative (announced in the fall of 1998). The international initiative will be jointly funded by the US and other nations' governments and organizations and will stress international cooperation and internationally distributed and accessible collections.

6.1.1 *The NSF/NASA/DARPA Funded Digital Library Initiative (DLI)*

In September 1993, the National Science Foundation (NSF) announced a jointly funded digital library initiative in conjunction with the National Aeronautics and Space Administration (NASA) and the Defense Advanced Research Projects Agency (DARPA). A total of $24.4 million was awarded to six research consortiums composed of major research universities, industry and other interested organizations. This initiative has been the flagship research effort for the National Information Infrastructure (NII) program.

The projects' focus was to be on dramatically advancing the *technology* involved in collecting, storing, and organizing digital information, and making that technology available for searching, retrieval and processing via communication networks. Particular emphasis was to be placed on user-friendliness and leveraging prior research in high performance computing and communications technology. The four-year projects (1994–1998) are centered at Carnegie Mellon University, the University of California, Berkeley, the University of California, Santa Barbara, the University of Illinois at Urbana-Champaign, the University of Michigan, and Stanford University.

The digital library effort was identified as a "National Challenge" (a fundamental application that has broad and direct impact on US competitiveness and the wellbeing of its citizens) in the Information Infrastructure Technology Applications component of the US High Performance Computing and Communications Program (HPCC). The key technological issue has been how to index, locate and display information of interest from very large, distributed (potentially internationally distributed) digital information repositories (or collections). In essence this involves developing infrastructures or architectures and tools for multimedia information retrieval on the Internet [88]. The six projects are briefly described in alphabetical order in the next six subsections.

6.1.1.1 Carnegie Mellon University: Informedia (www.informedia.cs.cmu.edu/) The title of this $4.8 million project is: "Full Content Search and Retrieval of Video." It is also called the Informedia project. The principal investigator is Howard Wactlar of the School of Computer Science. Industrial partners and collaborators include:

- *corporations*—BBC (British Broadcast Communications), Bell Atlantic Network Services, The Boeing Company, Digital Equipment Corporation, the Vira I. Heinz Endowment, Intel Corporation, Microsoft Inc., Motorola, QED Communications (WQED in Pittsburgh, Pennsylvania, a public television station); and
- *educational institutions*—the Fairfax Virginia County Public Schools, the Open University in the United Kingdom, and the Winchester Thurston School in Pittsburgh, Pennsylvania.

The testbed database consists of 1000 hours of digital video from the archives of public television station WQED/Pittsburgh, Electronic Field Trips on video from the Fairfax County Virginia public school system and video course material produced by the BBC for the Open University. The research is primarily concerned with creating and searching this interactive on-line digital video library by capitalizing on a number of techniques for which Carnegie Mellon University is famous [49]:

- *image analysis*—partitioned video, image feature, frame classification from the Machine Vision project,
- *speech recognition*—the Sphinx, HEARSAY projects,
- *face recognition*—matching faces to names in voice transcript, and
- *natural language understanding*—using content-based video retrieval, video paragraphs, and segmentation and experience from the Tipster, Ferret and Scout projects.

The project is also investigating various methods of protecting the collection's intellectual property data rights and techniques to provide security and privacy through network billing, variable pricing and access control (NetBill project).

The Informedia project digital video library system was created by Carnegie Mellon University and WQED/Pittsburgh and allows users to access, explore and retrieve video science and mathematics materials. Informedia integrates speech, image and natural language understanding technologies. The digital library is populated by automatically encoding, segmenting and indexing video data that is partitioned into scenes using image analysis (image and frame features), speech recognition (including both closed captioned transcripts and automatically generated transcripts using the CMU speech recognition abilities from Sphinx and HEARSAY projects), face recognition (matching faces with names in the transcripts), and natural language understanding. The collection is chunked or segmented into video clips or video "paragraphs" (using visual scenes or conversations as boundaries). Video data contains temporal and spatial information and is typically massive and unstructured, making it extremely difficult to segment. The ultimate goal is full-content and knowledge-based search and retrieval of the video collection. Usability studies are conducted using K-12-age children from the Winchester Thurston School in Pittsburgh, Pennsylvania and investigate questions related to human factors such as learning, interaction and motivation [96].

Video user-interface requirements are more challenging than those of other multimedia collections, so various new techniques and approaches are under investigation. Especially challenging is the need for data representation for video clips which will allow "video skimming" by users to determine the relevance of a particular video segment. Problems encountered in the resolution of video error and variability caused by music and noise mixed with speech, segmentation of long fragments of video, inappropriate language models, error-prone closed-captioned data and scripts, acoustic modeling, identification of speaker change, and speech recognition for keyword retrieval [96] are also being studied. Issues involving human-computer interaction, pricing and charging for digital video use, and privacy and security are being addressed as part of the research program. More detail is available from the project's homepage listed above, including links to a series of publications on the project [96].

6.1.1.2 University of California at Berkeley (elib.cs.berkeley.edu/) The title of this $4 million project is "Work-centered Digital Information Services." The principal investigator is Robert Wilensky of the Computer Science Department. Partners and collaborators in the project include:

- *University of California partners*—Office of the President, University of California at Berkeley Division of Computer Science, Office of Information Systems and Technology, Research Program in Environmental Planning and Geographic Information Systems, and School of Information Management and Systems;
- *State and federal agencies and organizations*—California Department of Fish and Game, California Department of Water Resources, California Environment Resources Evaluation System, California Land Use Planning Network, California Resources Agency, California State Library, San Diego Association of Governments, Shasta County Office of Education, Sonoma County Library, and USDA Forest Service;
- *Industrial partners*—Hewlett Packard (HP), Informix, IBM Almaden, Phillips Research, The Plumas Corporation, Ricoh California Research, The State of California Resources Agency, Sun Microsystems, and Xerox Corporation (Xerox PARC); and

The primary testbed database contains environmental information. Primary research topics for the project include [49]:

- *image content queries*—under investigation primarily at Xerox PARC using the Cypress engine and metadata or derived data generated at data acquisition time;
- *techniques for database extraction from documents*—including a variety of data formats from tables to spreadsheets or databases;
- *multivalent documents (MVD)*—a new digital document model which involves the creation of multiple potentially distributed layers of the same document, each containing different "behaviors;"

- *natural language processing (NLP)*—lexical disambiguation, and statistically based NLP; and
- *automatic categorization*—TileBars based on TexTile analysis which represent document content.

Other research areas include: automated indexing; intelligent retrieval and search processes; database technology to support digital library applications; new approaches to document analysis; and data compression and communication tools for remote browsing. A prototype digital library was created and contains environmental information to be used for the preparation and evaluation of environmental data, impact reports and related materials. The research prototype will eventually be deployed in the State of California as a full-scale CERES production system.

The project's main goal is to develop technologies for intelligent access to massive, distributed collections of photographs, satellite images, maps, full text documents, and "multivalent" documents. To accomplish this the project has focused on user needs assessments and a "simple" architecture consisting of information repositories, interoperable clients, indexing and searching techniques, interoperability mechanisms and protocols (e.g. ZQL—a Berkeley designed protocol that combines the SQL and Z39.50 protocol standards). The UC Berkeley project has created an interesting approach to image representation, "Blobworld," which locates objects by grouping low-level image properties (color, texture, symmetry) together into coherent units in a hierarchical manner. Users can query the image database by indicating on a given photograph regions central to their query. Blobworld returns images that contain regions that match user input. More detail can be obtained from the project's homepage mentioned above (including some very interesting demonstrations of their work and a list of publications) [100].

6.1.1.3 *University of California at Santa Barbara: Alexandria (alexandria.sdc.ucsb.edu/)*

The focus of this $4 million project is on "Spatially-referenced Map Information." It is also called ADL (Alexandria Digital Library). The principal investigator is Terrence R. Smith of the Departments of Computer Science and Geology. Partners include:

- *academic groups from the University of California at Santa Barbara*—Center for Remote Sensing and Environmental Optics, Departments of Computer Science and Electrical and Computer Engineering, Graduate College of Education, Map and Imagery Laboratory, and National Center for Geographic Information and Analysis (NCGIA);

- *academic researchers from NCGIA at other institutions*—University of Maine at Orono and State University of New York at Buffalo (SUNY-Buffalo);
- *libraries*—Library of Congress, St Louis Public Library, SUNY-Buffalo Library, UC Center for Library Automation and US Geological Survey Library; and
- *industrial partners*—AT&T, Conquest, Digital Equipment Corporation (DEC), Environmental Systems Research Institute (ESRI), and Xerox.

The project testbed database consists of maps, aerial photographs, atlases, gazetteers, and other spatially indexed information. The research focus is on [49]:

- *spatial indexing and retrieval*;
- *rapid response to image data queries*—including multi-resolution image storage and display;
- *image processing using features*—(i.e. texture, color, shape, location); and
- a variety of problems related to a distributed (i.e. across the Internet) digital library for geographically referenced information (i.e. all the objects in the library will be associated with one or more regions, or "footprints," on the surface of the Earth).

The primary project goal is to provide a comprehensive set of library services for spatially indexed and geographic information. The user interface supports both textually based and visually based queries focusing particularly on content-based searching. The collection is indexed using both top-down techniques (metadata based on a combination of USMARC, US Machine-Readable Cataloging, and FGDC, Federal Geographic Data Committee standards) and bottom-up automatic techniques.

The project is centered at the University of California, Santa Barbara (UCSB) because of its major map and image collections and its strong research focus in the area of spatially indexed information. The project is expected to expand to include other UCSB components, and other interested libraries such as the SUNY-Buffalo Library, the Library of Congress, the United States Geological Survey Library and the St. Louis Public Library. Facilities for geographical information interfaces, electronic catalogues, and information storage and acquisition will be included at each prototype test site. More detail is available from the project's homepage mentioned above (including a demonstration and a list of publications), and from [91] and [92].

6.1.1.4 University of Illinois at Urbana-Champaign: Interspace (dli.grainger.uiuc.edu/) The focus of this $4 million project is on "Federating Repositories of Scientific Literature." It is also called the Interspace Project. The principal investigator is Bruce Schatz of the Graduate School of Library and Information Sciences and the National Center for Supercomputing Applications at Illinois. Other principal researchers include: Ann Bishop, Hsinchun Chen (University of Arizona), and Bill Mischo. The principal partners include:

- *publishing partners and professional societies* (providers of digitized material)—Academic Press, Inc., American Association for the Advancement of Science (AAAS), American Astronomical Society (AAS), American Chemical Society (ACS), American Institute of Aeronautics and Astronautics (AIAA), American Institute of Physics (AIP), American Physical Society (APS), American Society of Agricultural Engineers (ASAE), American Society of Civil Engineers (ASCE), American Society of Mechanical Engineers (ASME), Institution of Electrical Engineers (IEE), Institute of Electrical and Electronics Engineers (IEEE), IEEE Computer Society, Institute of Physics, Tribune Company, US News and World Report, and John Wiley & Sons;
- *software and hardware providers*—BRS/Dataware, Hewlett Packard, Microsoft, Inc., NETBILL, OpenText, SoftQuad, Spyglass, and United Technologies; and
- *others*—The University of Arizona, Carnegie-Mellon University, CIC Consortium (members of the Big Ten Universities), Corporation of National Research Initiatives (CNRI), and National Center for Supercomputing Applications (NCSA).

The testbed database consists of a collection of engineering and science journals and magazines. The technology focus of the project includes investigating:

- conversion of SGML to HTML documents and the conversion of digital non-SGML documents to SGML;
- semantic retrieval using noun phrases—concept spaces;
- automatic semantic categorization using neural networks—category maps;
- supercomputing simulation of large-scale semantic analysis;
- information representation or visualization; and
- semantic interoperability or vocabulary switching across knowledge domains.

The project is based in the Grainger Engineering Library Information Center at the University of Illinois at Urbana-Champaign and is centered around engineering and science journals and magazines. The testbed prototype (called DeLIver) is a production facility at the university library. It contains hundreds of thousands of full-text documents and is accessible to thousands of users at the University of Illinois. The testbed software supports comprehensive indexing, searching, and display of the complete contents of the collection including text, figures, equations, and tables. The primary focus is on seamless *textual* information retrieval across distributed, heterogeneous digital information repositories. However, a joint project with the University California at Santa Barbara, The University of Arizona and the University of Illinois at Urbana-Champaign is also investigating information retrieval using *texture* for images (in particular, satellite images and aerial photographs) [17].

Indexing the collection combines humanly determined and assigned meta-data (a top-down approach) and automatically generated, statistical co-occurrence analysis based indexing terms (a bottom-up approach). Research includes sociological evaluation of the testbed (including a user evaluation study), technological development of *scalable* deep semantic retrieval and data visualization techniques, and a prototype design of a future scalable information system (the Interspace). The ultimate goals of the project are to bring professional quality index, search and display capability to a large digital collection that is accessible via the Internet and to develop and implement an Interspace prototype (a vision of future evolution of the Internet). More detail is available from the project's homepage listed above, which includes a list of publications and presentations on the project and from [86] and [90].

6.1.1.5 University of Michigan: UMDL (www.si.umich.edu/UMDL/) The focus of this $4 million project is on "Intelligent Agents for Information Location." It is also known as UMDL (University of Michigan Digital Library). The principal investigator is Daniel Atkins of the School of Information. Partners include:

- *library partners*—Ann Arbor Public Library and New York Public Library;
- *school partners*—in Ann Arbor, Michigan: Huron High School, Community High School, Roberto Clemente Student Development Center, and Ann Arbor Public School Administrative Staff and in New York: Hunter College High School, and Stuyvesant High School;
- *publishing partners*—American Mathematical Society, Association of Research Libraries (ARL), Cambridge University Press, Elsevier

Science, Encyclopedia Britannica Educational Corporation, Groliers, McGraw-Hill and University Microfilm (UMI); and

- *corporate and organizational partners*—Apple Computer, Bellcore, Eastman-Kodak, Hewlett-Packard, IBM, Sybase and UMI International.

The testbed database contains earth and space science multi-media information. The major focus of the research includes:

- *scalability*—using agents to help unify diverse collections and locate information; and
- *education*—investigating "inquiry-based education," a new teaching style that heavily utilizes the Internet and digital libraries.

Specific research topics include investigating: the use of agents to unify collections and services (scalability); conspectus search (i.e. finding appropriate collection); new educational methods (i.e. "inquiry-based education"); and several different approaches to copyright and intellectual property protection on the Internet. A critical component of the project is the testing and evaluation of the prototype system by a wide variety of users, including those from on-campus, local high schools and public libraries. Michigan has several other ongoing digital library projects (JSTOR, TULIP, and "The Making of America") that it is integrating with this effort [49].

The Michigan project is investigating the coordination of three levels or types of agents for intelligent information identification and retrieval: *user interface agents* which interact with the user to define and determine the scope of the inquiry, *mediation agents* which coordinate the requests of many user interface agents and their interactions with the collections (using a conspectus approach to searching) and *collection agents* which handle the specific details of searching various media types within a given collection. Another focus of the project is investigating the pricing of electronic access to knowledge using agents that negotiate access fees based on supply and demand or market information (see the Internet AuctionBot information on the project's web site). More detail can be obtained from the project's homepage (which also contains a list of publications, a demonstration, and links to other related sites) and [5] and [22].

6.1.1.6 Stanford University: InfoBus (walrus.stanford.edu/diglib/) The focus of this $3.6 million project is on "Interoperation Mechanisms among Heterogeneous Services." It is also called the InfoBus project. The principal investigator is Hector Garcia-Molina of the Computer Science Department.

Participating organizations include: Association for Computing Machinery (ACM), Bell Communications Research (Bellcore), Enterprise Integration

Technologies (EIT), Hewlett-Packard Labs (HP), Hitachi Corp., Hughes Research Laboratory, Interconnect Technologies Corporation (ITC), Interval Research Corporation, Knight-Ridder Information (Dialog Information Services—DIS), MIT Press, NASA/Ames Research Center (NASA/ARC) including the Advanced Interaction Media Group and the library, Naval Command, Control and Ocean Surveillance Center (NCCOSC), O'Reilly and Associates, WAIS Inc. and Xerox PARC.

The primary focus of this project is *interoperability*. It has a small testbed database consisting of computer science literature. Primary research topics include:

- investigating different database infrastructures for digital library support primarily focusing on CORBA (the distributed-object standard developed by the Object Management Group);
- investigating different network infrastructures for digital library support; and
- investigating a World Wide Web annotation service which allows "permanent" annotation of Web pages without modification to the original.

The problems involved in the integration of a variety of heterogeneous information sources (ranging from personal sources to public library or repository sources to private or proprietary domain specific sources) at a higher level than the current transport-oriented protocols is one of the primary research thrusts of the project. The goal is to develop enabling technologies for a single, integrated "virtual" library available over the Internet that provides uniform access to a large number of networked information sources and collections, thus requiring new forms of information sharing and communication models, client information interfaces and information finding services. Another research focus is on the legal and economic issues of a networked environment. Inter-Pay, for example, is an architecture designed to provide interoperability for fee-based services, while the Stanford SCAM and COPS (Copywrite Protection Service) projects are involved with the protection of intellectual property in electronic or digital forms [70].

The heart of the Stanford project is the "InfoBus" protocol, which provides a uniform way to access a variety of services and information sources through "proxies" acting as interpreters between it and each collection's native protocol and thereby allowing access to multiple, distributed information sources. The InfoBus protocol is supported by a variety of user-level applications, which provide powerful ways to find information by using cutting-edge user interfaces for direct manipulation or through agent technology. Examples of Stanford projects involved in information location include GlOSS (Glossary-of-Servers Server) which helps identify the most relevant information sources for a particular query and SIFT (Stanford Information Filtering Tool). More detail is available from the project's homepage mentioned above, which contains some excellent links to other digital library information [35, 70].

6.1.2 Other Major US Digital Library Research Activities

There are several other major US digital library research activities. Most of them involve large-scale content creation and user testbeds.

6.1.2.1 Library of Congress (lcweb.loc.gov/) The most ambitious national level project in the United States is the Library of Congress National Digital Library Program under the direction of James Billington. This project's goal is to digitize over 5 million multimedia items from the extensive Library of Congress collection with particular focus on historical information. There are currently 32 historical collections available for keyword searching and browsing (by titles, topics, and library division by collection type—i.e. photos and prints, documents, motion pictures, maps and sound recordings). This project includes: the *American Memory project* which will become one of the testbeds for the DLI-2 initiative; *THOMAS*, which provides full text access to bills under consideration in the US House of Representatives and Senate; and *Library of Congress exhibitions* which are now indefinitely available on the Internet. Many other databases and information resources are available on-line through the Library of Congress Internet gateway. Some of these resources include: Library of Congress catalogs (which contain over 110 million items); access to catalogs at over 200 other libraries both within and outside the US (through the Z39.50 Gateway); Library of Congress Thesauri; the Vietnam Era Prisoner of War/Missing in Action and Task Force Russia Databases; Science Tracer Bullets (SCTB) Online (bibliographic guides); GLIN—Global Legal Information Network; and US Copyright Office Records.

Other digital library projects at the Library of Congress include:

- *Standards*: The Library of Congress is the maintenance agency for several key standards used in the information community, including US Machine-Readable Cataloging (MARC) formats, the Z39.50 information retrieval protocol, the Encoded Archival Description (EAD) Document Type Definition (DTD) for Standard Generalized Markup Language (SGML), and the International Standard Serial Number (ISSN).

- *The Library of Congress/Ameritech National Digital Library Competition*: With a gift of $2 million from the Ameritech Foundation, the Library of Congress is sponsoring a competition to enable public, research, and academic libraries, museums, historical societies and archival institutions (except federal institutions) to create digital collections of primary resource material. These digital collections will complement and become part of the collections of the National Digital Library Program at the Library of Congress. The Library of Congress/Ameritech National Digital Library Competition will run for three consecutive years beginning in 1996, with the expectation that eight to ten awards of up to $75,000 each will be made annually.
- *The Digital Library Federation*: Fifteen of the nation's largest research libraries and archives agreed in 1995 to cooperate on defining what must be done to bring together digitized materials that document the building and dynamics of United States heritage and cultures and will be made accessible to people everywhere. Members include: Columbia University, the Commission on Preservation and Access, Cornell University, Emory University, Harvard University, The Library of Congress, National Archives and Records Administration, The New York Public Library, Pennsylvania State University, Princeton University, Stanford University, University of California at Berkeley, University of Michigan, University of Southern California, University of Tennessee, and Yale University.

  Some of the projects include: Computer Sciences Technical Reports (CSTR) Project at Berkeley, Emory University Virtual Library, Information Infrastructure Project (Harvard University), Making of America Project (Cornell University and the University of Michigan), Networked Computer Science Technical Reports Library (NCSTRL at Cornell University), New York State Museum Bulletins Project (Columbia University), Open Book Project (Yale University) and Stanford University Computer Science Electronic and Technical Reports Library.
- *The Global Information Society*: The Library of Congress is a member of the Global Information Society and participates in the G-7 "Electronic Libraries" Project (a part of the Group 7 Bibliotheca Universalis project) with France, Japan, Germany, Canada, Italy, and the United Kingdom. The Library officially represents the United States on issues of international importance to the future of libraries.

6.1.2.2 National Aeronautics and Space Administration—NASA (www.nasa.gov/) In addition to supporting the Digital Library

Initiatives, NASA provides the ability to search on-line its extensive collections, with particular emphasis on its earth science scientific data collection (predominantly satellite data and images). Some of the available NASA information includes: AVHRR 1 km Data Browser (Italy) Live Access to Climate Data (NOAA), Crustal Dynamics Data Information System (NASA), The Earth Science Data and Information System Project, EOS Data Information System Distributed Active Archive Centers, The Global Change Data Center, Global Hydrology Resource Center (GHRC), Guide to NASA Online Resources, National Climate Data Center (NOAA), NASA Documents on-line, NASA News Archives, NASA's Earth Observing System Earth Images and Data, NASA Facts on-line, NASA News Archives, Pathfinder Datasets, SeaWiFS Project—Ocean Color Data, The Total Ozone Mapping Spectrometer (TOMS) Project and Data, and many other NASA publications.

NASA's Digital Library Technology (DLT) project supports the development of new technologies to facilitate public access to NASA data via computer networks, placing greatest emphasis on scalable technologies that develop tools, applications, and software and hardware systems. The DLT project and the Public Use of Remote Sensing Data (RSD) project, are two related elements of the Information Infrastructure Technology and Applications (IITA) component of NASA's High Performance Computing and Communications Program (HPCC).

6.1.2.3 FedStats (www.FedStats.gov) The collection available in this multiple agency project includes statistical information (reports, charts, tables, etc.) from more than 70 different US governmental agencies. This information has two major data types (textual and numerical) and is Web accessible in a variety of formats from spreadsheets to relational databases. The major challenges include designing flexible presentation and intelligent user interfaces that can assist users in transforming raw statistics into useful information or knowledge.

6.1.2.4 IBM (www.software.ibm.com/is/dig-lib/) The IBM digital library is one of the first commercial digital library efforts in the United States. It consists of an array of products and services aimed at helping customers transform information into digital multimedia forms that can be shared with others via networks and includes tools to manage, present and protect that information. The primary focus of the IBM digital library is on *image retrieval* (e.g. Scriptorium) with particular emphasis on high quality image presentation, searching capabilities for multimedia information, and copyright and intellectual property management capabilities [34]. The IBM digital library integrates a wide array of scalable information storage, management, search, retrieval, and distribution technologies, many of which were already available, into a single architecture. Excellent examples of the IBM digital library are the El Archive General de Indies, the works of Andrew Wyeth, and the Vatican Library project, all of which can be found in [34].

6.1.2.5 The California Digital Library Project (www.lpai.ucop.edu) The California Digital Library or CDL is the tenth system-level University of California library and is administered from the Office of the President. The University of California system (nine campuses with nine-system level libraries) owns a collection of over 29 million volumes in its 100-plus individual libraries. With the exception of the Library of Congress, this collection is the largest on the American continent. Each of the nine University of California campuses has its own library system, but the CDL (formerly known as the Division of Library Automation) is the first system-wide library with a primary emphasis on accessibility and digitization. The CDL is responsible for leading several automation and digitization projects for the University of California, and for expanding both the accessible collections (to include more public and private libraries and museums and libraries and museums from California, around the United States, and eventually the world) and extending accessibility (to include the general public). Resource sharing is obviously one important focus of CDL, but another major research area includes efforts to reduce staggering increases in costs over the last five years by exploring electronic alternatives to scholarly publishing. The first target shared collection for digitization in the CDL is a Science, Technology and Industry Collection which will include access to approximately 1000 science and technology journals when it is complete. Plans are already underway to begin processing a Humanities and Social Sciences shared digitized collection.

CDL's responsibilities include the maintenance and continuing development and expansion of MELVYL, which can search on-line the entire system's catalog, the California Academic Libraries list of Serials and approximately 120 other databases, 25 of which are commercial products such as MEDLINE. The system can be accessed either by telnet or through the Internet. The original objective of the MELVYL system was to unite and provide on-line access to the library collections of the entire University of California system. The objective has since shifted to providing state-wide and wider access to a broad range of information sources for research and education. Several other public and private libraries in California (including museums and museum collections) like the Stanford University Library are now accessible, and the goal is to continue to expand accessibility. Richard E. Lucier was named to head the CDL in October of 1997. Funding for CDL was budgeted for $1.5 million in the fiscal year 1997–98 and approximately $4 million in the fiscal year 1998–99 [64].

6.1.2.6 The DARPA D-Lib Program (www.dlib.org) D-Lib is an excellent on-line digital library reference source coordinated by CNRI and sponsored by the Defense Advanced Research Projects Agency (DARPA). The Web site provides access to the D-Lib Magazine (an on-line monthly compilation of contributed stories, commentary, and briefings); Ready Reference (an exceptional clearinghouse of pointers to other sites on the Web of interest to researchers and users of digital libraries); Technology Spotlight (a collection of demos and interesting reports and papers related to digital library research); and other information.

6.1.2.7 JSTOR (www.jstor.org/) JSTOR is a not-for-profit organization initially established with funding from The Andrew W. Mellon Foundation. It is dedicated to helping the scholarly community take advantage of advances in information technologies, in particular electronic publishing capabilities, in an effort to ease the costs of scholarly journal storage and access. The project goals are to convert back issues of paper journals into electronic formats to save space (and the capital costs associated with that space) while simultaneously improving access to journal contents and offering a solution to preservation problems associated with storing paper-based information.

A pilot project at five library test sites provided electronic access to the back files of ten journals in two core fields, economics and history. Every issue of the ten participating journals published prior to 1990—approximately 750,000 total pages—has now been converted from paper into an electronic database that resides at the University of Michigan and is mirrored at Princeton University. Using technology developed at Michigan, high-resolution (600 dpi) bit-mapped images of each page are linked to a text file generated with optical character recognition (OCR) software which, along with newly constructed Table-of-Contents indexes, permits complete search and retrieval of the journal material.

Currently the collection contains the "back files" of over 50 journals and it permits access from more than 300 universities and colleges around the world [94]. Future plans include extending the collection to more than 100 journals (by the end of 1999).

6.1.2.8 Making of America—Cornell University and the University of Michigan (moa.cit.cornell.edu/ and www.umdl.umich.edu/moa/) The goal of the Making of America (MOA) Project is to create and make accessible (over the Internet) a distributed digital library of materials on the history of the United States. The Cornell University and University of Michigan libraries have cooperated in the initial phase of MOA, which has been funded by the Andrew W. Mellon Foundation and the Charles E. Culpeper Foundation. The MOA project is designed to select complementary journals and monographs on 19th-century US history from the two universities, create digital images of the material in a manner that ensures full capture of all significant information, and provide equitable access to the combined digital collection from both campuses.

6.1.2.9 Project Open Book—Yale University (www.library.yale.edu/preservation/pobweb.htm) Yale University Library's Project Open Book is a research and development program that is exploring the feasibility and costs of large-scale conversion of preserved material from microfilm to digital imagery. The project has three overall goals:

(1) *Conversion*—create a 10,000 volume digital image library from converted microfilm and evaluate issues of workflow, quality, and cost;
(2) *Intellectual access*—enhance intellectual access through the creation of document structure and page number indexes enabling scholars to go directly to a particular page or document structure element, such as a table of contents; and
(3) *Distributed physical access*—enhance physical access to the digital library by providing distributed access over the Yale campus network and eventually over the Internet.

6.1.2.10 The Red Sage Project (www.ckm.ucsf.edu/projects/RedSage/) The Red Sage project was a four-year (1992–96) experimental cooperative effort between the University of California at San Francisco, AT&T Bell Laboratories (Right-Pages), and Springer-Verlag to provide on-line access to 70 Springer-Verlag, American Medical Association (AMA) and other health-science related journals on molecular biology and radiology. The goal of the project was to allow researchers, clinicians, and students to search, read and print the full-page text, including graphics and photographs. Red Sage investigated technical, legal, business, and human factors issues related to network delivery of scientific journal literature.

6.1.2.11 The TULIP Project—Elsevier (www1.elsevier.nl/homepage/about/resproj/tulip.shtml) TULIP [7] was an Elsevier Science Publishers initiative (1991–95) created to explore the issues involved in electronic scholarly journal electronic distribution. Ten universities (Carnegie Mellon University—CMU, Cornell University, Georgia

Institute of Technology, Massachusetts Institute of Technology—MIT, Princeton University, University of California (all campuses), University of Michigan, University of Tennessee, University of Washington, and Virginia Polytechnic Institute and State University) were involved in the project, which also included several official *observing* institutions (California State University, Harvard University, Pennsylvania State University, Stanford University and the University of Southern California). Approximately 60 Materials Science journal titles comprised the testbed collection. The project was especially interested in exploring economic models for electronic access, investigating the technical feasibility of networked distribution of journals, and investigating usage patterns for electronic access. One of the most significant contributions made by this project is its extensive user evaluation effort [37]. It is now a commercial product.

6.2 Digital Library Activities in Other Countries

Most digital library activities in other countries focus on collections, particularly on improving access to collections of historical, cultural or artistic significance. Many of the social issues surrounding digital libraries, especially legal issues related to copyright laws and protection of intellectual property, are also of primary importance in these activities. As most of these efforts are at a national level, the primary lead organization is the national library of the country or government agencies.

6.2.1 Canada

There are several Canadian digital library initiatives. The most recent, the Canadian Initiative on Digital Libraries (CIDL) was initiated in 1998 and comprises an alliance of Canadian libraries. CIDL's goal is to promote, coordinate, and facilitate the development of Canadian digital collections and services, emphasizing interoperability and long-term access to Canadian digital library resources. (see www.nlc-bnc.ca/cidl/).

The University of Waterloo developed text-searching software based on the Patricia tree data structure that is now being commercially sold by OpenText [49]. Waterloo is also the site of a major community network prototype and study called Canada's Technology Triangle (CTT) Community Network. This project is community-based and provides local maps, information about local businesses and local historical information available on the Internet, focusing on communication and community-building infrastructures [21] (see CTTnet.uwaterloo.ca).

The National Library of Canada, much like the US Library of Congress, provides on-line access to some of its collection through its Web homepage.

It is also involved in an extensive historical information digitization project and in the Canadian Electronic Publications Pilot Project (see www.nlc-bnc.ca/ehome.htm).

6.2.2 United Kingdom

The Follett report begun in 1993 initiated digital library activities in the United Kingdom, resulting in the E-Lib project (UK Electronic Libraries Program), which funded approximately 60 programs for a total of £20 million starting in 1995. Some of the E-Lib projects include cataloging archives, providing UK-wide e-mail services, and providing access to Web multimedia information (a distinct challenge due to the congestion on the Internet between the US and the UK). The E-Lib project focuses more on providing electronic resources and services for UK higher education than on purely scientific research. One of the most significant E-Lib achievements has been persuading publishers to treat the *entire* UK academic community as a single site for the purposes of licensing information [59]. (see www.ukoln.ac.uk/services/elib/).

The British Library began a digital library program in 1993. Similarly to other national library programs, its emphasis is on collections, preservation and improved public access. The British Library is also heavily involved in the establishment of standards and legal guidelines for copyright law and intellectual property protection with respect to digital libraries [49]. Several fascinating key projects include: the Patent Express Jukebox (CD-ROM jukebox of over one million patent records), the Electronic Beowulf Project (a huge database of digital images of the Beowulf manuscript, related manuscripts and printed texts), the Electronic Photo Viewing System (10,000 images of historical significance that are hyperlinked to descriptive text), the Network OPAC (providing ability to search on-line the British Library's 6 million bibliographic records, which is connected to all other UK university and research institutes via the UK Joint Academic Network—JANET), and conversion of microfilm to digital form [75]. Several literary and artistic treasures are available on-line through the British Library Web site, including: the Lindisfarne Gospels, the Diamond Sutra, the Magna Carta, the Sforza Hours, a Leonardo da Vinci Notebook, and the Tyndale New Testament (see www.bl.uk).

6.2.3 France

The Bibliothèque Nationale de France, the national library of France, has led the digitization effort in France. It plans to make available through its Web site 110,000 digitized books (mostly in image format), 300,000 pictures and 3,000 recordings. The most interesting component of this digitized collection is a group of 1000 14th-century manuscripts [42, 49] (see www.bnf.fr/ or the English version at www.bnf.fr/bnfgb.htm).

Frantext/ARTFL (American and French Research on the Treasury of the French Language) allows full-text access to a digitized database of classic French literature (approximately 3500 items) [42, 49]. Frantext is managed by the Institut National de la Langue Française in Paris (see ciril.fr/mastina/FRANTEXT). ARTFL is managed through the University of Chicago (see humanities.uchicago.edu/ARTFL.html).

France is a co-leader with Japan on the Group 7 project (the Bibliotheca Universalis project) whose goal is to provide global access to digitized cultural, historical and scientific multi-media information from all over the world (see www.culture.fr/culture/bibliuni/engbu1.htm).

6.2.4 Germany

Multimedia Electronic Documents (MeDoc) is a digital library project sponsored by the German Informatics Society that involves several publishers, universities and research institutions. The goal of this project is to "stimulate the use of electronic media in academic education and in scientific research" [27]. The collection currently focuses on computer science literature. The system contains a mechanism that allows the retrieval of individual book chapters or journal paper components (Fulltext Storage System—FSS). (see medoc.informatik.tu-muenchen.de).

6.2.5 Gabriel

Gabriel is a gateway server that provides access to Europe's National Libraries. Information about the server is available in three languages (German, English and French). The objective is to try to bring the European community belonging to the Conference of European National Libraries (CENL) closer together by providing a single entry point through the Internet to all of their libraries, including their collections and services. Thirty-eight national European libraries (including the Vatican City library) are connected through Gabriel (see linnea.helsinki.fi/gabriel/index.html).

6.2.6 Japan

National Diet Library is sponsoring five Electronic Library Projects. They include: Pilot Electronic Library Projects, Electronic Library of Children's Books, Asian Information Supply System, G7 (Group 7) Electronic Library Project or Bibliotheca Universalis, and Full Text Database of the Minutes of the Diet. As with most national libraries, the focus is on the collections themselves, from an acquisition and preservation perspective with the goal of improving public access to national cultural and historical information. This project is extremely ambitious in that it is attempting to digitize a wide variety of multimedia information, and information in very diverse formats (e.g. scrolls, woodblock prints, microfilms) (see www.ndl.go.jp/index-e.html).

As one of the three main projects under the Center for Information Infrastructure (CII), the Electronic Library Pilot Project offers experimental electronic access to a vast collection of diverse books and other resources existing in libraries all over Japan. One of the main goals of this project is to test the latest advances in data-processing and network technologies in preparation for future digital libraries. This is a joint project of the Information-technology Promotion Agency and the National Diet Library (see www.cii.ipa.go.jp/el/index_e.html).

The MHTML (Multilingual HTML) project is a fascinating multilingual project that allows multilingual browsing of a digital library of Japanese folk tales on the Internet. This collection can be accessed via a gateway, and can process Japanese, Korean, Chinese, Thai, and several European languages, including French and English [55] (see mhtml.ulis.ac.jp/).

The Science Information Network (SINET) is one of several projects sponsored by the Japanese National Center for Science Information Systems. This system connects universities and research institutions all over Japan and in addition is interconnected with research networks in the United States, the United Kingdom, and Thailand. NACSIS-IR is an information retrieval service that provides access to 57 different databases, allowing a user to simultaneously browse, search and retrieve information from multiple databases. NACSIS also has an electronic library service (NACSIS-ELS) which provides access to an integrated bibliographic database from several distributed sources and an electronic document delivery service of Japanese scientific journals over the Internet (see www.nacsis.ac.jp/nacsis.f-index.html).

6.2.7 Korea

Korea has a digital library pilot program developed under the supervision and guidance of the Ministry of Information and Communication. The project testbed is accessible via the Internet and provides full-text, catalogue and abstract searching and browsing. Five major libraries are involved: (1) The National Library of Korea (includes a database of old rare books available on-line), (2) The National Assembly Library (provides legislative documents similar to the Library of Congress's THOMAS project), (3) The

Science Library in KAIST, (4) The Korea Research and Development Information Center in KAIST and (5) the Korean Research Foundation). The project uses SGML and Z39.50 standards to provide interoperability between the sites and users. All information is accessible in Korean and some of it is accessible in English. Much of the challenge involved in the Korean Digital Library project has to do with the problems associated with Korean text. The language is character-based and has many stylistic differences even within a given document collection, which can negatively impact information retrieval and exchange [62] (see www.dlibrary.or.kr).

6.2.8 Singapore

Singapore is aspiring to be the "Intelligent Island" with leading edge information management as one of its top national priorities. To this end, it has electronically linked all of the libraries in the country. As its primary focus is on linking to information via the Internet, most of the electronically linked libraries function essentially as gateways [49]. The emphasis in this country is on using digital information for business and for educational purposes. One of the other major projects in Singapore is the CLiB project. This project uses the Z39.50 protocol and provides multilingual (primarily Chinese and English) searching of several heterogeneous bibliographic databases. It focuses primarily on language support issues, especially those related to representation of Chinese characters [47] (see www.lib.gov.sg/nlb.html).

6.2.9 China, Hong Kong, and Taiwan

Most of the digital library projects in this region are centered around Chinese information retrieval and digitization of local and cultural content.

Digital library research efforts in China have focused on collections, and in particular on the preservation of national literary treasures. Other Chinese digital library projects involve tackling the challenge of electronically recognizing Chinese characters, Chinese cross-lingual retrieval and developing standards for digital libraries such as the Chinese MARC record standard [49] and the Net Compass Project of Tsinghua university.

In Hong Kong, special-purpose digital libraries have been created. One of these is the Financial Digital Library project at the university of Hong Kong that serves the unique need of the Hong Kong stock market and users. Similar to the research effort in China, the Chinese University of Hong Kong has continued to research NLP-based intelligent Chinese information retrieval for digital libraries.

The Academia Sinica in Taiwan probably has the best research teams in the area of Chinese information retrieval and voice recognition. Funded by the Taiwan government for about two decades, these teams have significantly advanced Chinese input, segmentation, indexing, and analysis. Many of their techniques have been adopted in digital libraries or Internet servers of Chinese content. In addition, several projects in Taiwan have focused on cultural content digitization, including the Digital Museum project of the National Taiwan University and the art collection digitization of Palace Museum in Taipei by IBM.

The First Asia Digital Library Workshop was held in Hong Kong in August 1998. With its focus on Asia digital library research projects, the workshop attracted more than 120 participants from nine Asia/Pan-Pacific countries and has served as the catalyst for Asia digital library collaborations. Several countries have expressed strong interest in sponsoring a Second Asian Digital Library Workshop. An Asia Digital Library Consortium is also under development to help foster long-term collaboration and projects in digital library related topics in Asia (see www.ssrc.hku.hk/sym/98/adl.html).

6.2.10 *Australia*

Australia has several major digitization efforts that are centered at the National Gallery of Australia and the National Library of Australia and focus on historical and cultural material. One effort is digitizing aboriginal language recordings and other multimedia material relating to Australian Aborigines. Another effort, the MetaWeb Project, is focusing on developing tools for metadata creation and maintenance [49]. The National Library of Australia is involved in a project called PADI (Preserving Access to Digital Information) that is investigating issues and solutions to problems of preservation and access to digital information in the future (see www.nla.gov.au/).

6.2.11 *New Zealand*

The New Zealand digital library system is a federally funded research project at the University of Waikato which contains several demonstration collections (primarily in computer science, e.g. computer science technical reports, literary works, Internet FAQs, the Computists' Communique magazine) and makes them available over the Web through full-text interfaces. The digital library has expanded beyond its original computer science-oriented collection to include literary, music, and image collections and an Arabic collection accessible in Arabic. There are interfaces to access the collections in five different natural languages (English, French, German, Arabic, and Maori) [101]. The goal of this research program is to explore the potential of Internet-based digital libraries by developing systems that automatically impose structure on anarchic, uncatalogued, distributed repositories of information, thereby providing information consumers with effective tools to locate what they need and to peruse it conveniently and comfortably [102]. The project includes a collaborative effort with the German MeDoc project [27] and the *Journal of Biological Chemistry* to explore novel browsing techniques. Access is based on full-text retrieval as opposed to metadata. The most interesting part of the collection is a musical collection of 9400 international folk tunes stored in musical notation. The melody index can be used to retrieve songs using sung, hummed or played musical input [60] (see www.nzdl.org/).

7. Digital Library Research Issues in Semantic Interoperability

7.1 Digital Library Grand Challenge: Semantic Interoperability

The Information Infrastructure Technology and Applications (IITA) Working Group, the highest level of the US National Information Infrastructure (NII) technical committee, held an invited workshop in May 1995 to define a research agenda for digital libraries.

The shared vision is an entire net of distributed repositories, where objects of any type can be searched within and across different indexed collections [88]. In the short term, technologies must be developed to transparently search across these repositories, handling any variations in protocols and formats (i.e. addressing structural interoperability [70]). In the long term, technologies must be developed to handle the variations in content and meanings (knowledge) transparently as well. These requirements are steps along the way toward matching the concepts requested by users with objects indexed in collections [87].

The ultimate goal, as described in the IITA report [53], is the Grand Challenge of digital libraries:

> deep semantic interoperability—the ability of a user to access, consistently and coherently, similar (though autonomously defined and managed) classes of digital objects and services, distributed across heterogeneous repositories, with federating or mediating software compensating for site-by-site variations ... Achieving this will require breakthroughs in description as well as retrieval, object interchange and object retrieval protocols. Issues here include the definition and use of metadata and its capture or computation from objects (both textual and multimedia), the use of computed descriptions of objects, federation and integration of heterogeneous repositories with

disparate semantics, clustering and automatic hierarchical organization of information, and algorithms for automatic rating, ranking, and evaluation of information quality, genre, and other properties.

Attention to semantic interoperability has prompted several of the NSF/DARPA/NASA funded large-scale digital library initiative (DLI) projects to explore various statistical, and pattern recognition techniques, e.g. concept spaces and category maps in the Illinois project [90, 15], textile and word sense disambiguation in the Berkeley project [100], voice recognition in the CMU project [96], and image segmentation and clustering in the UCSB project [56]. "Definition and use of metadata" and "clustering and automatic hierarchical organization of information," which require significant future research, are the key components needed to build classification systems for digital libraries automatically.

7.2 Research Towards Semantic Interoperability in Digital Libraries

Library classification systems and subject-specific thesauri such as the Library of Congress classification, Dewey classification, or the NLM's Unified Medical Language Systems (UMLS) are significant human efforts to have trained librarians, who are versed in classification scheme and domain knowledge, label knowledge consistently [39, 11]. Library classification systems and thesauri often capture nouns or noun phrases and represent only limited relationships (e.g. broader terms, narrower term, etc.). The representations are often coarse, but precise. The goal of supporting indexing and searching is practical. Significant human efforts are needed to create and maintain large-scale classification systems.

Artificial intelligence representations such as semantic networks, expert systems, or ontologies represent another approach to capturing knowledge, e.g. Lenat's CYC common sense knowledge base [45, 46, 36]. Such representations are often richer and more fine-grained and the goal of capturing human intelligence is ambitious and difficult. Due to the granularity required, knowledge creation is slow and painstaking. Only experimental prototypes in small, limited domains have been created. Their usefulness in large-scale digital library applications remains suspect.

The traditional approach to creating classification systems and knowledge sources in library science and classical AI is often considered top-down since knowledge representations and formats are pre-defined by human experts or trained librarians and the process of generating knowledge is structured and well-defined. A complementary bottom-up approach to knowledge creation has been suggested by researchers in machine learning, statistical analysis, and neural networks.

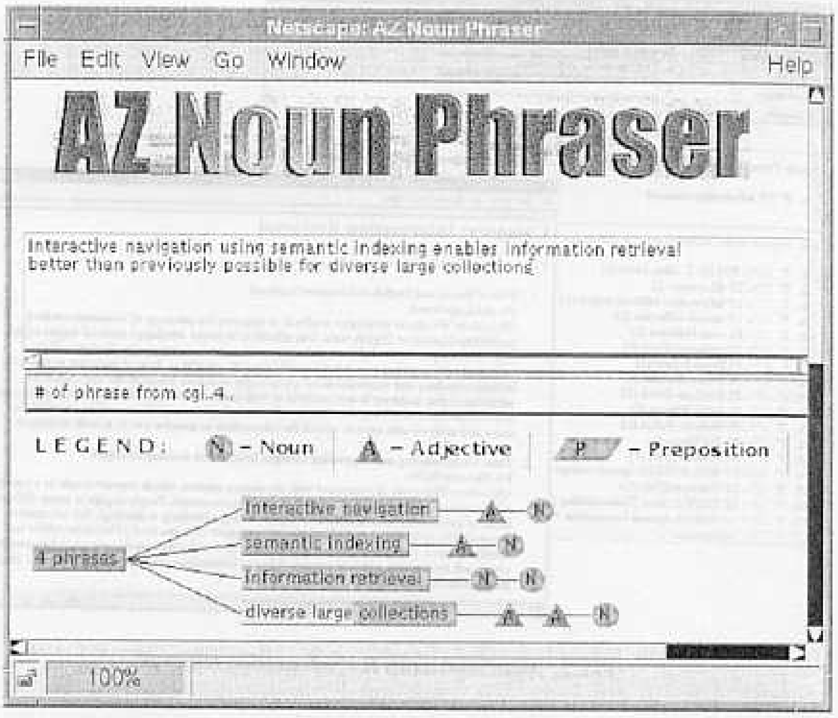
Based on actual databases or collections, researchers develop programs which systematically segment and index documents in various databases (text, image, and video) and identify patterns within such databases. Analyzing databases which contain structured and numeric data (e.g. credit card usage, frequent flyer program) is often referred to as data mining or knowledge discovery [73, 54]. Generating knowledge algorithmically from multimedia databases (especially text, e.g. customer complaint e-mail, machinery repair reports, brainstorming outputs) is considered the core of knowledge management [67].

Among the semantic indexing and analysis techniques that are considered scalable and domain independent, the following classes of algorithms and methods have been examined and subjected to experimentation in various digital library applications. We also provide examples from our own research for illustration purposes.

7.2.1 Object Recognition, Segmentation, and Indexing

The most fundamental techniques in information retrieval involve identifying key features in objects. For example, automatic indexing and natural language processing (e.g. noun phrase extraction or object type tagging) are frequently used to automatically extract meaningful keywords or phrases from texts [80]. Texture, color, or shape-based indexing and segmentation techniques are often used to identify images [56]. For audio and video applications, voice recognition, speech recognition, and scene segmentation techniques can be used to identify meaningful descriptors in audio or video streams [96].

As part of the Illinois DLI project, we have developed a noun phrasing technique for textual document indexing [38]. Noun phrase indexing aims to identify concepts (grammatically correct noun phrases) from a collection for term indexing. It begins with a text tokenization process to separate punctuation and symbols. It follows by part-of-speech-tagging (POST) using variations of the Brill tagger and 30-plus grammatic noun phrasing rules. Figure 1 shows an example of tagged noun phrases for a simple sentence. (The system is referred to as AZ Noun Phraser.) For example, "interactive navigation" is a noun phrase that consists of an adjective (A) and a noun (N). In [38], we have shown that the noun phrasing technique produces more accurate indices for digital libraries (than inverted word indexing or N-gram indexing) and helps in concept-based retrieval. By using such a scalable natural language processing technique, digital libraries will be able to efficiently (automatically) and precisely index its own collections.



Interactive navigation using semantic indexing enables information retrieval better than previously possible for diverse large collections

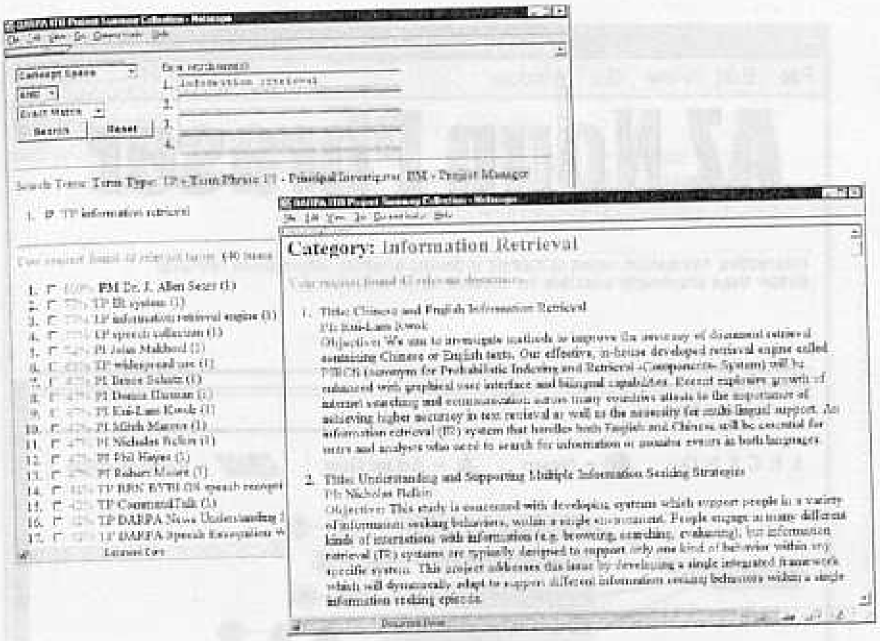
# of phrase from cgi..4..

FIG. 1. Tagged noun phrases.

7.2.2 Semantic Analysis

Several classes of techniques have been used for semantic analysis of texts or multimedia objects. Symbolic machine learning (e.g. ID3, version space), graph-based clustering and classification (e.g. Ward's hierarchical clustering), statistics-based multivariate analyses (e.g. latent semantic indexing, multi-dimensional scaling, regressions), artificial neural network-based computing (e.g. backpropagation networks, Kohonen self-organizing maps), and evolution-based programming (e.g. genetic algorithms) are among the popular techniques [10]. In this information age, we believe these techniques will serve as good alternatives for processing, analyzing, and summarizing large amounts of diverse and rapidly changing multimedia information.

The *concept space* technique, developed at the Illinois DLI project is an example of semantic, statistical analysis of large-scale digital library collections [15, 12]. Concept space, like an automatic thesaurus, attempts to generate weighted, contextual concept (term) association in a collection to assist in concept-based associative retrieval. It adopts several heuristic term


DARPA ITO Project Summary Collection - Netscape
File Edit View Go Communicator Help
Concept Space
AND
Exact Match
Search
Reset

Term(s) to search:
1. information retrieval
2.
3.
4.

Search Terms: Term Type: TP - Term Phrase PI - Principal Investigator PM - Project Manager

1. ☑ TP information retrieval

[illegible] (40 items)

1. ☐ 100% PM Dr. J. Allen Sears (1)
2. ☐ [illegible] TP IR system (1)
3. ☐ [illegible] TP information retrieval engine (1)
4. ☐ [illegible] TP speech collection (1)
5. ☐ [illegible] PI John Makhoul (1)
6. ☐ [illegible] TP widespread use (1)
7. ☐ [illegible] PI Bruce Schatz (1)
8. ☐ [illegible] PI Donna Harman (1)
9. ☐ [illegible] PI Kui-Lam Kwok (1)
10. ☐ [illegible] PI Mitch Marcus (1)
11. ☐ [illegible] PI Nicholas Belkin (1)
12. ☐ [illegible] PI Phil Hayes (1)
13. ☐ [illegible] PI Robert Moore (1)
14. ☐ [illegible] TP BBN BYBLOS speech recogni
15. ☐ [illegible] TP CommandTalk (1)
16. ☐ [illegible] TP DARPA News Understanding
17. ☐ [illegible] TP DARPA Speech Recognition

Document Done
DARPA ITO Project Summary Collection - Netscape
File Edit View Go Communicator Help

Category: Information Retrieval

Your request found [illegible] relevant documents

1. Title: Chinese and English Information Retrieval
   PI: Kui-Lam Kwok
   Objective: We aim to investigate methods to improve the accuracy of document retrieval containing Chinese or English texts. Our effective, in-house developed retrieval engine called PIRCS (acronym for Probabilistic Indexing and Retrieval -Components- System) will be enhanced with graphical user interface and bilingual capabilities. Recent explosive growth of internet searching and communication across many countries attests to the importance of achieving higher accuracy in text retrieval as well as the necessity for multi-lingual support. An information retrieval (IR) system that handles both English and Chinese will be essential for users and analysts who need to search for information or monitor events in both languages.

2. Title: Understanding and Supporting Multiple Information Seeking Strategies
   PI: Nicholas Belkin
   Objective: This study is concerned with developing systems which support people in a variety of information seeking behaviors, within a single environment. People engage in many different kinds of interactions with information (e.g. browsing, searching, evaluating), but information retrieval (IR) systems are typically designed to support only one kind of behavior within any specific system. This project addresses this issue by developing a single integrated framework which will dynamically adapt to support different information seeking behaviors within a single information seeking episode.

FIG. 2. Associated terms for "information retrieval".

weighting rules and a weighted co-occurrence analysis algorithm. Figure 2 shows the associated terms for "information retrieval" in a sample collection of project reports of the DARPA/ITO Program—TP (Term Phrase) such as "IR system," "information retrieval engine," "speech collection," etc. Such concept spaces have been computed for collections of the scale of 100,000 web pages [12], one million abstracts across engineering [15], and 10 million abstracts across medicine [3].

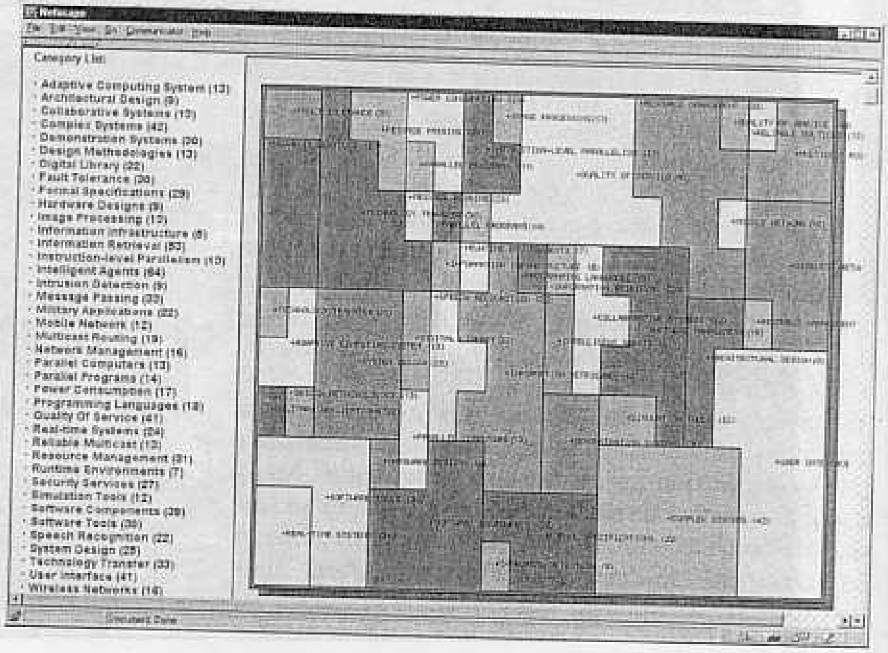
7.2.3 Knowledge Representations

The results from a semantic analysis process could be presented in one of many knowledge representations, including classification systems, semantic networks, decision rules, or predicate logic. Many researchers have attempted to integrate such results with existing human-created knowledge structures such as ontologies, subject headings, or thesauri [58]. Spreading activation based inferencing methods are often used to traverse various large-scale knowledge structures [14].

In [12] and [16], we reported a neural network-based textual categorization technique for digital library content classification. A category map is the result of performing neural network-based clustering (self-organizing) of similar documents and automatic category labeling. Documents that are similar to each other (in noun phrase terms) are grouped together in a neighborhood on a two-dimensional display. As shown in the colored jigsaw-puzzle display in Fig. 3, each colored region represents a unique topic that contains similar documents. Topics that are more important often occupy larger regions. By clicking on each region, a searcher can browse documents grouped in that region. An alphabetical list that is a summary of the 2D result is also displayed on the left-hand side of Fig. 3, e.g. Adaptive Computing System (13 documents), Architectural Design (nine documents), etc. Our current research has demonstrated the computational scalability and clustering accuracy and novelty of this technique [69, 12].

7.2.4 *Human–Computer Interactions (HCI) and Information Visualization*

One of the major trends in almost all digital library applications is the focus on user-friendly, graphical, and seamless HCI. The Web-based browsers for texts, images, and videos have raised user expectations of the rendering and



Category List:

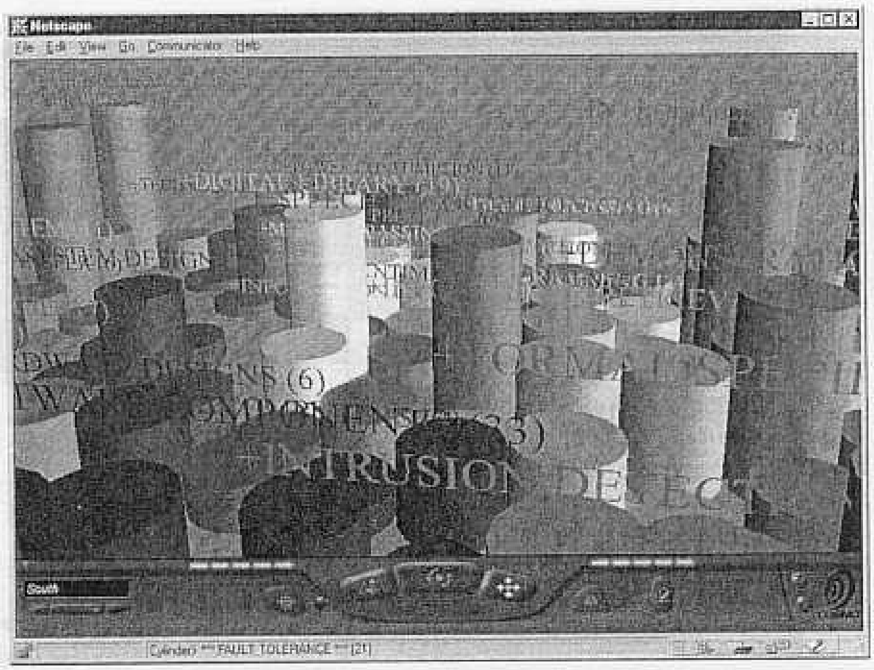
- Adaptive Computing System (13)
- Architectural Design (9)
- Collaborative Systems (13)
- Complex Systems (42)
- Demonstration Systems (30)
- Design Methodologies (13)
- Digital Library (22)
- Fault Tolerance (30)
- Formal Specifications (29)
- Hardware Designs (9)
- Image Processing (13)
- Information Infrastructure (8)
- Information Retrieval (53)
- Instruction-level Parallelism (13)
- Intelligent Agents (64)
- Intrusion Detection (9)
- Message Passing (33)
- Military Applications (22)
- Mobile Network (12)
- Multicast Routing (19)
- Network Management (16)
- Parallel Computers (13)
- Parallel Programs (14)
- Power Consumption (17)
- Programming Languages (18)
- Quality Of Service (41)
- Real-time Systems (24)
- Reliable Multicast (13)
- Resource Management (31)
- Runtime Environments (7)
- Security Services (27)
- Simulation Tools (12)
- Software Components (28)
- Software Tools (30)
- Speech Recognition (22)
- System Design (25)
- Technology Transfer (33)
- User Interface (41)
- Wireless Networks (16)

FIG. 3. Category map.

manipulation of information. Recent advances in development languages and platforms such as Java, OpenGL, and VRML and the availability of advanced graphical workstations at affordable prices have also made information visualization a promising area for research [24]. Several of the digital library research teams including Arizona/Illinois, Xerox PARC, Berkeley, and Stanford, are pushing the boundary of visualization techniques for dynamic displays of large-scale information collections.

In addition to the graphical, colored 2D display shown in Fig. 3, the same clustering results from the category map can also be displayed in a 3D helicopter fly-through landscape as shown in Fig. 4, where cylinder height represents the number of documents in each region. Similar documents are grouped in a same-colored region. Using a VRML plug-in (COSMO player), a searcher is then able to "fly" through the information landscape and explore interesting topics and documents. Clicking on a cylinder will display the underlying clustered documents.

Our initial lab experiments have confirmed the novelty and graphical appeal of such a 3D visualization metaphor, especially for the younger web generation. In particular, we found most users of digital libraries may exhibit different cognitive styles and tend to prefer one visualization



Cylinder "" FAULT TOLERANCE "" (21)

FIG. 4. VRML interface for category map.

metaphor over another (e.g. comparing 1D hierarchy, 2D jigsaw puzzle display, and 3D fly-through display). A more personalized interface may be needed for transmitting digital library content to different users. More HCI research in the context of digital libraries is critically needed in light of the richness of digital library content and media format and the diversity in user styles and needs.

8. Conclusions and the Future

The World Wide Web has made access to the Internet part of the structure of everyday life. At the same time, over the past few years, the primary interface to the Web has evolved from browsing to searching. Millions of people all over the world do Web searching every day, but the commercial technology of searching large collections has remained largely unchanged from its roots in US government-sponsored research projects of the 1960s. This public awareness of the Net as a critical infrastructure in the 1990s has caused a new revolution in the technologies for information retrieval in digital libraries, driven by the hardware revolution in network-based personal computers.

Digital libraries are a form of information technology in which social impact matters as much as technological advancement. It is hard to evaluate new technology in the absence of real users and large collections. The best way to develop effective new technology is in multi-year large-scale research projects which develop real-world electronic testbeds for actual users and which aim at developing new, comprehensive, and user-friendly technologies for digital libraries. Typically, these testbed projects also examine the broad social, economic, legal, ethical, and cross-cultural contexts and impacts of digital library research.

The May 1996 issue of the *IEEE Computer* magazine, edited by Bruce Schatz and Hsinchun Chen, was a special issue on digital libraries. The issue focused specially on the NSF/DARPA/NASA Digital Libraries Initiative (DLI). The six major projects supported by the DLI each was represented by a survey paper at the then halfway point in the initiative [88]. The February 1999 *IEEE Computer* issue, edited by the same guest editors, focuses on practical outcomes from research projects—major research testbeds and fundamental technology research that shows what the large-scale future infrastructure might become [89]. Digital libraries have become far more important nationally and internationally in 1999 than in 1996. Partially, this is due to the exponential growth of information in the Web, which the Web searchers are rapidly failing to handle successfully. This is a special case of the increasing dependence of modern society on information technology and the increasing failures of fundamental infrastructure stemming from lack of fundamental new technology.

The recently released PITAC report (President's Information Technology Advisory Committee Interim Report, August 1998) makes this point clearly. This was a report, in direct response to the President of the United States, by the leaders of information technology from industry and university. They concluded:

> Vigorous information technology (IT) research and development is essential for achieving America's 21st century aspirations. The technical advances that led to today's information tools, such as electronic computers and the Internet, began with Federal government support of research in partnership with industry and universities. All of these innovations depended on patient investment in fundamental and applied research.
>
> We have had a spectacular return on that Federal government research investment. Businesses that produce computers, semiconductors, software, and communications equipment have accounted for one-third the total growth in US production since 1992, creating millions of high paying new jobs. As we approach the 21st century, the opportunities for innovation in IT are larger than they have ever been—and more important. We have an essential national interest in ensuring a continued flow of good new ideas in IT.
>
> After careful review of the Federal programs, however, this Committee has concluded that Federal support for research in information technology is dangerously inadequate. Research programs intended to maintain the flow of new ideas in IT are turning away large numbers of excellent proposals. In addition, current support is taking a short-term focus, looking for immediate returns, rather than investigating high-risk long-term technologies. Significant new research on computers and communication systems serve our needs while protecting us from catastrophic failures of the complex systems that now underpin our transportation, defense, business, finance and healthcare infrastructure.
>
> The current Federal program is inadequate to start necessary new centers and research programs. Computers on university campuses and other civilian research facilities are falling rapidly behind the state of the art. The end result is that critical problems are going unsolved and we are endangering the flow of ideas that have fueled the information economy.
>
> To address these problems, the Federal budget for the year 2000 should include a commitment to sustained growth in IT research, along with a new management system designed to foster innovative research. The Federal IT research program must include vigorous support for fundamental and applied research and must ensure that the US research community is equipped with state-of-the art facilities.

The follow-on to the DLI is another NSF-led initiative, which builds on the successes of DLI and presages the even bigger efforts that will follow on the PITAC report. The NSF/DARPA/NLM/LC/NASA/NEH Digital Libraries Initiative Phase 2 (DLI-2) has recently made the initial awards for multi-year projects. DLI-2 will support a broader range of activities than the first DLI, including small projects and humanities topics. There will be an even stronger emphasis on working testbeds with real users and real collections.

The DLI and DLI-2 Program Director, Dr Stephen Griffin explains in [89]:

> The Digital Libraries Initiative-Phase 2 (DLI-2) supported by NSF, DARPA, NLM, LoC, NEH, NASA and other agency partners will address a refined technology research agenda and look to support new areas and dimensions in the digital libraries information lifecycle including content creation, access, use and usability, preservation and archiving. DLI-2 will look to create domain applications and operational infrastructure, and understand their use and usability in various organizational, economic, social, international contexts—in short, digital libraries as human-centered systems. DLI-2 involvement will extend far beyond computing and communications specialty communities and engage scholars, practitioners and learners in not only science and engineering but also arts and humanities. DLI-2 recognizes that knowledge access is inherently international and will actively promote activities and processes that bridge political and language boundaries, including funding through a new program in International Digital Libraries Collaborative. [see DLI-2 at www.dli2.nsf.gov]

The technologies of digital libraries will dominate the Net of the 21st century [87]. There will be a billion repositories distributed over the world, where each small community maintains a collection of their own knowledge. Semantic indexes will be available for each repository, using scalable semantics to generate search and navigation aids for the specialized terminology of each community. Concept switching across semantic indexes will enable members of one community to easily search the specialized terminology of another [13].

The Internet will have transformed into the *Interspace*, where users navigate abstract spaces to perform correlation across sources. Information analysis will become a routine operation in the Net, performed on a daily basis worldwide. Such functionality will first be used by specialty professionals then by ordinary people, just as occurred with text search on Internet. The information infrastructure will become the essential part of the structure of everyday life and digital libraries will become the essential part of information infrastructure in the 21st century.

ACKNOWLEDGMENTS

This project was funded primarily by:

- NSF/CISE "Concept-based Categorization and Search on Internet: A Machine Learning, Parallel Computing Approach," NSF IRI9525790, 1995–98.

- NSF/ARPA/NASA Illinois Digital Library Initiative project, "Building the Interspace: Digital Library Infrastructure for a University Engineering Community," NSF IRI9411318, 1994–98.
- National Center for Supercomputing Applications (NCSA), "Parallel Semantic Analysis for Spatially-oriented Multimedia GIS Data," High-performance Computing Resources Grants (Peer Review Board), on Convex Exemplar and SGI Origin2000, June 1996–June 1998 (IRI960001N).
- Department of Defense, Advanced Research Projects Agency (DARPA), "The Interspace Prototype: An Analysis Environment Based on Scalable Semantics," June 1997–May 2000 (N66001-97-C-8535).

We would also like to thank Dr Stephen Griffin of NSF, Dr Ron Larsen of DARPA, Dr Alexa McCray of NLM, Dr Larry Smarr of NCSA, and Dr Bruce Schatz of CANIS at UIUC.

References

[1] Adam, N., and Yesha, Y. (1996). Strategic directions in electronic commerce and digital libraries: Towards a digital agora. *ACM Computing Surveys*, **28**(4), 818–835.

[2] Agrawal, R., Imielinski, T., and Swami, A. (1993). Database mining: A performance perspective. *IEEE Transactions on Knowledge and Data Engineering*, **5**(6), 914–925.

[3] Alper, J. (1998). Taming MEDLINE with concept spaces. *Science*, 281, 1785, 18 September.

[4] Atkins, D. E. (1997). Report of the Santa Fe planning workshop on distributed knowledge work environments: Digital libraries. Supported by a Grant from the National Science Foundation (NSF-IRI-9712586) to the University of Michigan School of Information, March 9–11.

[5] Atkins, D. E., Birmingham, W. P., Durfee, E. H., Glover, E. J., Mullen, T., Rundensteiner, E. A., Soloway, E., Vidal, J. M., Wallace, R., and Wellman, M. P., (1996). Toward inquiry-based education through interacting software agents. *IEEE Computer*, **29**(5), 69–75.

[6] Atkinson, R. (1996). Library functions, scholarly communication, and the foundation of the digital library: Laying claim to the control zone. *Library Quarterly*, **66**(3), July.

[7] Borghuis, M., Brinckman, H., Fischer, A., Hunter, K., Loo, E., Mors, R., Mostert, P., and Zijlstra, J. (1996). *TULIP final report*. Technical report, Elsevier Science Publishers.

[8] Burke, C. (1992). The other memex: The tangled career of Vannevar Bush's information machine, the Rapid Selector. *Journal of the American Society for Information Science*, **43**(10), 648–657.

[9] Bush, V. (1945). As we may think. *Atlantic Monthly*, July, 101–108.

[10] Chen, H. (1995). Machine learning for information retrieval: neural networks, symbolic learning, and genetic algorithms. *Journal of the American Society for Information Science*, **46**(3), 194–216.

[11] Chen, H., and Dhar, V. (1990). User misconceptions of online information retrieval systems. *International Journal of Man-Machine Studies*, **32**(6), 673–692.

[12] Chen, H., Houston, A. L., Sewell, R. R., and Schatz, B. R. (1998). Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science*, **49**(7), 582–603.

[13] Chen, H., Martinez, J., Ng, D. T., and Schatz, B. R. (1997). A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the Worm Community System. *Journal of the American Society for Information Science*, **48**(1), 17–31.

[14] Chen, H., and Ng, D. T. (1995). An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): symbolic branch-and-bound vs. connectionist Hopfield net activation. *Journal of the American Society for Information Science*, **46**(5), 348–369.

[15] Chen, H., Schatz, B. R., Ng, T. D., Martinez, J. P., Kirchhoff, A. J., and Lin, C. (1996). A parallel computing approach to creating engineering concept spaces for semantic retrieval: The Illinois Digital Library Initiative Project. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18**(8), 771–782.

[16] Chen, H., Schuffels, C., and Orwig, R. (1996). Internet categorization and search: a machine learning approach. *Journal of Visual Communications and Image Representation*, **7**(1), 88–102.

[17] Chen, H., Smith, T. R., Larsgaard, M. L., Hill, L. L., and Ramsey, M. (1997). A geographic knowledge representation system (GKRS) for multimedia geospatial retrieval and analysis. *International Journal of Digital Library*, **1**(2), 132–152.

[18] Chen, S., Chien, Y., and Griffin, S. (1996). Agency perspectives on the digital library initiative. *IEEE Computer*, **29**(5), 23–24.

[19] Ching, N., Jones, V., and Winslett, M. (1996). Authorization in the digital library: Secure access to services across enterprise boundaries. *Proceedings of the Third Forum on Research and Technology Advances in Digital Libraries—ADL '96 Forum*, IEEE, pp. 110–119. 13–15 May, Washington, DC.

[20] Choy, D. M., Dwork, C., Lotspiech, J. B., Anderson, L. C., Boyer, S. K, Dievendorff, R., Griffin, T. D., Hoenig, B. A., Jackson, M. K., Kaka, W., McCrossin, J. M., Miller, A. M., Morris, R. J. T., and Pass, N. J. (1996). A digital library system for periodicals distribution. *Proceedings of the Third Forum on Research and Technology Advances in Digital Libraries—ADL '96 Forum*, IEEE, pp. 95–103. 13–15 May Washington, DC.

[21] Cowan, D. D., Mayfield, C. I., Tompa, F. W., and Gasparini, W. (1998). New role for community networks. *Communications of the ACM*, **41**(4), 61–63.

[22] Crum, L. (1995). University of Michigan digital library project. *Communications of the ACM*, **38**(4), 63–64.

[23] Davis, W. (1988). Documentation unfinished. *Bulletin of the American Society for Information Science*, **14**(5), 50–53.

[24] DeFanti, T., and Brown, M. (1990). Visualization: expanding scientific and engineering research opportunities. (1990). In *Visualization in Scientific Computing*, (eds G. M. Nielson, B. D. Shriver, and L. J. Rosenblum), IEEE Computer Society Press, NY.

[25] Dowler, L. (1997). Gateways to knowledge: A new direction for the Harvard College Library. In L. Dowler, editor, *Gateways to Knowledge: The role of academic libraries in teaching, learning and research*. MIT Press, Cambridge, MA.

[26] Drabenstott, K. M. (1993). Analytical review of the library of the future. Council on Library Resources, 1400 16th Street, N.W., Suite 510, Washington, DC 20036–2217. Research assistance by Celeste M. Burman.

[27] Endres, A., and Fuhr, N. (1998). Students access books and journals through MeDoc. *Communications of the ACM*, **41**(4), 76–77.

[28] Feldman, S. (1997). Advances in digital libraries '97. *Information Today*, **14**(7), 12–13.

[29] Ferguson, C. D., and Bunge, C. A. (1997). The shape of services to come: Values-based reference service for the largely digital library. *College and Research Libraries*, **58**(3), 252–265.

[30] Fox, E. A., and Marchionini, G. (1998). Toward a worldwide digital library. *Communications of the ACM*, **41**(4), 28–32.

[31] Frye, B. E. (1997). Universities in transition: Implications for libraries. In *Gateways to Knowledge: The Role of Academic Libraries in Teaching, Learning and Research* (ed. Lawrence Dowler). MIT Press, Cambridge, MA.

[32] Garrett, J. R., and Lyons, P. A. (1993). Toward an electronic copyright management system. *Journal of the American Society for Information Science*, **44**(8), 468–473.

[33] Ginsparg, P. (1997). First steps toward electronic research communication. In *Gateways to Knowledge: The Role of Academic Libraries in Teaching, Learning and Research* (ed. Lawrence Dowler). MIT Press, Cambridge, MA.

[34] Gladney, H. M., Mintzer, F., Schiattarella, F., Bescós, J., and Treu, M. (1998). Digital access to antiquities. *Communications of the ACM*, **41**(4), 49–57.

[35] The Stanford Digital Libraries Group. (1995). The Stanford digital library project. *Communications of the ACM*, **38**(4), 59–60.

[36] Guarino, N. (1995). The role of formal ontology in the information technology. *International Journal of Human-Computer Studies*, **43**(5/6), 623–624.

[37] Hitchcock, S., Carr, L., Harris, S., Hey, J. M. N., and Hall, W. (1997). Citation linking: Improving access to online journals. *Proceedings of the 2nd ACM International Conference on Digital Libraries* (Robert B. Allen and Edie Rasmussen eds.), pp. 115–122, ACM, New York, NY.

[38] Houston, A. L., Chen, H., Schatz, B. R., Sewell, R. R., Tolle, K. M., Doszkocs, T. E., Hubbard, S. M., and Ng, D. T. (1999). Exploring the use of concept space, category map techniques, and natural language parsers to improve medical information retrieval. *Decision Support Systems*, (forthcoming).

[39] Humphreys, B. L., and Lindberg, D. A. (1989). Building the unified medical language system. *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care*, pages 475–480, Washington, DC: IEEE Computer Society Press, 5–8 November.

[40] Huser, C., Reichenberger, K., Rostek, L., and Streitz, N. (1995). Knowledge-based editing and visualization for hypermedia encyclopedias. *Communications of the ACM*, **38**(4), 49–51.

[41] Kalakota, R., and Whinston, A. B. (1996). *Frontiers of Electronic Commerce*. Addison-Wesley Publishing Company, Reading, MA.

[42] Kessler, J. (1996). *Internet Digital Libraries: The International Dimension*. Artech House, Inc., Boston, MA.

[43] Kluiters, C. P. (1997). Delivering "building blocks" for digital libraries: First experiences with Elsevier electronic subscriptions and digital libraries in Europe. *Library Acquisitions: Practice and Theory*, **21**(3), 273–279.

[44] Knoblock, C., Koller, D., Shoham, Y., Wellman, M. P., Durfee, E. H., Birmingham, W. P., and Carbonell, J. (1996). The role of AI in digital libraries. *IEEE Expert*, **11**(3), 8–13.

[45] Lenat, D. B., Borning, A., McDonald, D., Taylor, C., and Weyer, S. (1983). Knoesphere: Building expert systems with encyclopedic knowledge. *Proceedings of the International Joint Conference of Artificial Intelligence*.

[46] Lenat, D. B., Guha, R., Pittman, K., Pratt, D., and Shepherd, M. (1990). CYC: Toward programs with common sense. *Communications of the ACM*, **33**(8), 30–49.

[47] Leong, M. K., Cao, L., and Lu, Y. (1998). Distributed Chinese bibliographic searching. *Communications of the ACM*, **41**(4), 66–67.

[48] Lesk, M. (1996). Digital libraries meet electronic commerce: On-screen intellectual property. *Proceedings of the Third Forum on Research and Technology Advances in Digital Libraries—ADL '96 Forum*, pages 58–64. IEEE. Washington, DC, 13–15 May.

[49] Lesk, M. (1997). *Practical Digital Libraries*. Morgan Kauffmann, Los Altos, CA.

[50] Levy, D. M., and Marshall, C. C. (1995). Going digital: A look at assumptions underlying digital libraries. *Communications of the ACM*, **38**(5), 77–84.

[51] Licklider, J. (1965). *Libraries of the Future*. The MIT Press, Cambridge, MA.

[52] Lyman, P. (1996). What is a digital library? Technology, intellectual property and the public interest. *Daedalus*, **125**(4), 1–34.

[53] Lynch, C., and Garcia-Molina, H. (1995). Interoperability, scaling and the digital libraries research agenda. (1995). *A Report on the May 18–19, 1995 Information Infrastructure Technology and Applications (IITA) Digital Libraries Workshop*, 22 August.

[54] Lynch, K. J., and Chen, H. (1992). Knowledge discovery from historical data: an algorithmic approach. *Proceedings of the 25th Annual Hawaii International Conference on System Sciences (HICSS-25), Decision Support and Knowledge Based Systems Track*, pp. 70–79, Kaui, HI, 7–10 January.

[55] Maeda, A., Dartois, M., Fujita, T., Sakaguichi, T., Sugimoto, S., and Tabata, K. (1998). Viewing multilingual documents on your local web browser. *Communications of the ACM*, **41**(4), 64–65.

[56] Manjunath, B. S., and Ma, W. Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18**(8), 837–841.

[57] Mansell, R. (1996). Designing electronic commerce. In R. Mansell and R. Silverstone, editors, *Communication by Design: The Politics of Information and Communication Technologies*. Oxford University Press, New York, NY.

[58] McCray, A. T., and Hole, W. T. (1990). The scope and structure of the first version of the UMLS semantic network. *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care*, pp. 126–130, Los Alamitos, CA: Institute of Electrical and Electronics Engineers, 4–7 November.

[59] McKnight, C. (1998). Many projects that depend on collaboration. *Communications of the ACM*, **41**(4), 86–87.

[60] McNab, R. J., Smith, L. A., Whitten, I. H., Henderson, C. L., and Cunningham, S. J. (1996). Toward the digital music library: Tune retrieval from acoustic input. *Proceedings of ACM Digital Libraries*, pages 11–18.

[61] Moen, W. E. (1998). Accessing distributed cultural heritage information. *Communications of the ACM*, **41**(4), 45–48.

[62] Myaeng, S. (1998). R&D for a nationwide general-purpose system. *Communications of the ACM*, **41**(4), 83–85.

[63] Nadis, S. (1996). Computation cracks semantic barrier between databases. *Science*, **272**, 1419, 7 June.

[64] Notice. (1997). University of California, November 1997. A publication of the Academic Senate, University of California, Volume 22, No. 2.

[65] Nyce, J. M. and Kahn, P. (1989). Innovation, pragmaticism, and technological continuity: Vannevar Bush's Memex. *Journal of the American Society for Information Science*, **40**(3), 214–220.

[66] Odlyzko, A. M. (1996). Tragic loss or good riddance? The impending demise of traditional scholarly journals. In R. P. Peek and G. B. Newby, (eds.), *Scholarly Publishing: The Electronic Frontier*. The MIT Press, Cambridge, MA.

[67] O'Leary, D. E. (1998). Enterprise knowledge management. *IEEE Computer*, **31**(3), 54–61.

[68] Olsen, J. (1997). The gateway: Point of entry to the electronic library. In L. Dowler (ed.), *Gateways to Knowledge: The Role of Academic Libraries in Teaching, Learning and Research*. MIT Press, Cambridge, MA.

[69] Orwig, R., Chen, H., and Nunamaker, J. F. (1997). A graphical, self-organizing approach to classifying electronic meeting output. *Journal of the American Society for Information Science*, **48**(2), 157–170.

[70] Paepcke, A., Cousins, S. B., Garcia-Molina, H., Hasson, S. W., Ketcxhpel, S. P., Roscheisen, M., and Winograd, T. (1996). Using distributed objects for digital library interoperability. *IEEE COMPUTER*, **29**(5), 61–69.

[71] Peek, R. P., and Newby, G. B. (1996). *Scholarly Publishing: The Electronic Frontier*. The MIT Press, Cambridge, MA.

[72] Peterson, I. (1996). Fashioning a world brain. *Bulletin of the American Society for Information Science*, **22**(5), 10–11.

[73] Piatetsky-Shapiro, G. (1989). Workshop on knowledge discovery in real databases. *Proceedings of the International Joint Conference of Artificial Intelligence*.

[74] Prentice, A. E. (1997). Copyright, WIPO and user interests: Achieving balance among the shareholders. *The Journal of Academic Librarianship*, **23**(4), 309–312.

[75] Purday, J. (1995). The British Library's initiatives for access projects. *Communications of the ACM*, **38**(4), 65–66.

[76] Rao, R., Pedersen, J. O., Hearst, M. A., Mackinlay, J. D., Card, S. K., Masinter, L., Halvorsen, P., and Robertson, G. G. (1995). Richer interaction in the digital library. *Communications of the ACM*, **38**(5), 29–39.

[77] Reddy, R. (1996). The universal library: Intelligent agents and information on demand. In N. R. Adam, B. K. Bhargava, M. Halem, and Y. Yesha (eds.), *Digital Libraries Research and Technology Advances*, Springer-Verlag.

[78] Rockwell, R. C. (1997). The concept of the gateway library: A view from the periphery. In L. Dowler, editor, *Gateways to Knowledge: The role of academic libraries in teaching, learning and research*. MIT Press, Cambridge, MA.

[79] Rush, J. E. (1996). Foreword. In R. P. Peek and G. B. Newby, (eds.), *Scholarly Publishing: The Electronic Frontier*. The MIT Press, Cambridge, MA.

[80] Salton, G. (1989). *Automatic Text Processing*. Addison-Wesley Publishing Company, Inc., Reading, MA.

[81] Salton, G., and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY.

[82] Samuelson, P. (1995). Copyright and digital libraries. *Communications of the ACM*, **38**(5), 15–21 and 110.

[83] Samuelson, P. (1998). Encoding the law into digital libraries. *Communications of the ACM*, **41**(4), 13–18.

[84] Saracevic, T., and Kantor, P. B. (1997). Studying the value of library and information services. Part I. Establishing a theoretical framework. *Journal of the American Society for Information Science*, **48**(6), 527–542.

[85] Saracevic, T., and Kantor, P. B. (1997). Studying the value of library and information services. Part II. Methodology and taxonomy. *Journal of the American Society for Information Science*, **48**(6), 543–563.

[86] Schatz, B. R. (1995). Building the interspace: The Illinois digital library project. *Communications of the ACM*, **38**(4), 62–63.

[87] Schatz, B. R. (1997). Information retrieval in digital libraries: Bringing search to the net. *Science*, **275**, 327–334.

[88] Schatz, B. R., and Chen, H. (1996). Building large-scale digital libraries. *IEEE Computer*, **29**(5), 22–27.

[89] Schatz, B. R., and Chen, H. (1999). Digital libraries: technological advancements and social impacts. *IEEE Computer*, **31**(2), 45–50.

[90] Schatz., B. R., Mischo, B., Cole, T., Hardin, J., Bishop, A., and Chen, H. (1996). Federating repositories of scientific literature. *IEEE Computer*, **29**(5), 28–36.

[91] Smith, T. R. (1996). A digital library for geographically referenced materials. *IEEE Computer*, **29**(5), 54–60.

[92] Smith, T. R., and Frew, J. (1995). Alexandria digital library. *Communications of the ACM*, **38**(4), 61–62.

[93] Tennant, R. (1997). The grand challenges. *Library Journal*, **122**(20), 31–33.

[94] Thomas, S. W., Alexander, K., and Guthrie, K. (1999). Technology choices for the JSTOR online archive. *IEEE Computer*, **31**(2), 60–65, February.

[95] Unsworth, J. (1997). Some effects of advanced technology on research in the humanities. In *Gateways to Knowledge: The role of academic libraries in teaching, learning and research* (ed. Lawrence Dowler). MIT Press, Cambridge, MA.

[96] Wactlar, H. D., Kanade, T., Smith, M. A., and Stevens, S. M. (1996). Intelligent access to digital video: Informedia project. *IEEE Computer*, **29**(5), 46–53.

[97] Weaver, W. (1955). Machine translation of languages. In *Translation* (W. N. Locke and A. D. Booth eds.). John Wiley, New York, NY, pp. 15–27. (Reprint of 1949 memo.)

[98] Wells, H. G. (1971). *World Brain*. Books for Libraries Press, Freeport, NY.

[99] Wiederhold, G. (1995). Digital libraries, value and productivity. *Communications of the ACM*, **38**(5), 85–96.

[100] Wilensky, R. (1996). Toward work-centered digital information services. *IEEE Computer*, **29**(5), 37–45.

[101] Witten, I. H., McNab, R., Apperley, M., Bainbridge, D., Cunningham, S. J., and Jones, S. (1999). Managing multiple collections, multiple languages, and multiple media in a distributed digital library. *IEEE Computer*, **31**(2), 74–80, February.

[102] Witten, I. H., Nevill-Manning, C., McNab, R., and Cunningham, S. J. (1998). A public library based on full-text retrieval. *Communications of the ACM*, **41**(4), 71–75.

[103] Wulf, W. A. (1989). The national collaboratory—a white paper. In Towards a National Collaboratory, the unpublished report of a workshop held at Rockefeller University, Joshua Lederberg and Keith Uncapher, co-chairs, March 17–18.