Building Large-Scale Digital Libraries

Building Large-Scale Digital Libraries

Bruce Schatz
University of Illinois

Hsinchun Chen
University of Arizona

In this era of the Internet and the World Wide Web, the long-time topic of digital libraries has suddenly become white hot. As the Internet expands, particularly the WWW, more people are recognizing the need to search indexed collections. As *Science* news articles on the US Digital Library Initiative (DLI) have put it, the Internet is like a library without a card catalog,[1] and the hottest new services are the Web search engines.[2]

The term "digital" is actually somewhat of a misnomer. Digital libraries basically store materials in electronic format and manipulate large collections of those materials effectively. So research into digital libraries is really research into network information systems. The key technological issues are how to search and display desired selections from and across large collections. While practical digital libraries must focus on issues of access costs and digitization technology, digital library research concentrates on how to develop the necessary infrastructure to effectively mass-manipulate the information on the Net.

Digital library research projects thus have a common theme of bringing search to the Net. This is why the US government made digital libraries the flagship research effort for the National Information Infrastructure (NII), which seeks to bring the highways of knowledge to every American. As a result, the four-year, multiagency DLI was funded with roughly $1 million per year for each project (see the "Agency perspectives" sidebar). Six projects (chosen from 73 proposals) are involved in the DLI, which is sponsored by the National Science Foundation, Advanced Research Projects Agency, and the National Aeronautics and Space Administration. This issue of *Computer* includes project reports from these six university sites:

- Carnegie Mellon University,
- University of California at Berkeley,
- University of California at Santa Barbara,
- University of Illinois at Urbana-Champaign,
- University of Michigan, and
- Stanford University.

PROJECT RANGE

The DLI projects are a good measure of the current research into large-scale digital libraries. They span a wide range of the major topics necessary to develop the NII. These projects, however, are not the only ongoing efforts, nor do they concentrate much on the practical issues of actually building large-scale digital libraries. (See the April 1995 special issue of *Communications of the ACM* on digital libraries for short descriptions of many major practical projects. This issue of *Computer* is intended to be deep rather than broad and will focus on infrastructure. For a discussion

Articles on the Web

As befitting the content of this theme issue, complete versions of the articles can be accessed through *Computer*'s home page at

http://www.computer.org/pubs/computer/computer.htm

0018-9162/96/$5.00 © 1996 IEEE

of the challenges involved in using AI to build digital libraries, see the June 1996 issue of *IEEE Expert*.)

The DLI projects address future technological problems. The overall initiative is about half over, so these articles describe issues and plans more than final results. The authors have tried to concentrate on concrete results and to cover the range of problems addressed within each project. Project details can be accessed through the home page of the DLI National Synchronization Effort (http://www.grainger. uiuc.edu/dli/national.htm).

VARIOUS APPROACHES

The DLI projects use many contrasting approaches. For example, the Illinois and Berkeley projects both plan full systems with many users, with the Illinois project focusing on manually structured text documents and the Berkeley project on automatically recognized image documents. These projects use complementary approaches, receiving materials in electronic format directly from publishers to take advantage of the embedded SGML structure, and receiving them in paper format in large volumes and automatically transforming the articles into digital form.

The Carnegie Mellon and Santa Barbara projects plan to provide the ability to manipulate new media that were previously impossible to index and search. Carnegie Mellon is investigating segmenting and indexing video, using automatic speech recognition and knowledge about program structure. Santa Barbara is indexing maps, using automatic image processing and knowledge about region metadata.

Finally, the Stanford and Michigan projects plan to investigate the intermediaries (gateways) necessary to perform operations on large-scale digital libraries. These projects are trying to find the representations needed, on one hand, to interoperate between the formats for different search services and, on the other hand, to identify the appropriate sources to be searched for a given query.

All projects are building testbeds with large collections to address their corresponding fundamental research questions into building large-scale digital libraries.

RESEARCH AGENDA

The Information Infrastructure Technology and Applications (IITA) Working Group, the highest level NII technical committee, held an invited workshop in May 1995 to define the 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 indexed collections. In the short term, technology must be developed to transparently search across these repositories, handling the variations in protocols and formats. In the long term, technology must be developed to transparently handle the variations in content and

Agency perspectives on the Digital Library Initiative

National Science Foundation
Su-Shing Chen, Yi-Tzuu Chien, and Stephen M. Griffin

The digital library concept can be traced back to the famous May 1945 *Atlantic Monthly* article by Vannevar Bush ("As We May Think") and the 1988 white paper on national collaboratories by William Wulf of NSF. The DLI provides the necessary impetus to realize the vision articulated in these statements. After the government has stimulated basic research in various enabling technologies and built several digital library testbeds, IT companies, traditional libraries, publishers, organizations, and users will join forces to develop knowledge repositories, which will play an essential role for society in the twenty-first century.

The general position of NSF, ARPA, and NASA on this Initiative was described in the NSF Research on Digital Libraries Announcement, NSF 93-141, as follows:

> To explore the full benefits of such digital libraries, the problem for research and development is not merely how to connect everyone and everything together in the network. Rather, it is to achieve an economically feasible capability to digitize massive corpora of extant and new information from heterogeneous and distributed sources; then store, search, process and retrieve information from them in a user friendly way. Among other things, this will require both fundamental research and the development of "intelligent" software. It is the purpose of this announcement to support such research and development by combining the complementary strengths of the participating agencies in basic research, advanced development and applications, and academic/industry linkage.

The projects, as appropriate to the research focus, were required to include the active participation of the following groups:

1. client groups (like specific research communities or other users of the information encompassed in the proposal);
2. commercial enterprises involved in the commercialization of a digital library system (like publishers, software houses, stock exchanges, equipment manufacturers, and communications companies);
3. archival establishments, either private or governmental (like libraries, data repositories, clearing houses, government or private information or data services); and
4. relevant computer and other science and engineering research groups (like academic departments, supercomputer centers, and industrial laboratories).

The projects were required to be university-led consortia with strong partnering relationships.

In the NSF/ARPA/NASA announcement, the following research areas were presented as illustrative examples:

1. systems for capturing information;
2. categorizing and organizing electronic information;
3. developing advanced software for searching, filtering, and summarizing large volumes of data, imagery, and information;

meaning as well. These are steps along the way toward matching the concepts requested by the users to the objects indexed in the collections.

The ultimate goal, as described in the IITA report,[3] 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.

At a stylistic level, the primary goal of networked digital libraries is to consider the entire Net as a single virtual collection from which users can extract relevant parts. Handling issues of scale, such as the number of objects and repositories and the range of types and subjects, is thus very important. This is why the DLI projects focus on large-scale testbeds. Indexing and searching technology must be not only effective for user needs, it must also scale up to large collections across large networks.

As the IITA report says,

> We don't know how to approach scaling as a research question, other than to build upon experience with the Internet. However, attention to scaling as a research theme is essential and may help in further clarifying infrastructure needs and priorities, as well as informing work in all areas of the research agenda outlined above. . . . There was consensus on the need to enable large-scale deployment projects (in terms of size of user community, number of objects, and number of repositories) and subsequently to fund study the effectiveness and use of such systems. It is clear that limited deployment of prototype systems will not suffice if we are to fully understand the research questions involved in digital libraries.

INDEXING AND FEDERATING

The process of using a digital library thus involves searching across distributed repositories. A repository is just an indexed collection of objects. Distributed searching

4. visualization and other interactive technology for quickly browsing large volumes of imagery;
5. networking protocols and standards needed to ensure the ability of the digital network to accommodate the high volume, bandwidth, and switching requirements of a digital library;
6. simplifying the utilization of networked information resources distributed around the nation and around the world,
7. individual and group behavioral, social, and economic issues in digital libraries.

The research areas outlined in the announcement require extensive research-and-development efforts that will take several years to accomplish. The six funded projects have targeted their research at a significant portion of these areas, with some intersecting, complementary, and orthogonal coverages.

Advanced Research Projects Agency

Barry Leiner and Robert Neches

Digital Libraries, from the ARPA perspective, are not merely about on-line access to documents, even multimedia documents. Rather, ARPA is most interested in research on digital libraries that produces enabling technology and infrastructure for wide-area information management, exchange, and collaboration. Capabilities for storing, finding, transmitting, viewing, and manipulating complex information—as well as for controlling access and accounting for it—have applications that go well beyond traditional uses of libraries. Investment in digital library research therefore contributes to the development of a broad base of information management tools in a wide range of domains. These include command and control, intelligence analysis, crisis management, and intelligent logistics. For this reason, ARPA is particularly concerned with the development of system architectures and protocols that facilitate interoperability between components being developed in the course of the research.

National Aeronautics and Space Administration

Nand Lal

NASA repositories of earth and space science data contain information interesting to non-NASA scientists and to the general public, especially for uses in environmental monitoring, formal education, and lifelong learning. This information has traditionally been unavailable outside the NASA science community due to lack of infrastructure and digital library technologies. These technologies are needed to implement NII services that enable content producers to publish materials, and to provide the public easy and timely access to useful information and knowledge derived from these materials at reasonable cost.

The NASA Information Infrastructure Technology and Applications (IITA) program aims to accelerate development of digital library technologies and applications to let the public obtain NASA information of interest via the NII. NASA participation in the NSF/ARPA/NASA joint initiative is an element of this program. NASA participation provides the projects meaningful access to the vast repositories of scientific information relating to multiple domains and in multiple representations. Since NASA data is public domain and its use is not encumbered by issues related to privacy and intellectual property, it thus offers a ready testbed for research in many areas.

involves "federating" (mapping together similar objects from different collections) in a way that makes them appear as one organized collection. The better the indexing, the better the searching.

Indexing process

Indexing was originally developed for text documents. Each document is segmented into significant words, and a table generated that indicates which words occurred where in what documents. A user can search by specifying a word (or words); the system then supplies the results by looking up the word in the table, and retrieving the documents containing it. For nontextual media, such as video programs or map textures, the segments differ from word phrases, but the process is quite similar. This traditional indexing is automatic but purely syntactic, matching only words that actually appear in the text.

An indexer (usually a human librarian expert in the subject matter) can also generate other words that describe the document, to improve the search. These subject descriptors, called A&I (Abstracting and Indexing) in the library business, capture some semantic content. A&I records are often called metadata, because they describe data properties and are important in indexing no matter the type of object. (That is, map librarians are as concerned with metadata as document librarians.)

A&I suffers from the economics and energies of human activity; that is, it is only available for large collections on major subjects and does not change as quickly as the words in the collections change. For this reason, much research in digital libraries concentrates on automatic or semiautomatic semantic indexing. As the repositories become more specialized (for small communities instead of large subjects), automatic indexing will become more important.

Federating process

The traditional form of federating was also developed by using collections of text documents. A common gateway is developed that transforms the user's query language into the query language of each search engine for each collection index. Current technology is largely syntactic: It concentrates on sending a query in the appropriate format to each engine, at best taking account of the metadata structure. So a user specification of an AUTHOR field could be mapped uniformly into variant field names, such as AU or AUT or AUTHOR, for different collections. However, different meanings for AUTHOR would simply be ignored by the mapping. A slightly more semantic federation uses a canonical document structure, like mapping together variants into a standard set of authors or equations. This structure mapping is seen today primarily with text standards such as SGML, since text is by far the most studied structure.

Semantic difficulties

A topic of active research is how to map content or meaning across collections—how to approach semantics. For example, a bridge designer concerned with the struc-

Digital Library Initiative project site information

The Digital Library Initiative is sponsored by the National Science Foundation, Advanced Research Projects Agency, and the National Aeronautics and Space Administration from September 1994 to August 1998. The initiative's focus is to dramatically advance the means to collect, store, and organize information in digital forms, and make it available for searching, retrieval, and processing via communication networks. Readers can contact individual projects at the following addresses. The Web pages contain extensive and current project information.

Carnegie Mellon University
Project: Full-content search and retrieval of video
Principal Investigator: Howard Wactlar, wactlar@cs.cmu. edu
Web site: http://fuzine.mt.cs.cmu.edu/im/informedia.html
Contact: Colleen Everet, cae@cs.cmu.edu, (412) 268-7674

Stanford University
Project: Interoperation mechanisms among heterogeneous services
Principal Investigator: Hector Garcia-Molina, hector@cs.stanford.edu
Web site: http://Walrus.Stanford.EDU/diglib/
Contact: Maryanne Siroker, siroker@cs.stanford.edu, (415) 723-0872

University of California at Berkeley
Project: Work-centered digital information services
Principal Investigator: Robert Wilensky, wilensky@cs.berkeley.edu
Web site: http://elib.cs.berkeley.edu/
Contact: Crystal Williams, crystal@cs.berkeley.edu, (510) 642-0930

University of California at Santa Barbara
Project: Spatially-referenced map information
Principal Investigator: Terrence R. Smith, smithtr@cs.ucsb. edu
Web site: http://alexandria.sdc.ucsb.edu/
Contact: Patty Towne, towne@alexandria.sdc.ucsb.edu, (805) 893-7665

University of Illinois at Urbana-Champaign
Project: Federating repositories of scientific literature
Principal Investigator: Bruce Schatz, schatz@uiuc.edu
Web site: http://www.grainger.uiuc.edu/dli
Contact: Susan Harum, dli@uiuc.edu, (217) 244-8984

University of Michigan
Project: Intelligent agents for information location
Principal Investigator: Daniel Atkins, atkins@umich.edu
Web site: http://http2.sils.umich.edu/UMDL/HomePage.html
Contact: Laurie Crum, lcrum@umich.edu, (313) 763-6035

tural effects of wind might want to compare the literature and simulations in the civil engineering digital library to those in the library concerning undersea cables, since the problems with the stability of long structures swaying in a fluid medium is similar. The difficulty is that the terminology and metadata are quite different for the fluid dynamics of air and of water, even though the concepts and ideas are quite similar.

Technology for solving the "vocabulary problem"[4] would enable users to search digital libraries in unfamiliar subjects by specifying terms in their own domain and having the system translate these to terms in the target domain. Over many years, researchers have tried many techniques to automatically translate vocabulary across domains. Natural language-parsing techniques, for example, have been extensively tried but are largely unsuccessful for effective search beyond a narrow domain where they can be hand-tuned. And little research has been done on similarity matching for objects beside text documents.

The most promising general techniques do statistical analysis for information retrieval. These are becoming computationally feasible as machines become faster. There are already instances of building similarity indexes across large collections.[5] These computations are being done on today's supercomputers, which are tomorrow's personal computers. These and other techniques must be developed to enable the Grand Challenge of Semantic Interoperability to be solved so that users can transparently and effectively search the Net.

THE FUTURE

The technology for information retrieval for large collections has remained basically unchanged for 30 years. The technology that ran on specialized research machines in the 1960s and on commercial on-line systems in the 1970s are still serving millions of Web users today. The government initiatives in the early 1960s spawned Dialog and Lexis/Nexis in the 1970s, and government initiatives in the early 1990s such as ARPA's CSTR (Computer Science Technical Report) produced the Lycos and Yahoo Web searchers.

The structure of the flagship DLI projects, with large testbeds and many partners, is again set up to encourage technology transfer of new developments. What the DLI projects promise is effective search of multimedia objects across multiple repositories in the Web. In the longer term, there is even hope for semantic interoperability, which is necessary to handle the coming variability and volume of electronic materials.

Just as propagating the data-access technology of packets in the ARPANET required adopting and evolving standards, so will propagating the Internet's information-organization technology. The D-Lib Forum (http://www.dlib.org) is acting as IITA's coordinating body for digital library research and development.

Finally, after searching transparently across collections becomes possible, research in the technology of network information systems will move to the next stage. This next wave promises to be information analysis: systems for cross-correlating items of information across multiple sources. Today on the Web you can fetch things by browsing documents. Tomorrow on the Web you will find things by searching repositories. In the new millennium beyond the Web, analysis environment technology will let you correlate things across repositories to solve problems.[6] ❙

References

1. R. Pool, "Turning an Info-Glut into a Library," *Science*, Oct. 7, 1994, pp. 20-22.
2. G. Taubes, "Indexing the Internet," *Science*, Sept. 8, 1995, pp. 1,354-1,356.
3. *Interoperability, Scaling, and the Digital Library Research Agenda*, IITA report, 1995; (http://www-diglib.stanford.edu/ diglib/pub/reports/iita-dlw/main.html).
4. H. Chen, "Collaborative Systems: Solving the Vocabulary Problem," *Computer*, May 1994, pp. 58-66.
5. B. Schatz et al., "Federating Diverse Collections of Scientific Literature," *Computer*, May 1996, pp. 28-36.
6. B. Schatz et al., *Building the Interspace*, 1996, (http://csl.ncsa.uiuc.edu/interspace.html).

Bruce Schatz *is principal investigator of the Digital Library Initiative project at the University of Illinois and a research scientist at the National Center for Supercomputing Applications, where he is the scientific advisor for digital libraries and information systems. He is also an associate professor in the Graduate School of Library and Information Science, the Department of Computer Science, and the Program in Neuroscience. He holds an NSF Young Investigator award in science information systems. Schatz has worked in industrial R&D at Bellcore and Bell Labs, where he built prototypes of networked digital libraries that served as a foundation of current Internet services (Telesophy), and the University of Arizona, where he was principal investigator of the NSF National Collaboratory project that built a national model for future science information systems (Worm Community System).*

His current research in information systems is building analysis environments to support community repositories (Interspace), and in information science is performing large-scale experiments in semantic retrieval for vocabulary switching using supercomputers. Schatz received an MS in artificial intelligence from Massachusetts Institute of Technology, an MS in computer science from Carnegie Mellon University, and a PhD degree in computer science from the University of Arizona.

Hsinchun Chen *is an associate professor of Management Information Systems at the University of Arizona and director of the Artificial Intelligence Group. He is the recipient of an NSF Research Initiation Award, the Hawaii International Conference on System Sciences Best Paper Award, and an AT&T Foundation Award in Science and Engineering. He has published more than 30 articles about semantic retrieval and search algorithms. Chen received a PhD in information systems from New York University.*

Readers can contact Bruce Schatz at NCSA, Beckman Institute, 405 N. Mathews, University of Illinois, Urbana, IL 61801; schatz@uiuc.edu. Hsinchun Chen's address is Dept. of Management Information Systems, McClelland Hall, University of Arizona, Tucson, AZ 85721; hchen@bpa.arizona.edu.