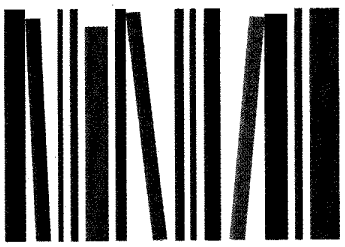
Guest Editors' Introduction

Digital Libraries: Technological Advances and Social Impacts



Public awareness of the Net as a critical infrastructure in the 1990s has spurred a new revolution in the technologies for information retrieval in digital libraries.

Bruce Schatz
University of Illinois at Urbana-Champaign

Hsinchun Chen
University of Arizona

The World Wide Web has made access to the Internet part of the structure of everyday life. Millions of people all over the world search the Web every day. But the commercial technology of searching large collections has remained largely unchanged since the 1960s, when it was developed in the course of US government-sponsored research projects.[1] This public awareness of the Net as a critical infrastructure in the 1990s has spurred a new revolution in the technologies for information retrieval in digital libraries.

Many believe that we are approaching the start of the Net Millennium, a time when the Net forms the basic infrastructure of everyday life. For this transformation to actually occur, however, the functionality of the Net must be boosted beyond providing mere access to one that supports truly effective searches. Collections of all kinds must be indexed effectively, from small communities to large disciplines, from formal to informal communications, from text to image and video repositories, and eventually across languages and cultures. The Net needs fundamentally new technology to support this new search and indexing functionality.[2]

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 by undertaking multiyear large-scale research projects that develop real-world electronic testbeds used by actual users and by aiming 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.

This special issue describes a wide range of research projects that investigate the development and usage of new information technology for substantial collections. The technologies contained within are a representative sample of the Net of the early 21st century. Particular emphasis is placed on retrospective papers from multiyear projects, which reflect actual experiences on an experimental basis with the use of new technologies. The issue thus also contains initial hints of the user experiences that will be common in the future Net.

RESEARCH INITIATIVES

In May 1996, a special issue of *Computer* focused specifically on a major new US government initiative—the Digital Libraries Initiative (DLI)—funded by the NSF, DARPA, and NASA. The six major projects supported by the DLI each had a survey paper at this halfway point in the initiative.

This issue focuses on practical outcomes from research projects—major research testbeds and fundamental research technologies that show what the large-scale future infrastructure might become. The papers are split between DLI and non-DLI projects. Digital libraries have become far more important nationally and internationally in 1999 than in 1996. This is largely due to the exponential growth of information in the World Wide Web, which Web searchers are increasingly failing to handle successfully. This is a special case of the increasing dependence of modern society on information technology and the increasing failure of fundamental infrastructure due to the absence of fundamental new technology.

The just-released PITAC report (President's Information Technology Advisory Committee) makes this point clearly.[3] In this report, the leaders of the US information technology research community concluded that "the current Federal program is inadequate to start necessary new centers and research programs....The end result is that critical problems are going unsolved and we are endangering the flow of ideas that have fueled the information economy."

0018-9162/99/$10.00 © 1999 IEEE

The committee went on to recommend that "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."

Digital Libraries Initiative-Phase 2 (DLI-2) is an NSF-led initiative that builds on the successes of DLI-1 and presages the even bigger efforts recommended in the PITAC report. DLI-2 has made the initial awards for multiyear projects that will support a broader range of activities than DLI-1, including smaller projects and topics in medicine and humanities. There will be an even stronger emphasis on testbeds with real users and real collections.

Many federal agencies are contributing to this initiative—namely NSF, DARPA, NASA, National Library of Medicine (NLM), Library of Congress, and the National Endowment for the Humanities. The "Funding Agencies" sidebar includes a contribution from the NSF program officer discussing DLI-2, as well as contributions from the lead agencies DARPA and NLM describing their agencies' other efforts to support digital library research.

The importance of digital library research is spread-

Funding Agencies

Digital Libraries: The View from NSF

Stephen Griffin, National Science Foundation

The Internet and WWW have demonstrated that scholars, students of all ages, and the general public have a boundless appetite for information of all types. Millions now regularly use the Web as a primary source of information, and as an inventive medium for communicating and sharing knowledge, enabling new relationships, collaborations, and intellectual communities.

The Digital Libraries Initiative (DLI), funded by NSF, DARPA, and NASA from 1994 to 1998, supported pioneering exploration into issues of organization, access, security, and use of distributed information resources. DLI demonstrated that large amounts of heterogeneous information can be organized into coherent, interoperable collections in computing laboratory settings, and that these can be searched and manipulated in new ways to yield useful knowledge. The six DLI projects addressed a broad range of fundamental research: new document models, video capture and indexing, geographic data spaces, image retrieval, concept spaces, agent-based synthetic global economies, and new tools for classroom education, to name a few. (See *DLI National Synchronization* at http://dli.grainger.uiuc.edu/national.htm.)

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 in the digital libraries information life cycle, including content creation, access, use and usability, preservation, and archiving. (See *DLI-2* at www.dli2.nsf.gov) DLI-2 will look to create domain applications and operational infrastructure, and understand their use and usability in various organizational, economic, social, and international contexts. In short, DLI-2 will investigate digital libraries as human-centered systems. DLI-2 involvement will extend far beyond computing and communications specialty communities to 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 sponsoring projects through a new program in International Digital Libraries Collaborative Research.

Many of the most important research questions regarding systems and use are bound into the process of building and using real-world operational systems. DLI was characterized by a single project model addressing a broad, technology-centered research agenda and building technology testbeds. Content was of secondary concern and acquired primarily through donations. DLI research illuminated the complexity and difficulty of fundamental issues of functionality, scalability, interoperability, reliability, and usability.

Investigations into these and related technologically grounded questions will continue in DLI-2, but until large-scale distributed systems are built, instrumented, filled with content of value, and open to use by large and diverse populations, many important questions will go unanswered. The DLI projects reached their most potent stage as research enterprises toward the end of their funded term as the testbeds matured and became heavily used. To scientifically understand how large-scale distributed digital libraries behave in a global information environment, and how they might be used to the good of society, we must first begin to build and use them.

The recent President's Information Technology Advisory Committee *Interim Report to the President* (Aug. 1998, www.ccic.gov/ac/interim) notes that current agency practices and modes of support for IT projects are ineffective in addressing research areas that require a large- or medium-sized team and a focused effort of more than a few years.

Digital library technologies are a natural outcome of earlier federal funding of high-end computing systems and high-performance networks. Digital libraries are among the first and most promising generations of applications to exploit and validate the continuing development of these basic technologies and services. Digital libraries fit the bill for new federal plans to provide for the nation's information needs in the 21st century through networked computing.

ing beyond the US. The "International Activities" sidebar includes contributions describing the developing activities in Europe and Asia, based on results from recent technical workshops. The sidebar concludes with the past president of the International Federation of Library Associations discussing political and economic difficulties of spreading research technologies into practical systems for searching across languages and across cultures.

The articles in this issue are careful retrospectives on multiyear digital library research projects, which discuss large-scale testbeds for text documents and fundamental technologies for semantic interoperability beyond text.

LARGE-SCALE TESTBEDS

Building an experimental testbed is an accepted methodology for evaluating networked information systems. A testbed is a prototype system with real collections and real users, but supported as a research rather than a commercial product. Many national policy committee reports such as the NRC National Collaboratories,[4] the NSF DLI-2 Planning,[5] and the PITAC[3] have emphasized the necessity of large-scale testbeds as the

Digital Libraries: The View from DARPA

Ronald Larsen, Defense Advanced Research Projects Agency

Speed and precision, central to DARPA's information management objectives, were addressed through vertical integration of information systems in the 1980s. But in the 1990s, this model broke down. Explosive growth of the WWW made vertical integration an inadequate response, and was replaced by an emphasis on digital libraries. DARPA's Information Management program (www.darpa.mil/ito/ research/im) addresses core digital library issues requiring revolutionary research technology.

- *Federated repositories.* The organization of distributed repositories into a coherent virtual collection is fundamental, as demonstrated by the Networked Computer Science Technical Reference Library (NCSTRL).
- *Scalability.* Managing billions of digital objects and millions of sources poses challenges in identifying, categorizing, indexing, summarizing, and extracting content.
- *Interoperability.* Digital libraries require semantic interoperability among heterogeneous repositories distributed across the network.
- *Collaboration.* Analysts work in distributed teams, building on each other's knowledge, experience, and, resources.
- *Testbed development.* DARPA is establishing four DLI sites and NCSTRL as partners in providing stable, accessible test collections to the research community. (www.dlib.org)
- *Communication.* Timely dissemination of research results is the focus of D-Lib, now mirrored internationally.

Defense requirements challenge digital library technology, both qualitatively and quantitatively. DLI illuminated the complexity and difficulty of scalability, interoperability, and usability. DARPA's information management research program reflects a continuing commitment to these issues.

Digital Libraries: The View from NLM

Milton Corn, National Library of Medicine

Health-related activities depend on vast seas of information. The physician's information needs differ from those of the molecular biologist, the patient, and the public health scientist, but whoever the questioner, the information is likely to be in a variety of locations and in multiple formats, including print, images, graphics, and video. Massive databases of clinical information are becoming ubiquitous, but use of digitized information is complicated by problems generic to digital libraries for any domain: behavior and cognition issues, lack of standards, legacy systems, distributed data, the need to network among heterogeneous systems, inefficient information retrieval, and, particularly for patient data, privacy concerns.

The National Library of Medicine, a component of the National Institutes of Health, is the world's prime repository of biomedical information. The Library maintains and distributes without charge on the Web a number of widely used databases, such as MEDLINE, the bibliographic reference of biomedical literature, and GenBank, a key resource for molecular biologists. The Library has been a pioneer in technology to manage information and supports research and development work in medical informatics.

As examples of large, complex databases in text and in images, NLM is offering two of its resources to the DLI-2 community:

- the Unified Medical Language System (www.nlm.nih.gov/research/umls/umlsmain.html), an ambitious project of fundamental semantic importance for digital libraries, which is mapping concepts across the myriad thesauri and vocabularies of biomedicine; and
- the Visible Human Project (www.nlm.nih.gov/research/visible/visible_human.html), by far the most comprehensive online image repository of the male and female human body.

Because digital libraries, utilization of distributed databases, and data-mining issues are of national importance to healthcare delivery and to research, the National Library of Medicine recognizes that a joint agency initiative, such as DLI-2, attacks problems generic to digital libraries, and can both profit from and benefit biomedical application. NLM has a particular interest in projects of value to consumers, but welcomes work relevant to computerized patient records, images, computational biology, education, public health, or any other health-related area.

only method for determining which information system features are actually useful in practice.

New technologies in digital libraries emerge from large-scale research testbeds. To obtain the requisite collections and users, these projects have concentrated on text documents, particularly articles already available in electronic form. Text dominates use of information in the scholarly world, where experiments could potentially be run. Thus, these representative papers on digital library testbeds concentrate on journal articles served to scholarly populations.

The Illinois DLI project was a classic testbed project, developing new technology and deploying it widely on an experimental basis. The Illinois project chose as its research paradigm the complete manipulation of structured documents—namely, the search and display of engineering journal articles encoded in Standard Generalized Markup Language (SGML). The project developed federated search of document structure across multiple repositories from multiple publishers, which was deployed in a testbed around campus.

The Illinois DLI project was a research project developing and experimentally testing new technology for federated search, by deploying real collections to real users on a production basis. The JSTOR project, in contrast, was intended to become a commercial service, now used by many academic institutions. They chose the mature technology of digitized bitmaps (page images) rather than the immature technology of SGML markup.

Many of the current generation of digital library research testbeds are turning into production services. For example, the DARPA D-Lib Test Suite[6] is providing continuing support for several of the DLI and related testbeds, and is actively seeking users to experiment further with these testbeds. These experiences give the first indication of usage patterns for search in the Net of the 21st century.

SEMANTIC INTEROPERABILITY

The challenge of digital libraries has remained unchanged from the goals described in the introduction to the 1996 special issue.[7] The DLI projects pursued deep semantic interoperability, making heterogeneous items in heterogeneous sources spread across the net-

International Activities

Digital Library Research in Europe

Alan Smeaton, Dublin City University, Ireland

European research into digital libraries (DLs) is funded by the European Union as well as by national sources. Several countries, such as the UK, have launched specific DL research programs. At the European level, the Fourth Framework Programme of the Commission of the European Union is now concluding without having had a specific research program in DLs, although DL projects have been supported by the Information Engineering in Europe (www.echo.lu/ie), Language Engineering (www.echo.lu/langeng/en/lehome.html), and Esprit (www.cordis.lu/esprit) programs.

DELOS (www.iei.pi.cnr.it/DELOS) is a DL working group and is part of the ERCIM (European Research Consortium for Informatics and Mathematics; www.ercim.org) Digital Library Initiative funded by Esprit within the Fourth Framework Programme. Its objective is to simulate DL research and collaboration. DELOS achieves this through workshops, including the series of European Conferences on Research and Advanced Technology for Digital Libraries.

One of the most exciting developments in European DL research is an NSF-EU collaboration that has formed five working groups in the key technical areas of interoperability, metadata, intellectual property rights, resource indexing and discovery, and multilingual information access. Emerging from these working groups will be a white paper driven by researchers and scientists in the DL area, which will recommend a research agenda for joint research in digital libraries. It is hoped that this will influence the content of the EU Fifth Framework Programme, the next major wave of technology funding in Europe.

Digital Library Research in Asia

Hsinchun Chen, University of Arizona
Jerome Yen and Chris Yang, University of Hong Kong

Since 1995 digital library research has become a national grand challenge in several countries in Asia. Most projects can be classified into the following categories:

- Nationwide digital library initiative and special-purpose digital libraries—for example, the Library 2000 Project in Singapore (to link all library resources in Singapore) and the Financial Digital Library at the University of Hong Kong (to serve the needs of the Hong Kong stock market and users).
- Digital museum and historical document digitization—for example, the Digital Museum Project of the National Taiwan University and the digitization of the art collection of the Palace Museum in Taipei by IBM.
- Local language and multilingual information retrieval—for example, the Net Compass Project of Tsinghua University in China, Chinese Information Retrieval at the Academia Sinica, Taiwan, and New Zealand's multilingual project.

Local language processing and historical cultural content could be the most immediate Asian contribution to the international DL community. There is significant interest among Asian DL researchers in exchanging research ideas and collabo-

work appear to be a single uniform federated source.

Federating the search at a semantic level is an area of active research in the digital library community. Statistical approaches in particular are leading the way toward *scalable semantics*—indexing deeper than text word search that is computable on large real collections. For example, *concept spaces*, which capture contextual information, have been computed for collections of millions of documents.[8,9]

Semantic interoperability beyond federated search also involves making multiple sources appear as a single source, or making single systems with multiple functions. The Carnegie Mellon DLI project searched multimedia, particularly video segments, by generating text indexes using speech understanding. The New Zealand project searched multilingual documents, as well as nontextual search by singing a musical phrase into a folk-song database. The Stanford DLI project searched across different engines using multiprotocol gateways. These articles represent a good sample of current research technology. Other even harder issues remain untouched, such as multicultural search across context and meaning.

THE NET OF THE 21ST CENTURY

In the Net of the 21st century, there will be a billion repositories distributed over the world, where each small community maintains a collection of their own knowledge.[1] Semantic indexes will be available for each repository, using scalable semantics to generate search 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.[10]

The Internet will have been transformed into the *Interspace*, where users navigate abstract spaces to perform correlation across sources.[11] Information analysis will become a routine operation in the Net, performed on a daily basis worldwide.[12] Such functionality will first be used by specialty professionals and then by ordinary people, just as has occurred with text search. Information infrastructure will become the essential part of the structure of everyday life, and digital libraries will become the essential part of information infrastructure.

This issue of *Computer* gives retrospectives for a representative sample of the major research projects in digration on projects. The University of Library and Information Science in Japan has sponsored a series of International Symposia on Digital Libraries (www.DL.ulis.ac.jp). Commercial companies have sponsored DL workshops, such as Sun Microsystem's 1997 workshop in Beijing, and IBM's Asian workshops.

More recently, the First Asia Digital Library Workshop was held in Hong Kong in August 1998. The workshop, which focused on Asian DL research projects, attracted more than 120 participants from nine Asia/Pan-Pacific countries. It has served as the catalyst for Asian DL collaborations. Several countries have expressed strong interest in sponsoring a Second Asian Digital Library Workshop. An Asia Digital Library Consortium is fostering long-term collaboration and projects in DL-related topics in Asia (see www.cyberlib.net/adl).

Digital Libraries: A Global Connection

Robert Wedgeworth, University of Illinois at Urbana-Champaign and Int'l Federation Library Associations

Digital libraries are emerging in many parts of the world to give access to the world's scientific developments. Searching across these multidisciplinary repositories will be a daunting task. Some of the research in the Digital Libraries Initiatives is beginning to indicate that this can be done. But what about access to the generations of scientific knowledge that already exist in the world's libraries?

A global interconnected library network of existing collections has begun to emerge. Its Web site (www.ifla.org) is maintained by the International Federation of Library Associations and Institutions (IFLA). Recognizing some years ago that the resources would never be available to replicate the collections, facilities, staffing, and technologies of modern libraries in all parts of the world, IFLA launched an electronic network strategy in 1994 as an experiment to connect its members to its conference in Havana, Cuba. By 1995, libraries in more than 70 countries were connected. At the beginning of 1998, libraries in more than 100 countries were connected to this global library network, sharing information and expertise.

Although there are concerns that a technology-based strategy for library development could exclude some parts of the world, many libraries report that needing to connect to the international world has assisted them in obtaining the resources to connect to the Internet. Others have found that rapid communications with libraries in their own region has facilitated access to scientific information as well as access to advice and consultations that previously would have taken many weeks.

Connectivity and training continue to be the principal barriers to integrating the global network of libraries with the emerging digital libraries. However, the existence of a global communications network facilitates training as well as access to scientific information. The community of learners and researchers desiring access to scientific information will require many librarians and information specialists to assist in navigating complex search and retrieval systems across heterogeneous repositories. Further integration of existing repositories of scientific information with the emerging digital information systems will be necessary for the scientific community to retain access to the full record of scientific progress. Utilizing such a global network will render scientific communications completely independent of space and time.

ital libraries. The fundamental new technology surveyed here stands a good chance of becoming a fundamental part of everyday life in the foreseeable future. ❖

References

1. B. Schatz, "Information Retrieval in Digital Libraries: Bringing Search to the Net," *Science*, Vol. 275, Jan. 17, 1997, pp. 327-334.
2. J. Alper, "Assembling the World's Biggest Library on Your Desktop," *Science*, Vol. 281, Sept. 18, 1998, pp. 1,784-1,786.
3. K. Kennedy and W. Joy, chairs, President's Information Technology Advisory Committee (PITAC), *Interim Report to the President*, Aug. 1998, www.ccic.gov/ac/interim.
4. V. Cerf and W. Wulf, eds., *National Collaboratories: Applying Information Technology for Scientific Research*, National Academy Press, Washington, D.C., Mar. 1993.
5. D. Atkins, ed., "Digital Libraries: Report of the Santa Fe Planning Workshop on Distributed Knowledge Work Environments," Mar. 1998, www.si.umich.edu/SantaFe.
6. W. Arms, ed., *D-Lib Test Suite*, Summer 1998, www.dlib.org/test-suite.
7. B. Schatz and H. Chen, "Building Large-Scale Digital Libraries," *Computer*, Vol. 29, May 1996, pp. 22-26.
8. H. Chen et al., "A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Project," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 18, Aug. 1996, pp. 771-782.
9. J. Alper, "Taming MEDLINE with Concept Spaces," *Science*, Vol. 281, Sept. 18, 1998, p. 1,785.
10. H. Chen et al., "A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System," *J. Am. Soc. Information Science*, Vol. 48, Jan. 1997, pp. 17-31.
11. B. Schatz, "High-Performance Digital Libraries: Building the Interspace on the Grid," *Seventh IEEE Int'l Symp. High-Performance Distributed Computing*, July 1998, pp. 224-234.
12. *Interspace Prototype*, www.canis.uiuc.edu.

Bruce Schatz is director of the Community Architectures for Network Information Systems (CANIS) Laboratory at University of Illinois at Urbana-Champaign and a professor in the Graduate School of Library and Information Science.

Hsinchun Chen is a professor in the Department of Management Information systems at the University of Arizona and director of the Artificial Intelligence Lab.

Contact the authors at schatz@canis.uiuc.edu or hchen@bpa.arizona.edu.