Digital Libraries: Technological Advances and Social Impacts

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. 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.

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. 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.”
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-
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 (N CSTR L).
- 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 N CSTR L as partners in providing stable, accessible test collections to the research community. (www.dlib.org)

...
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 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. 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 Arizona

Hsinchun Chen, University of Arizona

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.5,6

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.7 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.8,9

The Internet will have been transformed into the Interspace, where users navigate abstract spaces to perform correlation across sources.10 Information analysis will become a routine operation in the Net, performed on a daily basis worldwide.11 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 dig...
Digital libraries. The fundamental new technology surveyed here stands a good chance of becoming a fundamental part of everyday life in the foreseeable future.

References

Bruce Schatz is director of the Community Architectures for Network Information Systems (CANIS) Laboratory at the 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.