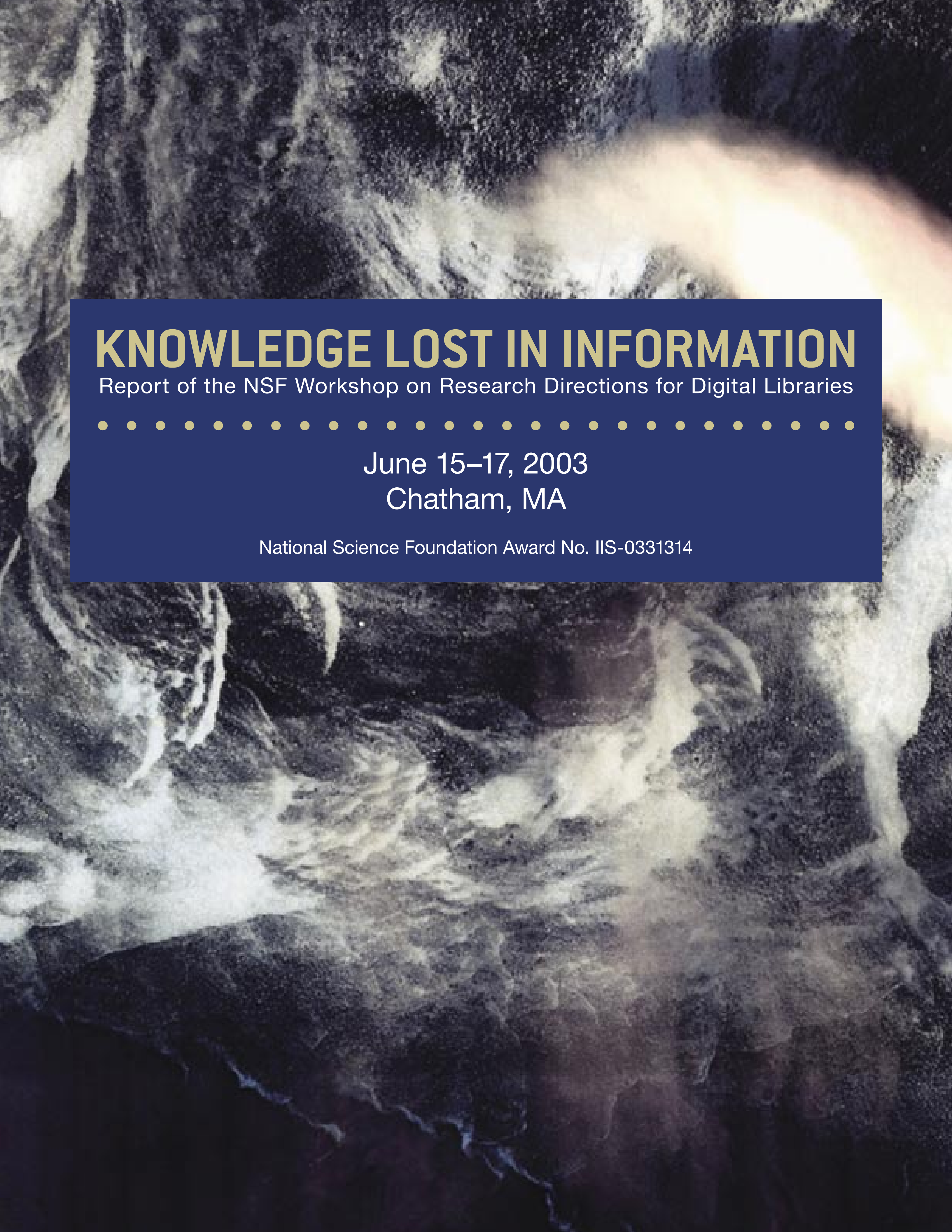


KNOWLEDGE LOST IN INFORMATION

Report of the NSF Workshop on Research Directions for Digital Libraries

June 15–17, 2003
Chatham, MA

National Science Foundation Award No. IIS-0331314

Acknowledgements

The authors are grateful to the National Science Foundation for the support that afforded us the opportunity to convene an inspiring group of individuals who care deeply and have thought rigorously about the future of digital libraries and the research necessary to realize such a future. We are also indebted to the proprietor and staff of the Wequassett Inn, who provided a splendid venue on Cape Cod for creative thought and reflection.

The authors are also indebted to Amy Friedlander, who edited this report, and to the participants who invested post-workshop time in reviewing earlier drafts. Assembling the contributions of such a diverse group of participants into a coherent narrative is nothing if not challenging.

The cover image is provided courtesy of the National Aeronautics and Space Administration.

The workshop and this report were made possible by grant NSF-IIS-0331314 to the School of Information Sciences at the University of Pittsburgh.

The University of Pittsburgh is an affirmative action, equal opportunity institution. Published in cooperation with the Department of University Marketing Communications. UMC4662-0304

Preface

The first decade of digital library research provided ample evidence that humankind's ability to generate and collect data exceeds our ability to organize, manage, and effectively use it. This trend is unlikely to abate without continued research and development. For the foreseeable future, data of many types will be increasingly abundant and "technologically" available. But these data will continue to seem chaotic, lacking sufficient organization, stability, and quality control. Moreover, individuals and communities may lose the ability to control access to and manage their own data. The effective use of data and information resources must scale with their ever-increasing abundance and variety; this will take continued research and technology development.

Together with the recently articulated Cyberinfrastructure effort as well as digital libraries programs in state and federal agencies and the contributions of major philanthropic organizations, the National Science Foundation (NSF) and its community of researchers can point to broad support for digital libraries. Disciplines from the social to the physical sciences are engaged in aspects of digital library development for populations of users from preschoolers to seniors, addressing national priorities from education to health care and defense. The integration of digital libraries, data grids, and persistent archives is actively under way, enabling an infrastructure that supports collection, publication, sharing, storage, and preservation of information in a variety of forms. While each community focuses on material and functions that are crucial to its domain, a common set of principles is emerging.

An ecology of projects and programs that meet near- and long-term goals lets us attack many of the hard problems that still remain while not losing sight of the more elusive questions that lurk at the frontiers of research. Progress in the near term will continue through collaborative work exemplified by programs such as the Library of Congress National Digital Information Infrastructure and Preservation Program, the National Archives and Records Administration persistent archive prototypes, and the NSF National Science Digital Library integration infrastructure, among many others. While we continue to evolve the critical infrastructure and technology required to sustain the digital information revolution, we must also accelerate research to transform this infrastructure into the advanced digital knowledge environment that will enlighten and empower the next generation. This report, based on a workshop held in Chatham, Massachusetts, in June 2003, sets forth the near- and long-term steps leading us toward realizing this vision.

Ronald L Larsen

Ronald L. Larsen
Dean, School of Information Sciences
University of Pittsburgh

Howard D. Wactlar

Howard D. Wactlar
Vice Provost for Research Computing
Carnegie Mellon University

Table of Contents

Executive Summary

Digital libraries are transforming research, scholarship, and education at all levels. Vast quantities of information are being collected and stored online, and organized to be accessible to everyone. Substantial improvements in scholarly productivity are already apparent. Digital resources have demonstrated the potential to advance scholarly productivity, easily doubling research output in many fields within the next decade. These resources can also become primary resources for education, holding the potential for advances in lifelong learning that have been sought for many years. But such progress will not be achieved without investment. This report details the nature of the federal investment required to sustain the pace of progress.

The National Science Foundation (NSF) conceived and supported this research, resulting in U.S. leadership in digital information, a position that would be easy to lose to international competition but is not expensive to retain. During the last few years, digital library research has become the most interdisciplinary area at NSF, including researchers from 35 different academic departments. The program has also engaged international partners, with several U.S. projects coordinated with counterpart projects in the United Kingdom and Germany, as well as with broader international projects involving the European Union and Asian countries. Moreover, the kinds of information created and examined have moved well beyond text and book-like objects to include CT-scans of fossils, images of dolphin fins, cuneiform tablets, and videos of human motion, potentially enabling more sophisticated analysis in domains that range from archaeology and paleontology to physiology, while exploring the engineering problems that such investigations expose.

To maintain national expertise and achieve necessary gains in research and education, the NSF should:

1. Provide $20 million annually for innovative and exploratory research to address challenges in the creation, collection, organization, use, and long-term availability of digital resources of all sorts in a rapidly evolving global information infrastructure;

2. Provide $40 million annually for transformative change to infrastructure and practice, including:

   a) Integrative deployment, enhancement, and empirical evaluation of sustainable digital resources and services;

   b) Periodic review by a multidisciplinary advisory panel of progress and priorities in light of evolving technological and societal conditions; and

   c) Research on the organizational, social, and policy implications of advancing information infrastructure.

Each of these components is required, and experience over the past decade demonstrates the high return on investment from both forms of investment in digital libraries.

New digital information resources pose issues different and distinct from those of conventional resources, and, hence, require specific attention. Progress toward a future built on digital knowledge will require a commitment to not only establishing but also sustaining these critical resources over an extended period.

Our ability to generate and collect digital information continues to grow faster than our means to organize, manage, and effectively use it. This trend is likely to continue without focused research and development. For the foreseeable future, data of many types will be increasingly abundant and “technologically” available. But these data will continue to seem chaotic, lacking sufficient organization, stability, and quality control. Individuals and communities may lose the ability to control access to and manage their own data, thus compounding the problem. The effective use of data and information resources must scale with their ever-increasing abundance and variety; this will take continued research and technology development.

Successfully addressing these issues is vital for us to continue to function effectively and responsibly as individuals, as members of organizations, and as a society across decades and generations. Research breakthroughs in these areas can enhance our ability to conduct scholarship and science, improve education and learning, make our industries and government more effective and more competitive, and give birth to entirely new technology-driven industries. They can also ensure that our cultural memory organizations continue to evolve and function in a responsive and appropriate way.

While major progress has been made in indexing, searching, streaming, analyzing, summarizing, and interpreting multimedia data, the more that is accomplished exposes the more that remains to be done. Interactive environments for knowledge creation, use, and discovery need to move out of the laboratory and be broadly deployed in society. Systems for information access, delivery, and presentation are in a continual state of catch-up as they scale to the ever-increasing generative capabilities of sensor networks and related information sources. Increasing demands are being placed on knowledge access, creation, use, and discovery across disciplines, and on content interpretation across linguistic, cultural, and geographic boundaries. The opportunities are unlimited, but they will remain only challenges unless a continued commitment by NSF sustains and accelerates research into the most fundamental of our intellectual assets—*information*.

Developing the infrastructure to support new modes of scholarly inquiry and communication requires more than software development. Experience has already shown that it engages the best minds and talents of both domain specialists and computer and information scientists. Research and technology challenges have emerged across a spectrum of computing disciplines, including network and systems design, human computer interaction, artificial intelligence, information retrieval, information organization, machine translation, database systems, and complexity theory.

Solutions for many of these challenges begin with the information itself—understanding, organizing, managing, disseminating, and preserving it. Digital libraries provide these services, and when done best, they are barely noticed. Technologists describe these types of services as “transparent.” But transparent does not mean nonexistent. Indeed, they are achieved only by long-term, focused research and development, but, in the end, they enable users to enrich their lives in ways never before achieved. Appropriately constituted, research in digital libraries embraces a broad range of topics, from those traditionally found in computer science and engineering to those that arise in the social and behavioral sciences.

The real challenge, therefore, is to build systems supporting scholarly inquiry and communication that yield new capabilities and capacities so effectively and efficiently that they are intuitive and transparent in their operation. Indeed, a serious measure of success may be how simple the resulting infrastructure appears to operate to the casual (and serious) user. This is what we refer to here as the Ubiquitous Knowledge Environment, or the “information ether.”

A technology-centric vision of a Ubiquitous Knowledge Environment draws in part on human-centric computing, the cornerstone of which is individualized, customized, human-centric information. Realizing such a vision is dependent upon long-term advances in the disciplines of information retrieval, storage and communication system technologies, and human-computer interaction that will leverage and harness the methods and mechanisms of artificial intelligence: knowledge representation, language understanding, cognitive processing, and machine learning.

As data, information, and knowledge play increasingly central roles in personal, organizational, and social practices, the next phase of digital library research should focus on:

- Increasing the scope and scale of information resources and services;
- Employing context at the individual, community, and societal levels to improve performance;
- Developing algorithms and strategies for transforming data into actionable information;
- Demonstrating the integration of information spaces into everyday life; and
- Improving availability, accessibility, and, thereby, productivity.

Key to achieving productivity gains is reducing the human overhead required to obtain and use information. Digital libraries offer unparalleled access to information for a far broader range of users than prior physical and organizational arrangements. But gathering, organizing, utilizing, and sharing these information resources requires a scalable, interoperable infrastructure that includes embedded knowledge about services, storage repositories, and content, and is able to bridge context, culture, and language. An appropriate infrastructure program will provide sustainability of digital knowledge resources along five dimensions:

- Acquisition of new information resources;
- Effective access mechanisms that span media type, mode, and language;
- Facilities to leverage the utilization of humankind's knowledge resources;
- Assured stewardship over humanity's scholarly and cultural legacy; and
- Efficient and accountable management of systems, services, and resources.

Digital libraries and the knowledge made available therein have immense potential to contribute to issues of national priority. The realization of this potential requires commitment to a long-term, systematic research program, buttressed by sustainable infrastructure. Outcomes from this research will have untold social and economic payoffs, focusing directly on sustainable and rigorous processes and structures for creating, managing, utilizing, and preserving society's most valuable digital and digitized information resources.

1 Transforming the Information Landscape

Where is the wisdom
we have lost in knowledge?
Where is the knowledge
we have lost in information?
—*T.S. Eliot*

The January 2003 National Science Foundation (NSF) report "Revolutionizing Science and Engineering through Cyberinfrastructure"[1] observed that "multiple accelerating trends are converging and crossing thresholds in ways that show extraordinary promise ... in how we create, disseminate, and preserve scientific and engineering knowledge." That study concluded that the "National Science Foundation should establish and lead a large-scale, interagency, and internationally coordinated Advanced Cyberinfrastructure Program (ACP) to create, deploy, and apply cyberinfrastructure in ways that radically empower all scientific and engineering research and allied education." It envisioned a program "to build more ubiquitous, comprehensive digital environments that become interactive and functionally complete for research communities in terms of people, data, information, tools, and instruments and that operate at unprecedented levels of computational, storage, and data transfer capacity." The report's conclusions resonated with prior recommendations, such as those in the November 2002 NSF/DARPA/NASA report "Technology Requirements for Information Management"[2] which expressed particular interest in the pursuit of research and development to support applications typified as "highly distributed with very large amounts of data and a high degree of heterogeneity of sources, data, and users."

Digital libraries were envisioned a decade ago as the answer to networked knowledge environments. Much progress has been made toward that goal, but in its pursuit, the goal has transformed into something much more dynamic than was originally envisioned. Certainly, the idea of curated, network-accessible repositories—the original notion of "digital libraries"—was (and remains) a fundamental need of scholarly inquiry and communication, as is the notion that these repositories should support information in multiple formats, representations, and media. But not until serious efforts were made in the last decade to build such resources, particularly for non-textual digital content (audio, image, and video, for example) and to reach into a broad range of disciplines in the sciences, social sciences, and the humanities did it become apparent that this venture would stretch the bounds of computer and information science, and, indeed, require the articulated confluence of multiple computer and information science disciplines.

It now appears that these emerging tools and capabilities hold the potential to transform the conduct of disciplinary research itself, and even to foster the creation of new areas of investigation at the interstices of existing disciplines. The fundamental challenge has become building systems supporting scholarly inquiry and communication that yield new capabilities and capacities so effectively and efficiently that they are intuitive and transparent in their operation. Indeed, a serious measure of success may be how simple the resulting infrastructure appears to operate to a wide range of users with diverse information needs. This is what we refer to here as the Ubiquitous Knowledge Environment, or the "information ether."

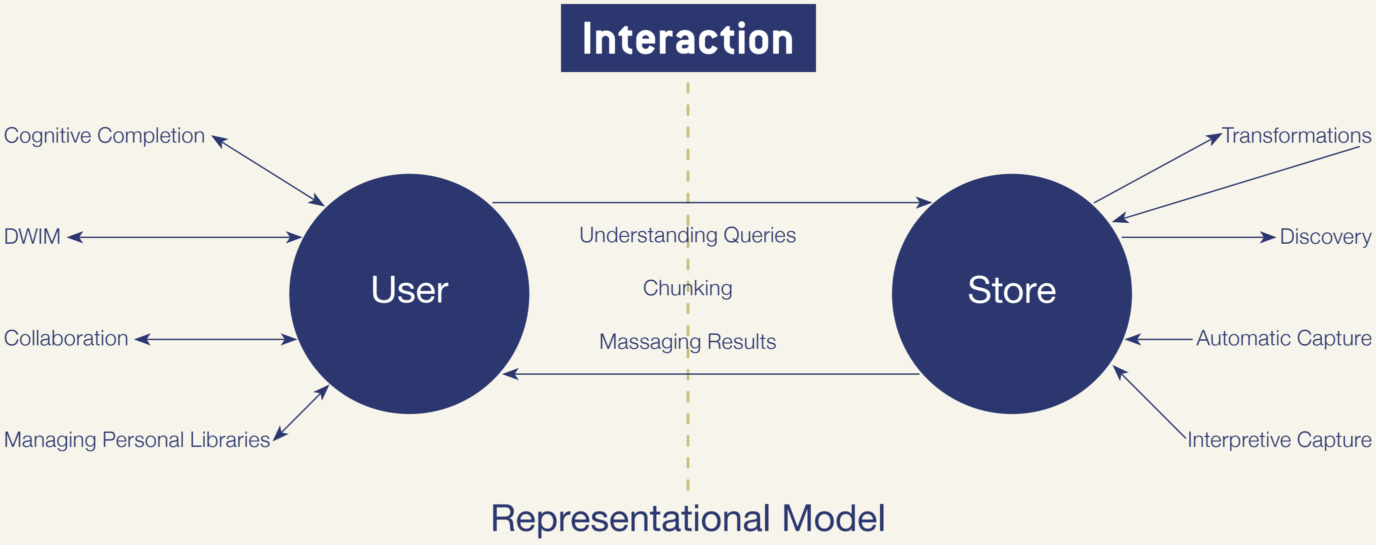
In June 2003, the National Science Foundation sponsored a workshop in Chatham, Massachusetts, involving recognized national and international scholars and researchers (Appendix A) to frame the long-term research necessary to realize such a scholarly communication infrastructure. The panelists recognized that while many difficult problems with important near-term application remain, we must simultaneously address "grand challenge" questions that promise to transform the underlying science. Specifically, continued research is required in the following areas:

- Acquisition of new information resources;
- Effective access mechanisms that span media type, mode, and language;
- Facilities to leverage the utilization of humankind's knowledge resources;
- Assured stewardship over humanity's scholarly and cultural legacy; and
- Efficient and accountable management of systems, services, and resources.

Projects undertaken within these rubrics are likely to have near-term implications. But as important as these broad topics are, order of magnitude advances toward a Ubiquitous Knowledge Environment, existing like an information "ether" are more likely to occur when the results of near-term projects are cross-fertilized with the results of projects that address attributes of a Ubiquitous Knowledge Environment. That is, a suite or ecology of projects that address near- and long-term research questions are more likely to produce richer results than a program that fostered one dimension to the exclusion of the other.

A technology-centric vision of information existing like "ether" draws in part on a call for human-centric computing[3], the cornerstone of which is individualized, customized, human-centric information. Realizing this vision is dependent upon long-term advances in the disciplines of information retrieval, storage and communication system technologies, and human-computer interaction that will leverage and harness the methods and mechanisms of artificial intelligence: knowledge representation, language understanding, cognitive processing, and machine learning. The representational model (Figure 1) of the information space is composed of processes that mediate a continuous relationship between the itinerant user and the ubiquitous information store to deliver the right information at the right time in the right format and language, and within the appropriate context and at the right level of complexity and comprehensiveness.

Figure 1: Representational Model



User

- Cognitive Completion
  - > Extension of spell-checking to fact-checking
  - > Task and user context-sensitive
- Do What I Mean (DWIM)
  - > Find what I need
  - > Be aware of what I know
- Collaboration
  - > Identify collegial context
  - > Provide contextual guidance
- Managing Personal Libraries
  - > All that is seen and heard
  - > Personal memory assistant

Interaction

- Query
  - > Question-answering
  - > Natural language processing
  - > Context aware
- Chunking
  - > Cognitive patterns
  - > User and process aggregation
- Massaging Results
  - > Intelligent responses
  - > Summarization and filtering
  - > Multidimensional representations

Store

- Transformations
  - > Formats
  - > Tables, charts, text, …
- Discovery
  - > Novelty, anomalies, the unexpected
- Automatic Capture
  - > "Reading" emerging content
  - > Extracting relevant information
- Interpretive Capture
  - > Beyond rhetoric (e.g., bias detection)
  - > Beyond words (e.g., intent identification)

In this model, information systems serve the user in ways beyond conventional, on-demand response, to proactive provision of as-and-when-needed services (left side of the figure). Information may either be explicitly summoned or implicitly requested. "Cognitive completion" may be viewed as the potential extension of a *spelling corrector* that operates as you type to a *subject advisor.* It correlates a model of the content matter being written about with a model of the maturity of its user and what he/she is already likely to know. Advanced information stores will continuously ingest content of all forms, and process, index, integrate, and summarize it with other like and dissimilar sources contemporaneously (right side of the figure). Active interaction mediators (center of the figure) will interpret the context and semantics of queries and the constraints of available space and time in which to deliver results.

Numerous functions, capabilities, and aids, some of which may be developed in the context of infrastructure or near-term research, will either contribute to or result directly or indirectly from this domain of research. A long sought after “do what I mean” (DWIM) function that interprets what is needed in the context of what is already known will be approachable. The late Michael Dertouzos refers to this class of capability and function as the “ascent to meaning.”[4] The ability to capture electronically all that one sees, reads, and hears, experienced either digitally or live; to index, search, and summarize it on demand; and to customize its delivery for the situation and device at hand, provides a potential lifelong personal memory assistant. Such a vision, dating back to Vannevar Bush’s 1945 Memex machine,[5] may finally be within our grasp.

The ability to ask a machine more than a factual question has long been pursued by cognitive psychologists, computer scientists, and linguists[6] and is now a vision approachable through this focus of research in information science and natural language processing.[7] These systems need to be “context-aware,” knowing and matching both the context of the question and that of the source of the potential answer. Though researching a problem may be novel to an individual, the requirement has common global attributes. If systems were able to capture the problem-solving sequence and correctly anticipate and deliver the same end result as a “chunk,” thereby shortcutting the discovery process, considerable time could be saved and net productivity increased. “Chunking” is an implicit process in most human learning and reasoning and has been implemented in a number of systems.[8]

One of the most significant challenges for using digital information is not finding relevant information, but coping with the very large number of potential results from confederated library searches, particularly when there are many that are relevant but not overlapping. This may be alleviated through filtering and summarization across a variety of media types and source languages at query time.[9] Again, many of the tools that contribute to this vision are already under development, but advancing them beyond relatively simple applications mandates conceiving them in a far larger theoretical context. Thus a search and retrieval system does not merely succeed if it yields relevant results; it will succeed more richly if in building it, we apply lessons from and contribute to related domains in cognitive science, linguistics, and psychology as well as from information and computer science.

Commercial technology will provide very large distributed stores and very inexpensive attached local caches and personal libraries in the terabytes. Combining these advances with anticipated improvements in commodity Internet networking will, for example, enable the commercial digital video library vision of access to “all the movies ever produced, available on demand from anywhere, anytime.” But the storage systems for digital libraries envisioned, rather than passive repositories, will be proactive agents, like the user interfaces discussed earlier that gather and filter information in anticipation of need. These systems will continuously capture emerging content as it is created and generate corresponding metadata to exploit it. There remains the potential, subject to the limits of available technology, to analyze and interpret content as it is ingested, going beyond its rhetoric, words, and imagery to identify its intent and characterize its perspective and bias. On a less complex level, proactive storage systems can be expected to interpret and transform the information contained in graphs, charts, tables, and maps as documents containing them are added. The goal of automatically transforming text from multiple documents into such visual forms may present a greater challenge. Finally, as these systems accrue content, could they eventually perform proactive *discovery*? Data mining applied to finding patterns and detecting anomalies across content and over time and source may be a good starting point.

Research in digital libraries envisions the development and widespread deployment of immersive knowledge environments that are open and accessible to all. Long-held public aspirations, such as facilitating lifelong wellness, providing more equitable education opportunities, developing life skills, and advancing literacies of all forms need not be only aspirations. Research on digital libraries is not the entire solution, but it is part of the solution. People can learn to create, manage, and share personalized knowledge spaces and to link these into other formal and informal information stores. People can learn richer, more natural interactions with information systems through immersion in higher order domains such as concept spaces. Highly curated and well-structured digital libraries can provide a framework for organizing knowledge, and new content can be incorporated via adaptive mechanisms. Coordinated with the Cyberinfrastructure initiative and other initiatives such as those undertaken by the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP), the National Archives and Records Administration (NARA), and NASA, digital libraries not only can provide organization to knowledge, but also can address stewardship, including preservation mechanisms that are self-sustaining and scalable, assuring the preservation of shared content for future use.

We have already begun to see how this vision might work. Undergraduates today can study the Beowulf manuscript, for example, in its Web representation as created by researchers at the University of Kentucky (www.rch.uky.edu) with NSF funding. The scanned version, made using ultraviolet and fiber-optic backlighting, is more readable to the human eye than the fragile manuscript in its present state. Moreover, because of the computational analysis, students can now see the effect on the artifact of generations of conservation and how that has affected interpretation of its content and meaning. Similarly, students use the Space Physics and Aeronomy Research Collaboratory (http://sparc-1.si.umich.edu/sparc/central) at the University of Michigan to see original data from upper-atmosphere sensors and to converse with the scientists collecting and studying this data, enabling them to understand at a relatively young age that science is an investigation of hypotheses, testing, and interpretation. Additional research, on a large scale, is needed to help transform education in the way that digital resources are transforming science along the lines evolving under the leadership of the National Science Digital Library (NSDL).[i]

Research in digital libraries has heretofore drawn largely from the union of computer science and library and information science, including such topics as network and systems design, human-computer interaction, artificial intelligence, information retrieval, information organization, machine translation, database systems, and complexity theory.[10] [11] Future research aims to develop an integrative theory. Institutionally, many parts of NSF and many agencies of the government and beyond have engaged in research and development.[ii] But the core funding for the development of new algorithms comes from CISE, and this funding is critical to sustain an emphasis on research. CISE funding is also critical to ensure U.S. leadership in the development of new systems for digital libraries.

In the next sections, we review development and progress of the Digital Libraries Initiatives, the research issues that are required to develop an effective infrastructure, and then the "grand challenges" that constitute the frontier. We conclude with a brief discussion of workshop participants' experiments in predicting the future and the implications of the workshop's deliberations for the next phase of the program.

[i] NSDL (http://nsdl.org) seeks to build an operating educational digital library to be widely deployed. Its focus is appropriately on education, and even though it includes a small research component, NSDL research will necessarily be tied directly to its near-term objective of delivering an operational capability.

[ii] Digital library research has been funded not only by CISE, but also by other directorates within NSF, including most particularly MPS (e.g., the National Virtual Observatory), EHR (NSDL), GEO, and so on. Project funding has been shared with other government agencies such as DARPA, NASA, NIH, NEH, and others. Still other federal and state government activities (such as the Library of Congress' *National Digital Library* Program http://memory.loc.gov/ammem/dli2/html/lcndlp.html and the University of California's *California Digital Library* www.cdlib.org) contribute, as do private foundations (most notably the Andrew W. Mellon Foundation, the Alfred Sloan Foundation, the Internet Archive, and the Packard Humanities Institute) and industrial research (including many companies, of which IBM is perhaps most visible). Projects have also been funded in cooperation with many international governments, particularly the United Kingdom, Germany, India, and the European Union.

2 The Framework for Digital Libraries Research

Knowledge requires that we can find the information we need when we need it. It remains an elusive goal. But the success of the last decade suggests that we are making substantial progress and that another generation of substantial advances lies within our grasp. In this section, we review the achievements of the first decade of research and the implications of that research for the next 10 years.

2.1 A Decade of Advancing Capability

By the early 1990s, advances in information retrieval, database systems, artificial intelligence, human-computer interaction, geospatial systems, and related disciplines convinced senior program managers in the federal research agencies that much could be learned through an interagency program of integrated, interdisciplinary, project-oriented research. A seminal series of workshops and their respective reports (Box 1) led to the NSF/DARPA/NASA Digital Libraries Initiative, which framed a vision based on an initial set of four goals:

- A digital network of knowledge systems—connecting computing, information, and human resources
- A set of enabling technologies—for creating, distributing, and using knowledge in human-centered multimedia, multimodal environments
- New information services—in networked education, commerce, health care, transportation, government, and others, beyond those provided by traditional libraries and information sources
- Ubiquitous, public, and personal—open 24 hours and network-accessible

> All citizens anywhere anytime can use any Internet-connected digital device to search all of human knowledge. ... In this vision, no classroom, group, or person is ever isolated from the world's greatest knowledge resources.
>
> *—PITAC Digital Library Panel*
>
> *Digital Libraries: Universal Access to Human Knowledge*[12]

Box 1

Digital Libraries Program Planning and Management History Major Community Planning Input

1990	Preliminary Planning Workshops [i]
1995	Information Infrastructure Technology and Applications (IITA) Working Group Workshop on Digital Libraries [ii]
1996	Workshop on Social Aspects of Digital Libraries [iii]
1997	Planning Workshop on Research Agenda for Distributed Knowledge Work Environments [iv]
1998	NSF/EU Working Groups on Digital Libraries Research [Round 1] [v]
2001	President's Information Technology Advisory Committee (PITAC) Report [vi]
2002	NSF/EU Working Groups on Digital Libraries Research [Round 2] [vii]
2003	Blue-Ribbon Advisory Panel on Cyberinfrastructure [viii]

Details of other digital library workshops and meetings can be found at www.dli2.nsf.gov/workshops.html.

[i] Fox, et. al., Digital Library Source Book, http://fox.cs.vt.edu/DLSB.html

[ii] Lynch, C., and Garcia-Molina, H., Interoperability, Scaling, and the Digital Libraries Research Agenda, http://diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html

[iii] Borgman, C., et. al., Social Aspects of Digital Libraries, http://is.gseis.ucla.edu/research/dl/index.html

[iv] Atkins, D., Report of the Santa Fe Planning Workshop on Distributed Knowledge Work Environments: Digital Libraries, www.si.umich.edu/SantaFe

[v] NSF/EU Working Groups Report on Future Developments for Digital Libraries Research, http://si.umich.edu/UMDL/EU_Grant/home.htm

[vi] Digital Libraries: Universal Access to Human Knowledge, www.hpcc.gov/pubs/pitac/pitac-dl-9feb01.pdf

[vii] DELOS/NSF Joint Working Groups on Digital Libraries Research, http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/joint-wgs.html

[viii] Atkins, D., et. al., Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, www.cise.nsf.gov/sci/reports/toc.cfm

The subsequent decade of digital libraries research continued to be guided by related workshops and expert panels to gauge progress and to re-assess research directions. With this ongoing guidance, NSF shaped its digital library research into two U.S. Digital Libraries Initiatives and two International Cooperative Research Initiatives (Box 2). These initiatives extended digital libraries research far beyond the social, behavioral, and institutional contexts of conventional libraries. Digital library projects engaged many partners from many disciplines, from music to geography to the audio recordings of Supreme Court proceedings, as well as embracing and extending advances in the engineering and computer sciences. As observed in the previous chapter, the integrated, increasingly interdisciplinary effort drew broadly on directorates within NSF as well as other federal agencies, state agencies, philanthropic organizations, commercial entities, and international partners.

Box 2

Digital Libraries Initiatives Program Implementation History

NSF 93-141	Research on Digital Libraries [Digital Libraries Initiative—Phase 1] Agency Sponsors: NSF, DARPA, NASA Submission Date: February 4, 1994 Proposals Received: 76 Awards: 6 totaling $25 million (FY94–98)
NSF 98-63	Digital Libraries Initiative–Phase 2 Agency Sponsors: NSF, DARPA, NIH/NLM, NASA, LoC, NEH Agency Partners: IMLS, Smithsonian Institution, NARA Submission Date: July 15, 1998 and May 17, 1999 Proposals Received: ~300 Awards: 34 totaling $48 million (FY98–03)
NSF 99-6	International Digital Libraries Cooperative Research Initiative Submission Dates (target): January 15 of each year 1999, 2000, and 2001 Proposals Received: ~60 Awards: 16 totaling $6 million (FY99–03) Joint Awards: 6 by NSF and JISC; 4 by NSF and DFG
NSF 02-085	International Digital Libraries Cooperative Research Initiative and Applications Submission Dates (target): April 15 of each year 2002, 2003, 2004, … Proposals Received: ~50 Awards: in process

The program has been strikingly successful (Box 3), adapting to rapid advances in technology as well as reaching out to information problems in new domains enabling advances in edge detection, three-dimensional imaging, and simulation, enabling more sophisticated analysis in domains that range from archaeology and paleontology to physiology, while exploring the engineering problems that such investigations expose. Moreover, over the last few years, the most important change in digital libraries has been the inclusion of new kinds of information, users, and researchers. The 20 DLI–2 awards went to PIs from 35 different academic departments, and the program has reached out to international partners, with several U.S. projects coordinated with counterpart projects in the United Kingdom and Germany, as well as with broader international projects involving the European Union and Asian countries.

All of these projects are distinguished by large, multidisciplinary teams of researchers in experimental development of large-scale, engineered systems. The problems they address cannot be done on a small scale, for it is scale and heterogeneity that make them both useful and interesting. These projects present many challenges for digital library research, not the least of which is keeping the existing resources available and usable as test beds for investigators, as well as, increasingly, for larger populations of more generalist users, exemplified by the long-standing relationship between the Alexandria Digital Library research project and the Map and Imagery Library at UCSB. Challenges also arise from discontinuities presented by new technologies and changing national needs, including:

- Sensor systems that automatically flow vast quantities of data into large-scale digital databases;
- The rapid spread of wireless communications, most obvious among young people in Japan and Europe, that introduces ubiquity and geography into all information systems;
- The continuing decline in the price of disk space, now below a dollar per gigabyte (Figure 2);
- The dependence of national security on timely information acquisition, extraction, correlation, summarization, and mining;
- The need to transform teaching and learning so that U.S. students remain competitive in an increasingly global economy; and
- The need to accelerate research in key areas of national needs such as medicine, health, and environmental sciences.

U.S. leadership in the world requires continued support for a broad array of research areas. Those directly related to digital libraries include:

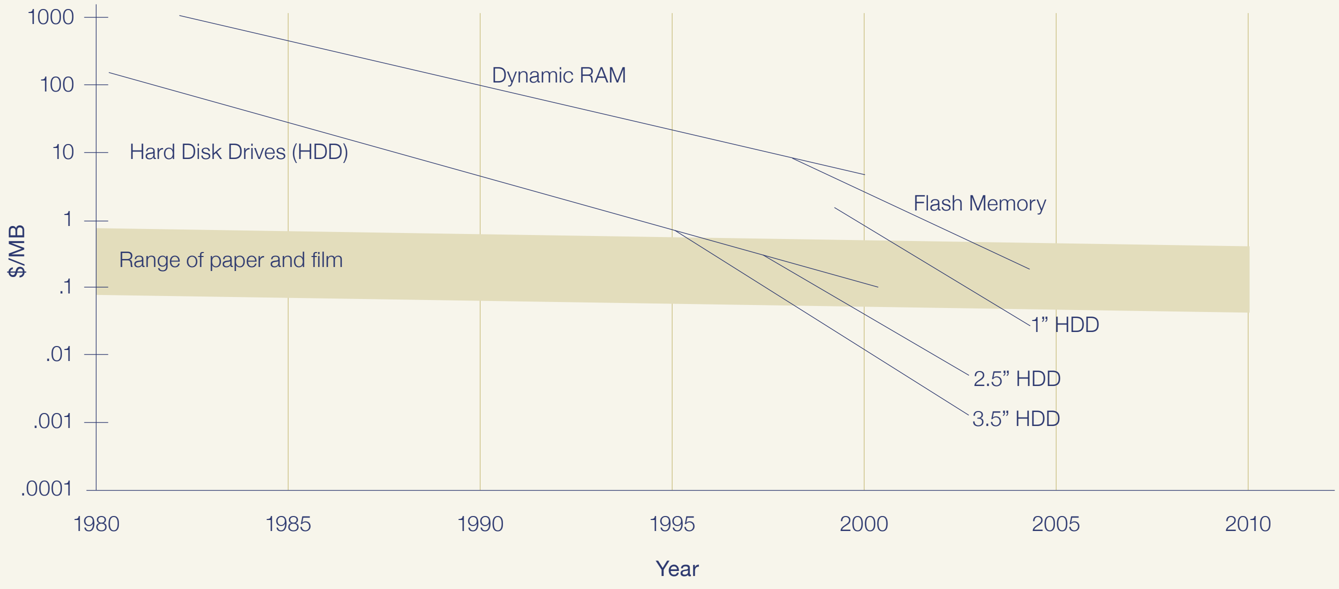
- Learning to absorb and "mine" the vast amounts of data that are coming from sensors. The earth sciences, for example, now have more than a petabyte of sensor-derived data.
- Continuing the expansion in scale, complexity, and diversity of digital resources and the algorithms that can be used with them. For example, 3-D searching is vital to drug design. Analysis of image and motion data to identify people is of value for national security. Medical data mining is important both for immediate patient care and for epidemiological studies. And many media are still resistant to automatic searching; with some irony, software itself stands out as an area still dependent on human categorization.
- Studying users, to understand better how best to employ digital libraries in research and education. Systems need to both be adaptable to user needs and to make the best use of the information users can provide to help others.
- Sustaining the collections and finding economic models that solve the "tragedy of the commons" problem in which many wish to use the world-wide information resources now available, but fewer are willing to pay the cost of maintaining them.

Box 3

- The Google search engine (www.google.com), based upon ideas created and explored in the Stanford University database group using digital library research funding, now does the majority of searches on the Web. One can only imagine the gains that could be achieved through more advanced tools.
- LOCKSS ("Lots of Copies Keep Stuff Safe" http://lockss.stanford.edu) is another research project started at Stanford with an NSF Small Grants for Exploratory Research (SGER) award. It was written up in the *Economist* (June 21, 2003, page TQ-8) and also in a paper accepted in the 2003 ACM Symposium on Operating System Principles (SOSP) conference; the need to build a system to share the responsibility of preserving data led to new insights and methods for peer-to-peer system reliability.
- The National Virtual Observatory (www.us-vo.org/), using the data from the Sloan Digital Sky Survey, enables astronomers to find the results of proposed experiments without having to take new pictures (based on funding from an NSF ITR award and the Alfred Sloan Foundation).
- The protein and genome data banks (www.rcsb.org/pdb) transform molecular biology into a science in which database lookup augments and, in some cases, replaces laboratory experiments, creating an entire field of "computational molecular biology" (using funding from NSF and NIH).
- The Alexandria Digital Library (www.alexandria.ucsb.edu), specializing in location-indexed information, at the University of California at Santa Barbara is used in teaching geography courses at UCLA (again supported by NSF digital library funding).
- Image searching software is able to assign index keywords to pictures, based only on image analysis. David Forsyth, Kobus Barnard and Pinar Duygulu at the University of California at Berkeley have explored and developed this capability with NSF digital library funding (http://elib.cs.berkeley.edu/vision.html).
- Audio searching and storage systems, such as the National Gallery of the Spoken Word (www.ngsw.org) at Michigan State University, the music analysis and searching system at Indiana University, and the Oyez project (Supreme Court oral arguments) at Northwestern University were developed using NSF funding from both the digital library and ITR programs.
- The UC Berkeley Digital Library Project (http://elib.cs.berkeley.edu) demonstrated the potential for digital libraries in real-time crisis management by making significant contributions to California's flood recovery efforts during 1997 in the Russian River Basin.

Figure 2: Average Price of Storage

Source: Ed Grochowski, Hitachi GST



- Extending our interdisciplinary and international collaborations, to gain the greatest efficiencies and advantages by bringing the best resources to each problem. Computer science gains as well as the applications area as new algorithms are invented to cope with new problems that arise in new contexts: for example, the image processing algorithms that were created to help read burned manuscripts, the pattern finding methods invented to help search music, or the speech recognition systems used to build video libraries.
- Leveraging our investment in these digital libraries (DLs) by making them available to other users and other applications, from science to education and commerce.

2.2 The Next Decade of Research in Digital Libraries

The first decade of digital library research provided ample evidence that our ability to generate and collect data exceeds our ability to organize, manage, and effectively use it. This trend is unlikely to abate without continued research and development. For the foreseeable future, data of many types will be increasingly abundant and "technologically" available. But these data will continue to seem chaotic, lacking sufficient organization, stability, and quality control. Moreover, individuals and communities may lose the ability to control access to and manage their own data. The effective use of data and information resources must scale with their ever-increasing abundance and variety; this will take continued research and technology development.

Organizations, individuals, and societies are served, yet challenged, by the explosion of networked information and affordable computational and storage resources. Data, information, and knowledge play an increasingly central role in personal, organizational, and social practices. "Information overload" is one cryptic but popular reaction to these developments. Digital library research focuses directly on this issue. A growing body of evidence suggests the need to include user-centric perspectives at all levels, recognizing that individuals, as well as organizations, now hold very large digital collections of records, and that these collections, via the network, are linked to other information.

It follows that a range of questions emerge about how to structure and manage personal information, from the personal to the societal and transnational, over long periods of time (human lifetimes and beyond) and to integrate this information usefully and respectfully with organizational and "public" information. The processes of authoring, structuring, using, and re-using information and knowledge are becoming increasingly important. The consequences are particularly significant in terms of ensuring future success in the conduct of research, scholarship, understanding, and other areas of human endeavor.

Social communities grow up around, interact with, create, and structure information and knowledge; information comes from many sources and is often contradictory, redundant, or inconsistent. Tools to construct, analyze, model, simulate, and support social communities in conjunction with the information life cycle will always be needed. Issues of trust, reputation, belief, consistency, and uncertainty of information in a distributed digital environment will continue to dominate, where assumptions about underpinnings such as identity and provenance are in question. Complicating the situation immensely, network-based communities also interact with economics, business models, and markets in ways that are not yet well understood.

Finally, there are the entire areas of stewardship, preservation, and curation of data, information, discourse, knowledge, and culture. We need to consider these issues not only in the small but also in the large. The potential importance of stewardship, preservation, and curation as public policy goals is still to be fully appreciated. But some of the relationships among these activities, national security, and the protection of a nation's cultural heritage were illustrated in Iraq following the fall of Baghdad, an instance in which we can imagine that the existence of digital surrogates would have gone far to reduce the informational loss as well as enable identification of pirated objects and their return to the Iraqi people. There are tremendous technical, economic, legal, and political problems here; much progress has been made in mapping these problems, but much less in developing solutions. And again, these issues need to be translated into a personal, user-centric perspective, as well as exploring them within the existing institutional frames. In the DL programs to date, prototypes (primarily large-scale prototypes) have proven to be very valuable. In subsequent DL research programs, work on prototypes will need to be complemented by a new investment in a range of experiments including models and simulations, while also leaving room for long-term research projects.

Solutions for many of these challenges begin with the information itself—understanding, organizing, managing, disseminating, and preserving it. Digital libraries provide these services, and when done best, they are barely noticed.

Technologists describe these types of services as "transparent." But transparent does not mean nonexistent. Indeed, they are achieved only by long-term, focused research and development, but, in the end, they enable users to enrich their lives in ways never before achieved. Progress to date suggests that digital library research should focus on (1) expanding what can be searched; (2) employing "context" in information retrieval at the technical, individual, and societal levels; (3) integrating information spaces into everyday life; (4) reducing data to "actionable" information; and (5) improving productivity through efficient information access. Thus, research in digital libraries embraces a broad range of topics from those traditionally found in computer science and engineering to those that arise in the social and behavioral sciences.

2.2.1 Expand What Can Be Searched

People discover information in many ways. What is typically called "search" is distinguished by three characteristics: it is intentional, it is a monologue rather than a dialogue, and it places the initiative on the recipient. Search is a process pursued by human and machine together, not performed by a machine alone. Machines bring three unique capabilities: speed, scale, and repeatability. Humans also bring three: intention, intelligence, and decision. Without machines, the scope and effectiveness of human search would be severely constrained. The challenge, however, is to envision, create, and assess ways in which machines can support this human activity, leading to retrieval of information that includes traditional and multimedia data (text, audio, images, and video), but extends further to include complex combinations of these forms, as well as data, software, models, and information forms yet to be envisioned.

During the last 10 years there has been enormous progress. Text searching is now used effectively every day by millions, while research is active on searching 2-D images, including extracting text from images of ancient, handwritten text, and on sound recordings, including both music and voice. But it is now known how to convert not just the traditional books, pictures, sounds, and video to digital form, but also 3-D artifacts, including fossils, buildings, and sculptures. Thus, the current research frontier in organizing and searching is in 3-D images and in the combinations of techniques needed to search video.

In contemporary implementations, search in virtually any medium usually starts with text search, or a variant thereof. A typical information seeker with network access will likely begin with Google. But the results of such searches—even well-refined text searches—often result in more responses than the searcher can grasp. Early experiments are under way to engage in searches that rely on different modalities (e.g., searching by shape or color, search by "looks like" or "sounds like").

Despite the undeniable achievements of this research, the volume, complexity, and heterogeneity of new information outpaces even the most advanced of current approaches. Part of the solution may lie in better tools for envisioning information spaces. Research in this area posits that search can be improved by exploring new ways (or old ways in new data environments) to visualize media-rich information. Visualizing quantitative data has long been a fundamental tool for scientific research, rendering what would otherwise be an inconceivably complex set of numbers and relationships into a clearly understandable display of some phenomenon (or, thought of in terms of a familiar contemporary complaint, replacing information overload with an intuitively understandable visualization that captures the essence of a situation). While an appealing metaphor, the problem of visualizing information becomes much more challenging when text, audio, video, and other media forms are added to the information pool.

2.2.2 Use Context for Information Retrieval

Information retrieval attempts to match what can be inferred about information content with what is known about an expressed need. This is fundamentally a qualitative judgment that requires knowledge not only of the topic, but also of how the topic's treatment aligns with a user's need and the context of both user and information. Context has two dimensions: the relationship among information objects and the relationship between information objects and users' needs or desires. The first can be pre-analyzed and is typically represented in metadata. Techniques for content analysis have typically focused on analysis of textual objects and inferences from their contents. But then Google demonstrated the importance of link structure among items for ranking search results. Projects such as Google image search and the CLiMB project at Columbia (www.columbia.edu/cu/cria/climb/) show how context can provide 'aboutness' information relative to an object. There is substantial opportunity for further advancement in the state-of-the-art of automated analysis (and thereby break through a major scalability issue in information management) by understanding how to fully exploit information *context* for dynamic generation of descriptive information traditionally associated with the more static metadata.

Along the lines mapped by Google, link-based search algorithms such as page rank and Hypertext Induced Topic Selection (HITS)[13] attempt to infer quality by recursively quantifying links. Research leading to better understanding the structure underlying links, both in terms of node granularity and edge polarity, may improve these automatic quality assessments. Additionally, important open questions remain about reputation systems that facilitate "reviewing the reviewers." While a number of algorithms exist to assess reputation, they have proven fragile to intentional attacks and need to be improved before they will be a reliable basis for automated quality assessment.

Link analysis and other efforts to describe information items and their contextual relationships form one dimension of context; understanding users and their frameworks is the second. Early efforts to infer user intent aspire to profile users (manually or automatically) and then attempt to adapt to their behaviors. Any user of Amazon will notice the sensitivity of search results to both global user behavior and individual user behavior. Techniques for understanding how to effectively exploit user behavior (without violating user privacy) are still at their initial stages and need to be addressed through focused research, rather than simply hoping for commercial progress.

Information and users come together in libraries, among other places, and libraries have offered a rich societal organizational context in which to examine problems in information retrieval. In particular, libraries consist of *collections* and *users* of information (both managers and consumers). Context may be established in the aggregate of the collection rather than at the finest granularity of an item or a subset of an item (for example, a subset of a database). Research work on automatic inference of these aggregations can be extrapolated to the Web, where it can make an important contribution to our ability to build a more effective knowledge environment. Even the most commonly used information aggregation on the Web—the notion of a "site"—remains poorly understood. Work to automatically discover more abstract aggregations, such as groupings into semantically based collections or equivalence classes has shown some progress but needs considerably more focus before it can be effectively deployed in real information systems.

Finally, the recording of special knowledge in distributed collections requires different technologies than are customary in conventional knowledge management. In particular, federation across collections is necessary for navigation among collections. This requires knowledge representations that are comparable across collections, leading to a new paradigm of analysis across repositories.[14] This new paradigm of cross-analysis is a modern restatement of the classical problem of information retrieval called vocabulary switching, where the difficulty was mapping across subject thesauri, or across document contents.

2.2.3 Integrate Information Spaces into Everyday Life

Even as information becomes ubiquitous and our lives transcend local surroundings, human beings remain rooted by a strong sense of place. Connecting history to place has, historically, been a difficult task. Those who live in long-established communities may have only the dimmest understanding about the generations who lived, struggled, suffered, laughed, walked, and died in the areas through which they now walk each day. We can, however, imagine systems that deliver customized information to individuals as they move through a particular space. A system may alert the visitor that an object of interest (a Greek revival house built in 1847, a statue commemorating a hero of the antislavery movement, the home of a jazz musician) is only a few steps away, provide a survey of the names and occupations of those who lived in a particular building or historic images of their current location, or provide dynamic VRML representations of the location in a particular period.

Such dynamic spatially customized data clearly have broad military, commercial, and personalized applications, but these needs, however acute or potentially lucrative, necessarily have a strong cultural context. While some crucial questions may be relatively easy to answer (e.g., location of geographic entities, identification of customers for a particular product), long-term success will likely depend on much more diffuse information about local practice and knowledge. The ability to successfully link relevant aspects of culture to the information needs and cognitive context of the user could be both intellectually satisfying and economically lucrative. Such a capability remains well beyond our grasp and could well be a "grand challenge."

As digital information becomes more pervasive, the boundaries between physical and cyber spaces, and between public and private spaces, grow less distinct and more permeable. For example, surveillance cameras are already ubiquitous in public spaces and the collection of data created by the cameras becomes part of a "cyberinfrastructure" of which most of the population is only vaguely aware and which is then explicitly accessed under presumably well-controlled conditions. Well-designed user interfaces can help make explicit our passages between these spaces. (User interface is taken here to mean a set of resources and tools that help people to interact with information in these various spaces.)

Unfortunately, the user interfaces we use today assume inordinate uniformity—the same screen, keyboard, and browser are used regardless of whether we are writing, communicating, banking, or using a digital library. In contrast, customized and specialized interfaces arrive with every new personal device (digital camera, GPS, PDA, home entertainment system, programmable thermostat, etc.) as well as scientific device (e.g., MRI, mass spectrometer, sensor array, radio telescope, especially when shared in a collaboratory setting). These examples suggest that we seem destined to have as many ways to interact with digital information as we do with the physical. Integrating the results with other information (e.g., medical records, tables of material properties, maps, telescopic images) adds tremendous value. But in spite of excellent developments in cross-platform languages and user interface design, customized user interfaces for digital libraries appropriate to the variety of users, devices, information types, and objectives present a set of topics that remains largely underexplored.

Clearly, as devices and information proliferate, the demand for the cyberinfrastructure to support diversity among platforms, information, and users intensifies. Consider a small set of interface needs that are required in a permeable space defined by private—public and physical—cyber dimensions. Interfaces are required to coordinate and manage our many devices and services made possible by the cyberinfrastructure. Whether we are moving from scientific data sets in a DL to streams of new data from instruments or from amazon.com to our files in our LAN to the different e-mail and messaging clients on our phones, computers, and intelligent buildings, our preference settings, collaborative filtering profiles, histories and bookmarks should move with us. We might even imagine common personalized interfaces that move with us among various applications.

2.2.4 Reduce Data to Actionable Information

Key to the effective use of ever-growing information is reducing the human overhead required to obtain information from digital libraries and the Web more generally, while retaining the valuable contents, as Nobel Laureate Herbert Simon recognized more than 30 years ago:

> What information consumes is rather obvious; it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.[15]

The total of available attention in the world may well be less than the total available information. Already, it has become easy to talk about billions of online Web pages, and a hidden Web that is yet larger. And yet, because so much potentially valuable information is lacking, many initiatives are funded to put *more* on the Web. A crucial task for digital libraries is hence to support reducing available information to actionable information, i.e., the specific information that will cause a change in behavior, a reduction in further work, or the making of decisions.

Many technologies to enhance the accessibility and utility of information have been investigated in the past. Some are listed in Box 4, generally in an order from the easier to the harder and more speculative tasks. Several are used routinely today.

Many of the tasks itemized in Box 4 may be handled semi-automatically, i.e., with human interaction, before full automation can be achieved. But semi-automatic systems should have the capability to learn from those interventions, to reduce the human load over time and to accommodate the wealth of new information. All these tasks can be expanded by adding adjectives as distributed, multimedia, or ubiquitous, but those won't change the scientific import greatly.

Box 4

Enhancing Information Accessibility and Utility

1. Rank by document contents
   Assumes the consumer will consider only a few documents on the top of list.
2. Rank by authority
   Gives preference to documents valued in a particular context
   (e.g., a journal versus a workshop report).
3. Rank by reference authority
   Google's page ranking algorithm, for example, extracts communal knowledge
   as evidenced by references given.
4. Eliminate redundancy
   Similar retrieved documents are filtered, either by currency or relevance.
5. Determine differences among documents
   May be as simple as looking for additional material in a new version
   or as complex as requiring deep analysis.
6. Identify novelty of a new document
   Maximum marginal relevance is an example of ranking by novelty.
7. Determine novelty with respect to an individual
   Domain emphasis is needed in order to be feasible.
8. Abstract textual documents to retain essentials
   Selecting sentences that appear to represent the content; better abstractions result
   for domain-specific texts.
9. Abstract the content of document collections
   Integrate and semantically compare sources, using ontologies.
10. Data-mine
    Link data-mining results with information from textual sources
    to strengthen explanatory capabilities.
11. Reduce textual information to a visual presentation
    Place an abstraction into an appropriate temporal or spatial model.
12. Populate analytic model
    Use an analytic model not only to discern novelty, but also to represent normal behavior.
13. Support predictive tasks
    Information systems not only analyze the past, but also extrapolate to possible futures.
14. Discover abnormal situations
    Apply pre-existing model to identify and isolate unusual or abnormal behavior.

Box 5

Cost/Benefit and the Consideration of Errors

Assessment of costs and benefits of alternative technologies requires characterization of the setting, or context. In some settings the cost of missing a source entry (a “Type 1” error) is high; in other settings the cost of having to reject irrelevant entries (a “Type 2” error) is high. For instance, the cost of missing a terrorist is indeed high, but many schemes now being considered fail because technologies that have a low rate of Type 1 errors are typically associated with an excessive rate of Type 2 errors, so that even at a low cost per error rejection, the cost/benefit ratio will not be acceptable.

To support Web-based businesses, as envisaged in the semantic net, a very low rate of Type 2 errors (false hits) will be needed. Businesses already routinely prequalify suppliers. The potential cost of getting the wrong material, geting it late, or obtaining the wrong information about material is much higher than the benefits of “getting a good deal.” Saving 5% on supplies may be negligible when considering the cost of filtering misleading deals. For Web-based businesses, a very low rate of Type 2 errors will be essential for business automation, as envisaged for the semantic web. [Tim Berners-Lee, Jim Hendler, and Ora Lassila: “The Semantic Web”, Scientific American May 2001]

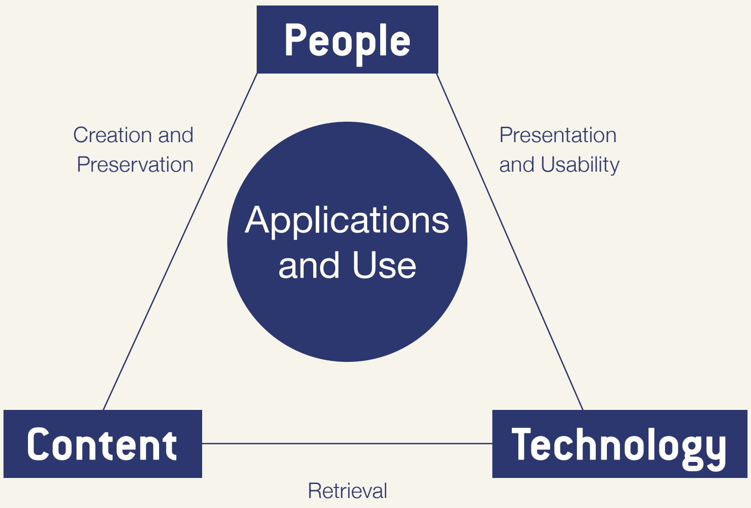
Assignment of costs to these two types of errors may also depend on one’s generation or back ground. Senior people, having grown up in an information-poor setting will want to get all the information. It is often the generation in the trenches that realizes that there is already too much to which one can devote attention.

2.2.5 Improve Productivity through Information Access

Ultimately, digital libraries will offer unparalleled access to information for a far broader range of users than existing physical and organizational structures. But even after a decade of sizable investment in digital libraries by the National Science Foundation (NSF) and other federal agencies in which considerable progress has been made, clearly, we are a long way from realizing significant productivity gains. Yet a mounting body of evidence of partial gains in a broad range of disciplines has intensified the belief that such achievements are possible. As the writers of *The Report of the DELOS–NSF Working Group on Digital Imagery for Significant Cultural and Historical Materials* observed:

> Our interdisciplinary research will develop **technologies** to enhance the way **people** create and access the **content** … People encompass all users, from curators and library and information scientists, to scholars, teachers, and students in all areas of the humanities, to citizens of all cultures. Content is the vast array of significant … materials throughout the world. Technologies are the enabling research and development in all related technical areas such as information retrieval, image processing, artificial intelligence, and data mining. [16]
>
> The conclusions of [the NSF–DELOS] workshop are fully consistent with the NSF–DELOS recommendations for focused, interdisciplinary research programs along the three edges and the center of the triangle [Figure 3], areas that traditional research programs currently neglect. The research area between people and content is the area of digital imagery **creation and preservation**. The area between content and technologies is the efficient and effective retrieval of the content using technologies. Research into **presentation and usability** will enhance the ability to access the content. Effective **applications and use** of the research results, under lifecycle management, will integrate research of the three related areas.

Figure 3: Interdisciplinary Digital Library Research



3 Sustaining Key Digital Resources

New digital information resources pose issues different and distinct from those of conventional resources, and, hence, require specific attention. Progress toward a future built on digital knowledge will require a commitment to not only establishing but also sustaining these critical resources over an extended period. Given the geometric growth rate of digital information from a wide array of sources and the growing population of users of such information, efficient means of acquisition, access, usage, stewardship, and management will pose continuing challenges for the foreseeable future. Whereas this report endorses the research recommendations of both the PITAC and the Blue Ribbon Advisory Panel on Cyberinfrastructure (advocating $20 million annually for advanced research in digital libraries, discussed subsequently in Section 4), we also stress the need for continued investment in the underlying infrastructure at a level of $40 million annually in order to sustain the growing inventory of resources and capabilities into the future and to nurture the research agenda. A fundamental lesson of digital libraries research is that advanced research and a rich information infrastructure are both mutually supportive and mutually dependent.

The infrastructure recommended includes a near-term research program with five major components: acquisition, access, usage, stewardship, and management. They apply to all five elements of the framework set forth in Section 2. This section lays out the basic issues and requirements for each of these topics.

3.1 Acquisition

Many issues in acquiring information are bound up in metadata. Advances in metadata, its generation, use, and management, are required to sustain digital library progress. Content must be usable and readily re-usable by multiple audiences. Studies in these areas will also contribute directly to long-term research questions in metadata, ontologies, knowledge organization, and information management—issues that crop up with respect to sustainability and management as well as to the “grand challenge” questions set forth in Section 4.

Better understanding of methods to translate metadata between schemas is needed. Current strategies for manual translation founder in a morass of detail because they aim for perfection. However, an imperfect system may still be useful, particularly when there is no practical alternative, like when the volume of data is very large or when one is attempting to federate a set of collections. Various statistical methods should be studied to build lexicon-like structures that can automatically align different types of metadata.

Another promising direction is to develop methods for obtaining metadata that are absent from an existing digital object. The current approach—in which skilled personnel attach metadata to objects—is infeasible for large collections. The first natural strategy to explore involves using existing, labeled collections to build classifiers that can attach tags to unlabelled collections. This approach is viable when there are many large collections that are partially labeled, sufficient that it is practical to learn and evaluate classifiers for many kinds of metadata tags. This class of strategies is likely to be successful only for metadata that can be inferred from the object itself. For example, it may be possible to determine the names of those present from a picture, but it is likely to be impossible to determine the time and date at which the picture was taken. One can deal with this difficulty by attempting to compel creators or editors of objects to attach metadata at the time of creation or editing, but in practice this very likely will not succeed. An alternative is lightweight metadata collection. Objects would become “sticky”; when an object is created, edited, downloaded, read, etc., it would opportunistically collect various forms of information that might be helpful in building classifiers later. This is a generalization of the current approach where digital cameras insert a date and time into picture header files; in the not-too-distant future, one might expect to find GPS information there, too.

The likelihood of a single approach to creating metadata is decreasing, largely due to the success of multiple methodologies aligned to specific contexts. Though nearly all of these seek to reduce human effort, whether exerted by expert catalogers or others, the approaches run the gamut from complete automation to a variety of leveraging strategies. Further increasing this diversity, some systems make very effective use of strongly typed metadata fields (such as for geospatially indexed resources) while other successful efforts (such as systems built with OAI–PMH, www.openarchives.org/OAI/openarchivesprotocol.html, or based entirely on content analysis) employ few data types beyond text. Hence an important but open question is how to fulfill the bibliographic functions associated with acquisition in contexts where the joining of multiple libraries and collections yields a mixture[18] of metadata approaches. Progress on this challenge must be coordinated with efforts to advance numerous schemas, vocabularies, and ontologies.

Wide deployment of an overarching model that embraces multiple forms of metadata seems essential. This would serve as a key component in a common technical fabric for linking independent digital libraries and creating coherence on national and international scales. At the heart of the problem is the need to organize information for diverse purposes such as management, discovery, and browsing. Strategies wholly reliant on human-generated metadata will not scale to the digital libraries envisioned, suggesting the need to accelerate progress on alternative approaches for automatic metadata generation and collection aggregation.

3.2 Access

Data and information captured and represented in various digital formats are proliferating rapidly. However, the techniques for accessing these materials remain rudimentary and imprecise, based largely on simple keyword indexes, relational queries, and low-level image or audio features.

In considering how better to approach issues of information access, it is instructive to note that cognitive psychologists distinguish between two fundamental types of memory: recognition and recall. Recognition occurs when you see something familiar, while recall requires that you remember something and are able to articulate it. Most information retrieval depends upon recall skills—the user has to describe what he/she wishes to retrieve, formulate a query, and submit a search. Browsing, another common strategy users employ to find information, depends more on recognition skills—looking around until you find something of interest that you recognize as useful. But most browsing still requires that the user describe a starting point. Recall approaches are effective with text-based systems because words can be spelled and matched against a corpus of documents. Describing images and sounds is vastly more difficult, both for the indexer and the retriever. Recall also depends on the availability of rich metadata or on sufficient amounts of matching text.

In the future, recognition approaches are likely to be more effective in digital libraries for at least two reasons: (1) the proliferation of nontextual digital documents (e.g., still and moving images, sound) and (2) the lack of metadata on which to base recall algorithms. We need ways to summarize nontextual data so that people can recognize relevance easily, such as the video "fast forward" experiments reported at the Joint Conference on Digital Libraries (JCDL) 2003.[17] Metadata is expensive to produce and cannot serve all of the uses of any given document. Individuals are unlikely to invest the effort in rich description of everything they add to a digital library. They need ways to summarize and browse repositories quickly and easily. Research in the past decade has shown that children can learn to use a recognition-based catalog of science materials quickly and easily.[18] In trying to explain the difficulty in describing images they seek, users are quick to note how they rely on serendipity and on "knowing it when I see it," i.e., recognition, not recall.

Information visualization is a recognition technique of proven value for quantitative data that becomes more challenging when we add text, audio, and video content. A major objective is to explore new ways (or old ways in new data environments) to visualize media-rich information. We perceive much through abstractions. We know that information visualization (in the full multimedia sense) can be effectively employed to summarize content and provide the means to display it in new and novel ways. Imagine an information room in which users can specify an initial domain (e.g., American domestic politics, 1964) and "walk through" the information space. In one domain there will be audio, in another domain, video. Perhaps users are interested in particular individuals or events. The space could be reconfigured to provide access to all public data associated with them. Through a process of sampling and refining subsequent searches, users can locate what they seek and, perhaps, discover additional information that bears on their interests as seen through the prism of the visualized or sensed world.

Content-based relevance feedback (CBRF) extends already proven relevance feedback (RF) to include relevance judgments based on features extracted from the content of multimedia objects, such as colors in images, melodies in music, structures in architectural graphics, and choreography in dance. RF techniques have a long history of successful deployment in text-based environments.[19] Popular search engines now routinely include RF; Google, for example, includes a "similar pages" retrieval option. CBRF tools, in contrast, are necessary[20] in multimedia DLs because general users do not typically possess the domain-specific (usually text-based) vocabularies to express their multimedia information needs. In the domains of image (primarily photographic) and video retrieval, significant CBRF advances are already being made.[21] [22] [23] [24]

Disambiguation poses major challenges: what set of features, components, or facets of an object will the user deem to be relevant? In music, is it the tempo, the melodic line, the orchestration, the lyrics, or the rhythm of a given work upon which the user is basing the relevance assessment? Once an object has been disambiguated, CBRF interfaces can enable users to see and hear the constituent components of the object(s) of interest. CBRF extends beyond the intrinsic features found *within* each media type to include extrinsic similarities across media types. For example, three pieces of music might have no melodic sequences, no rhythms, no harmonies nor lyrics in common, but still might be deemed to be in some sense “similar” and “relevant” if all three had “similar” dance gestures associated with them.

In the grander scheme of things, it is inappropriate and limiting to conceive of knowledge as exclusively scientific and of digital libraries as mere repositories of scientific data. The disciplines whose expertise encompasses language and literature, history and philosophy and religion, film, art and music, cultural studies, psychology and political science and sociology, anthropology, geography, architecture and archeology have much to contribute to analyzing issues of national priority, such as homeland security, and much to contribute to developing systems to support abstract, open-ended investigation and analysis. Because they work closely with the relevant cultural and historical sources, scholars in the humanities understand the vagaries of language subsumed into the technical term “disambiguation” of both text and images and are familiar with strategies based on context, experience, and structure for automatically disambiguating the source. Moreover in specialized contexts, metadata that incorporate translation, interpretation, analysis, and criticism—the digital library equivalents of the books and articles written about primary sources in traditional libraries—enhance and extend the use of material.

The important difference is that, if provided with proactive interfaces, scholars can continually enhance the primary data of digital libraries with searchable metadata of essential secondary information. This explanatory information will forever be linked to the primary digital sources and always be available for accurate, comprehensive, instantaneous searching. While applicable to all digital resources, these are most obviously true for the millions of historic, heterogeneous hand-written materials on stone, clay, cloth, wood, canvas, papyrus, animal skin, paper, walls, or any other medium carrying text of any kind. A transcription of the text on these objects is metadata. Without such metadata, the data are in many cases (e.g., ancient texts, metaphorical language, unfamiliar scripts, and foreign languages) virtually meaningless. Now confined to the relatively specialized venues of museums and galleries and long-gone civilizations, the challenge is to enable understanding of contemporary objects and expression so that radically different contemporary cultures may be made equally accessible to modern analysts.

3.3 Usage

Digital libraries focus on leveraging knowledge resources. The rate and the breadth of knowledge advancement presently is less than what could be achieved and sustained with improved infrastructure in five key areas: bibliographic systems; cognition-leveraging tools; social factors; information architectures; and engineering, operations, and evaluation. The proposed work has precedents on which one may project a high likelihood of success. Indeed, important results have been achieved in each of the five areas as the following summaries of current work suggest:

3.3.1 Bibliographic Systems

Current practices of knowledge organization, for the general public and for scholars or others with special interests, have been developed and refined over a 150-year period. Four decades of this has incorporated computer technologies, including 10 years of NSF-supported research on digital libraries, some of which explicitly targeted bibliographic practices.

Despite this rich history, technologies for creating, exchanging, managing, and presenting information have outstripped the capacities of bibliographic systems to 1) paint comprehensive views of the knowledge landscape or 2) fully address the needs of the growing number of people who now expect to navigate this landscape without the expert assistance of librarians.

Rather than a criticism, this is an observation of the challenges wrought when profoundly difficult changes occur with astounding speed. The basic functions of a bibliographic system are *finding, collocating, choosing, acquisition, and navigation*[25], and each is being affected by rapid technological change. Some key challenges to be addressed are discussed below:

Dynamic Content and Atomicity

Resources that are commonplace in the computer age do not necessarily come packaged neatly as indivisible (atomic) units, and some of the most expressive media, such as real-time observational data, are highly dynamic. In *The Intellectual Foundations of Information Organization*, Svenonius points out: "Documents with uncertain boundaries, which are ongoing, continually growing, or replacing parts of themselves, have identity problems. It is not possible to maintain identity through flux ('One cannot step twice into the same river' [referencing Heraclitus]). … A snapshot cannot accurately describe information that is dynamic. This is not simply a philosophical matter, since what is difficult to identify is difficult to describe and therefore difficult to organize."

Unfortunately for bibliographers, such documents are an increasingly important aspect of the knowledge landscape, especially in science and engineering. Some progress has been achieved in respect to this problem, but an important and achievable goal should be to gain broad acceptance and use of an *operational* definition for uniquely identified digital entities across most or all NSF-sponsored cyberinfrastructure. The associated identifiers would foster interoperability and form crucial foundations for all types of bibliographic systems.

Universality versus Specialized Expressiveness

A better understanding is needed—from the perspectives of users as well as of content and metadata providers—of the tradeoffs between universality and specialization. Every bibliographic system designer faces difficult choices between maximizing compatibility with other systems and maximizing expressiveness, relative to the needs of a target audience, often comprising specialists in particular fields. Solid data in these matters would greatly improve both planning and effectiveness for a vast array of information- and knowledge-handling systems.

Scalability of Bibliographic Systems

Scalability of effective services remains a problem. The problem derives from new user expectations (driven in part by their experiences with Google and its kin) and from increased creation rates of materials to be included in science-related bibliographic systems. Understanding and exploiting user contexts and information semantics is essential to keep the act of searching effective as the breadth and volume of available materials increase.

A tempting solution might be to rely entirely on Google or the like, but studies[26] indicate that users expect levels of bibliographic functionality that are beyond what Google or other current systems can deliver, at least for now. NSF could chart a course for bibliographic systems to address the foregoing challenges and simultaneously to exploit the scalability of automated approaches to information organization and discovery.

3.3.2 Cognition-Leveraging Tools

Fundamentally, all technologies leverage human capabilities, but computing and networking are especially rich in the forms of augmentation they offer. Of these forms, the leveraging of cognition is one of the most important, enhancing the abilities of humans to understand one another, to gain new perspectives on the universe they occupy, and, most fundamentally, to learn.

Among the best outcomes of the NSF's Digital Libraries Initiatives have been demonstrations of such impacts (ADEPT, http://eil.bren.ucsb.edu/projects/proj-adept.htm), but the affected audiences remain relatively small, and the full potential appears far greater than what has been achieved to date. Further, learning and cognition represent areas where research and education meet, so emphasis on these matters will extend the impact to encompass NSF educational goals and to affect very large audiences, including a significant fraction of the American workforce. Hence digital library outcomes should include new and improved tools for collaboration and for working (interactively) with all artifacts of scientific progress, including: observed and simulated data; taxonomies; mathematical expressions; molecular, chemical, and genomic expressions; structural, physical, and computational models; tables, graphs, charts, maps, and images; field and laboratory notebooks; monographs and other scholarly documents; critical reviews and discourse; ontologies; and bibliographic references to scholarly literature. Implicit in the need for tools is the need for widely agreed, nonproprietary, digital representations for such artifacts.

Functions addressed by cognition-leveraging tools might include: spontaneous online meetings, collaborative editing (of rich scientific documents), bibliography sharing, curriculum architecting, semantic tagging, knowledge mapping, visualization sharing, (large) data-set structuring, and creating (shared or personal) logs or diaries of experiments and studies.

Such tools, embedded within digital libraries and elsewhere, will help ensure that the nation's cyberinfrastructure is an active, not a passive, place for learning by students as well as researchers. This in turn would promote educational enhancements based on inquiry, constructivism, and group learning.[27]

3.3.3 Social Factors

To gain alignment with key social and institutional needs, digital libraries should explicitly support the formation of communities of practice, built initially by change agents and early adopters of digital library systems. Properly chosen and supported, such communities will increase the impact on both specialized and general-purpose domains, fostering improved science accessibility for all citizens.

These communities are critical to another aspect of digital libraries: selective, intelligent, targeted collection development.[28] Though no single individual or group can hope to perform this function for the entire community of users, the problem is tractable within smaller communities of practice, where domains of interest are more limited. Stated another way, smaller communities will create opportunities for excellence where global approaches are likely to yield mediocrity. Such communities can help determine which entities most deserve resources to assure their long-term preservation. This is a well-established responsibility of libraries. In a related matter, digital libraries should seek a broad geographic presence, with a commitment to embracing a very large number of educational institutions, including those that traditionally have not received significant NSF support. Wide geographic distribution would enable emerging virtual communities to be seeded and strengthened by face-to-face interactions. This idea is informed by the success of the regional networks used to deploy the NSFnet, and it may be advantageously linked to the nascent Institutional Repositories movement.[29]

Finally, the NSF should consider anthropological studies, documenting the emergence of social norms and communities of practice around digital libraries and other aspects of cyberinfrastructure.[30]

3.3.4 Information Architectures

Though progress has been achieved in important areas such as OAI–PMH and DC metadata (http://dublincore.org), a common architecture for digital libraries remains elusive. Yet to be realized fully on a large scale is the goal articulated by Besser for digital libraries: to deliver information to multiple clienteles, using the same collection to serve many different groups of users, each with its own level of knowledge and modality of learning and interacting.[31] The obstacles have technical and political dimensions, and NSF has the potential to address both.

Technically, digital libraries should include one or more frameworks explicitly designed to support data mining within selected segments of the Internet, utilizing content, metadata, and an increasingly rich array of contextual information, such as links, citations, and usage data by audience-type; this might be realized by developing and operating one or more large-scale data warehouses that place particular emphasis—beyond characterizing individual entities and proxies—on relationships among these entities, relationships that are themselves determined by numerous diffuse and diverse processes, with and without human mediation. Underpinnings for such warehousing—which need both advancement and widespread deployment—include unique-identifiers, nationwide authentication of users (with anonymity protections), high-level semantic markup and associated registries, and representations to deal effectively with programmatic services (and with entities that may be accessed only via such services).

To ensure success, the NSF must deal with the reality that some degree of standards enforcement may be required. It may be wise to make grantees' awards contingent upon adherence to some (small) set of protocol and interface standards, such as the warehouse underpinnings described above. Explicit linking to the emerging NSDL core integration system may be part of an appropriate level of standardization. To ensure scalability, dependability, and persistence of resources, the architecture also must enable distributed replication and synchronization of entities, including services.

3.3.5 Engineering, Operations, and Evaluation

Good software engineering, software support, and reliable computer operations are difficult and expensive. They also are important, as they determine the end-user experience, but such characteristics often are given little support in NSF grants (because achieving them is not considered research). The systems that are developed by the NSF investment, however, will have to serve very large audiences dependably, on multiple platforms, and they must be well-supported in response to user problems and needs for incremental enhancement. This suggests ample funding for organizations that are committed to both engineering excellence and software support, focused on needed tools and representations. Explicit attention to contemporary approaches to software engineering and support, especially user-centered design and development, should be part of digital libraries development.

An important means for gaining an appropriate emphasis on customer service, dependability, and end-user effectiveness is to foster a culture of evaluation. Projects should include both summative and formative evaluation efforts. A difficult but worthy objective is to embed, within digital libraries and related projects, instruments for observing the learning process in relation to resources employed.[32] Doing so would also help the Cyberinfrastructure initiative embrace large-scale education enhancement as a first-class goal. An understanding of digital libraries usage can yield a common philosophical and technical fabric for coordinating and using independent digital libraries.[33]

3.4 Stewardship

Stewardship over digital materials presents new and daunting challenges to which conventional preservation approaches do not apply. There is little value in preserving an 8-inch floppy disk, for example, unless a working disk reader is also preserved, and a computer that can accommodate the reader, and the software to interpret the data. But the value of the disk is more likely to reside in the data it holds than in the physical mechanism of its storage (except, of course, for the museum that collects old computing equipment).

Digital information is fragile in the short term and difficult to preserve in the long term. It is constantly at risk, particularly as computers become increasingly networked with large numbers of software components. This presents an enormous vulnerability, posing both major security and stewardship issues (although it is more frequently considered a security threat). Denial-of-service attacks provide the most spectacular examples of risk to date, but a more subtle and equally dangerous risk is data corruption. Effective stewardship must include policies and practices to ensure the integrity of information.

Institutions, particularly universities, libraries, and museums, pride themselves on their special resources, capabilities, and collections. These shape their identity and differentiate them among their peers. They are also typically located in one place. Digital technology makes them more accessible, but it is not yet being widely used to mitigate risk to these resources. Very few have found the resources to digitize precious materials. Even materials that are digitized are rarely replicated effectively. More active replication is necessary, not only for preservation purposes, but also for risk avoidance.

Geographically distributed digital repositories are already available that mirror information relatively automatically. These could be employed for the replication of digitized cultural resources. While simple distributed storage is not enough for preservation, it is a start. Preservation strategies also need to replicate control, for centralized control over distributed collections also introduces risk of loss (from bad decisions) that may not be recoverable. A robust preservation strategy will account not only for accidents and maliciousness, but also for well-intentioned but ill-conceived decisions.

On a local level, a tremendous amount of digital content is being created by individuals. Personal libraries of significant size are beginning to emerge, and new software systems will be needed to preserve these materials. Much of community history will be found here. These materials will be of significant interest to historians and cultural researchers of the future.

Affordable and reliable digital preservation requires new tools and technologies. Digital preservation will not scale without tools and technologies that automate many aspects of the preservation process and that support human decision making. Decision models are needed to support selection, choice of preservation strategies (normalization, migration, emulation), and the costs and benefits of various levels of description and metadata. Digital preservation strategies are “metadata intensive.” Therefore there is a critical need to develop tools that automatically supply core metadata, extract metadata from resources at ingest, and restructure and manage metadata over time.

It is important to recognize that metadata, schemas, and ontologies are dynamic—subject to frequent extension and revision. It will be essential for future users of archived materials to recover and relate the metadata schema used when the entity was created. Managing schema evolution is a major issue. Likewise, managing the identity of preserved digital objects over time is a challenge for digital archives because the identifiers assigned to digital objects can be changed easily and the technologies for naming and tracking digital objects evolve over time. Issues in the area of naming and authorization include development of methods for unique and persistent naming of archived digital objects, tools for certification and authentication of preserved digital objects, methods for version control, and interoperability among naming mechanisms used by different content providers. Tools are also needed to automatically transform preserved digital objects from obsolescing to contemporary formats, standards, and data models and to document the effects of these transformations.

3.5 Management

As NSF and the scientific communities contemplate the creation of a cyberinfrastructure enhanced by digital libraries, the need for systematic attention to questions of organizational design appears in at least three areas. First, the research agendas for the digital libraries and related programs have regularly made assumptions about the organization of various social factors. The RIACS report on information management, for example, characterizes use, privacy, security, and usability as “attributes of interoperation,” interpreting these factors largely as independent variables that define technology requirements.[34] These key social factors, however, are themselves highly variable, depending at least in part on the type, purpose, and structure of the organization in which information and technology users are embedded. Because so much technical research in this field depends logically on assumptions about the social organization of information, the validity of the research—and its potential for further development and implementation—depends on much more careful investigation than has been achieved to date of these organizational assumptions, including those about business models and modes of operation.

Second, the report on cyberinfrastructure by Atkins, et. al., wisely distinguishes among the processes of research, development, and operations as essential for promoting and sustaining the software and hardware products of an Advanced Cyberinfrastructure Program (ACP). These three broad classes of activity, of course, are interrelated and ideally feed back options and requirements to one another. Such feedback would be an essential component of the ongoing vitality of the program, but it is not sufficient. To administer and help sustain an ACP, the report further recommends both an internal organization within NSF and a community-based structure of centers. These centers would foster development activities as well as user support and other operations at both generic and disciplinary levels. Although the recommendations for organizational change at NSF are quite specific, the design of the community-based centers is left unspecified and will presumably vary, in part, with the needs of the academic communities they are intended to serve. However, if the centers are to help sustain and support the products of an ACP, they must be designed in such a way that they operate in a businesslike fashion and can sustain themselves, eventually independent of NSF support. Such design should not fail to be informed by current expert understandings and additional targeted research that would indicate how certain types of mission, leadership, governance, organizational structure, legal arrangements for intellectual property, and financing, especially in the context of public goods economics, could contribute to—or undermine—the success of such centers.

Third, many digital library and related research projects depend on the use or creation of substantial databases of content from one or more subject domains. Moreover, subject-based research within specific academic disciplines also often yields significant collections of data and other content that may themselves be valuable outcomes of the funded research, and essential ingredients of an emerging cyberinfrastructure. However, such databases are often created only with narrow research uses in mind, and efforts to move them into more broadly based, self-sustaining, operational use are often stymied.

A highly fertile area for support to the curation and communication of scholarly information in the sciences and other disciplines falls under the broad topic of organizational design. Electronic resources do not need to be managed within existing organizational structures, but to persist they must be managed within some organizational context. On the one hand, with investment in technology, barriers to entry for the creation and management of digital resources can be lower than they are when the storage of physical items requires large capital investments in physical objects and buildings to house them, but small institutions that want to develop, provide, and manage electronic resources often lack the sophisticated curatorial, legal, financial, and other organizational skills that are necessary. On the other hand, the huge economies of scale that are possible with digital databases are difficult to manage over current institutional boundaries. Clearly, new organizations and organizational models are needed that are sensitive to the dynamics of particular scientific communities, driven by academic mission, and able to sustain themselves over time as integral parts of the broader cyberinfrastructure. To foster the development of appropriate organizations and organizational models, NSF should institute the following programs and features as part of the digital libraries agenda to support the broader Advanced Cyberinfrastructure Program (ACP):

There would be two broad objectives. The first is to identify the organizational variation within an academic community or set of academic communities that would affect the requirements and parameters for research, development, and operation of new technology funded as part of the ACP. A second objective would be to take various scenarios of research, development, and operation, and explore the advantages and disadvantages for the emerging cyberinfrastructure of different mixes of organizational features.

The work on organizational design should focus on the following organizational variables: types of mission such as commercial and nonprofit; types of governance, including membership, board, and partnership models; leadership qualities; structural dimensions, such as size; policy issues, such as privacy, security, and risk management approaches to the ownership and use of intellectual property; and financing options, taking into account the importance of common or public good economics for the emerging cyberinfrastructure. Funded work could employ a mix of empirical case studies and theoretical approaches, and it could be embedded as part of a larger project or conducted as a stand-alone initiative.

In order to create a sustainable cyberinfrastructure, the new centers and content-management organizations should have access to a highly specialized organization—or set of organizations—that can provide expert advice on questions of mission, leadership, governance, and general business practices so that when created, new organizations created or struggling to survive in the cyberinfrastructure have a reasonable chance of operating in a businesslike fashion. The supporting organization(s) must not operate in a “cookie-cutter” fashion, but must be sensitive to variations in need among academic communities, as well as to differences in size and trajectories of growth. Ideally, to economize on the costly duplication of services, the supporting organization(s) might also take direct responsibility for providing a set of common services, such as accounting, human resources, board governance, and legal advice, thereby helping to create a family (or families) of efficiently run organizations.

4 Emerging Research Opportunities and the 'Grand Challenge' Problems

The first 10 years of research into digital libraries have been notably successful, but may be even more noteworthy for clarifying what remains beyond reach. The continuous challenge, however, is to understand how digital libraries research interacts with and enhances other sciences. We envision a Ubiquitous Knowledge Environment. In the previous section, we laid out a well-defined agenda for near-term research that will reinforce and advance the information infrastructure that a decade of research has already engendered. However, the successful realization and continued advancement of the next-generation information infrastructure must be supported by scientific, social, and technological research that will yield new generations of knowledge environments that evolve in pace with advances in the computing and communications infrastructure. In this section, we examine the requirements of the long-term advanced research agenda and the grand challenge problems in information that still elude us.

4.1 Basic Themes in Long-Term Research

The long-term research agenda outlined here presumes a future in which information is pervasive, shaping the way individuals live and societies function. The challenge is to provide the means for people and institutions to understand and use information to their and society's advantage. It is developed from the notion of a ubiquitous information environment in which information envelopes users who understand—or can learn to understand at point of use—how to access the resources they need. This might range from programming home climate controls to conserve energy to conducting large-scale advanced experiments in a laboratory. The same family of technologies can and should support the entire range of possible uses.

This information-rich environment requires research in the following broad themes, which iterate and expand upon many of the topics previously described.

Understanding information and its uses. Within this topic fall questions concerning heterogeneity of information systems, sources, and users; seemingly unbounded scale of data and users; incomplete and uncertain information; comprehensive metadata allowing automatic interoperation of diverse resources; and semantics of and correlative relationships among data.

Appropriate stewardship over information. Although preservation and management are near-term needs, preservation and records management over the long term will remain a research problem that will continue to evolve as the volume and heterogeneity of digital information increases and the expectations of users grow.

Fitting technology-enabled opportunities into the social fabric. A number of vexing issues are already apparent in the evolving legal and social frameworks. Privacy, confidentiality, and intellectual property rights are three obvious examples. But others will surface as more and more users and organizations of varying organizational capability, behavior, and tradition become embedded in the digital environment.

Matching system capabilities to user needs. As data become voluminous beyond current capacities, real-time operation and collaborative, synchronous processes become pressing needs. These issues are already apparent in scientists' interactions with immense data sets such as astronomical or geospatial data where the power of the new science lies in the ability to integrate across multiple heterogeneous data sets but the performance of the system has not yet scaled. If the goal is to enable ubiquitous access to equally large or larger data sets of equal or greater heterogeneity, this problem, rather than being a relatively esoteric one, will become common.

Interoperability. In some ways, the grail of digital libraries research since the early 1990s—interoperation across disparate data and different disciplines—remains a fundamental requirement, and one for which the needs are better known than the solutions. Some characteristics of interoperation for common applications such as collaboration, visualization, and decision making include interoperation: (i) of sources, services, and ontologies; (ii) with information about possible futures; (iii) in spite of incompleteness, inconsistency, and uncertainty; and (iv) that supports authentication, privacy, and security.

4.2 Question Answering as a Grand Challenge

Question answering (Q-A) remains a grand challenge area for research. It is both extraordinarily valuable and extraordinarily difficult, and its future depends upon services and resources that are the product of digital libraries research. Substantial success has been achieved in developing systems that can answer questions for which factual data is available. ("Who is the current CEO of XYZ Co.?") This is the simplest form of a Q-A system. The next level of complexity in Q-A requires the derivation of relationships among items of factual data. ("What position did the current CEO of XYZ hold prior to becoming CEO?") Questions of this type can be handled by current systems, but questions exhibiting more complexity are beyond the current capacity of most Q-A systems. See http://trec.nist.gov/data/qa.html (question-answering track of the TREC conference) for the results of recent work. Consider, for example, the following set of increasingly difficult types of questions:

- Context
  "How has NAFTA affected the decisions rendered by the current CEO of XYZ?"
- Interpretation
  "How do the current XYZ CEO's opinions on foreign trade compare to positions he/she espoused prior to becoming XYZ's CEO?"
- Inference
  "Does the XYZ CEO have a track record of independent decision making or is he/she strongly influenced by positions taken by the Board of Directors?"
- Observation and discovery
  "Of the CEOs of the Fortune 100 companies, who has demonstrated the most objectivity in their decision making, particularly on issues where the public good and corporate profitability may conflict?"

Question answering is a very high level application that builds on sophisticated, large information stores. The techniques used to extract information relevant to a question vary by the complexity of the question. One of the reasons that Q-A systems to date have only been successful at the lowest levels of question answering is that information retrieval depends on matching search terms, rather than higher-level abstractions such as concepts. To address the more complex types of questions, more sophisticated search capabilities will be needed, including search by analogy and similarity. Much more powerful capabilities will also be needed to improve search in nontextual media and across media types.

Question answering, in many ways, is an intensely personal activity, requiring an understanding of the questioner's context, the domain of inquiry, and information resources appropriate to the inquiry. Whereas digital libraries research focused on the development of large-scale, centralized information resources a decade ago, the opportunities are now becoming apparent for more personalized systems and even for personal digital libraries. But this transition, while beneficial to individuals seeking answers, requires a relatively seamless information infrastructure, from the personal to the global. Interoperability among repositories, from the small to the very large, is required. Support for a vast array of disciplines, and for the interactions among them, is also needed. Personal digital libraries, in particular, should support the personal preferences of their users, posing substantial challenges regarding the management of potentially idiosyncratic metadata at both the personal and the community level.

4.3 Embedding Socioeconomic Values as a Grand Challenge

Learning and scientific advancement are fundamentally social activities, and the Internet is decreasing whatever degree of isolation previously existed. The Internet is increasing the pace of discourse and interaction underlying scientific progress and is generating challenges to long-held ethical traditions championed by academic and public libraries: fair use, equal access, free speech, and patron privacy. Cyberinfrastructure and digital libraries research must extend these library values, as well as those of customer service, dependability, and longevity[35], into the digital era. Doing so will nourish realization of the Internet as a shared commons for creative work, which is critical for scientific progress[36], and protecting intellectual property as an economic good, which is critical for social progress.

Digital libraries research, beyond providing the purely functional tools to support social communities (including the construction, analysis, modeling, simulation, and ongoing support of the information infrastructure supporting their activities) must also be attentive to the qualitative attributes upon which much of this work depends. Issues of trust, contributor reputation, belief systems, and consistency of findings are fundamental attributes that contribute to progress. Complicating this situation further, current scholarly communication processes do not take advantage of their capability to automatically capture the transitions through which manuscripts go as they progress from proposed articles to published reports (thereby preserving assessments of trust, reputation, etc.). At each step through the process, relationships need to be re-derived by computationally- or labor-intensive services. There is no unambiguous, recorded, visible trace of the evolution of a scholarly asset. A true grid, analogous to the data grid, is needed to support scholarly communication. In addition, temporal concerns ranging from assuring the accessibility of information tomorrow to ensuring its usability in the next century raise unanswered questions of stewardship, preservation, and curation.

5 Knowledge at Hand

Since the days of Vannevar Bush, Americans have supported basic research in the well-justified faith that investments today will pay off decades hence in national advantage. The full realization of this potential requires progress on many fundamental technical, social, and policy issues. Continued progress calls for a long-term systematic research program, coordinated with the Cyberinfrastructure program, to develop digital libraries into an institution of human knowledge. A long-term NSF research program with an annual funding level of at least $20 million is needed to promote the field and develop the science. An additional investment of $40 million annually should be provided to transform and sustain the information infrastructure, resources, and services based on the evolving science. But financial investment is not sufficient. Clear goals and a timetable are also required, as well as creative strategies to leverage the investments of other organizations. (The Tipster TREC program provides an excellent example of how this can be done to great advantage. See http://trec.nist.gov.)

In three to five years, for example, significant advances should be expected in domain-specific knowledge creation techniques, using advanced algorithms and advanced human-computer interaction principles. Instead of accessing low-level, fragmented data and individual items of information, more high-level abstract and decision-relevant knowledge should be accessible in a more seamless manner.

In eight to 10 years, a systematic science with strong theoretical underpinnings for digital library knowledge creation can be developed and validated. With appropriate rigor in the underlying science, it should be possible to make major progress toward a multilingual, multimedia, mobile, and semantics-based digital library knowledge network.

Successful outcomes from research on digital libraries will have social and economic payoffs. If individuals can find, select, organize, use, and re-use digital content in new and effective ways, the promise of digital libraries to increase productivity and foster innovation and creativity can be achieved. Indeed, some participants in this workshop posited that doubling human productivity in information-intensive activities within the next decade is possible. Too much time is currently invested in learning to use too many single-purpose tools, none of which comprehensively supports information management. Digital libraries are information-integrating tools that can enable the management of creative resources more effectively and with lower overhead than is possible today. Digital libraries provide the mechanisms to leverage the substantial economic resources being invested in building large repositories of digital content. Leverage will be achieved in several ways. Clearly, content can, in principle, be used and re-used by multiple users for multiple purposes. But interoperation among large DLs and personal DLs is required in order to achieve this, such that individuals can download data for local manipulation and can upload annotated data to share both the content and the metadata.

In a world where all the knowledge of humankind is readily accessible, information's value will come not from its existence but from its use. The limiting resource will be human attention. Research on digital libraries aspires to minimize the attention required to find needed or desired information, to optimize the effort required to understand and use it, and to facilitate and enhance creative intellectual departures to transform information, leading to discovery.

A coordinated approach that links research in digital libraries with e-science, e-learning, e-government, and e-business can maximize the benefits that accrue to both public and private endeavors, while minimizing duplication or, more seriously, the risk of developing incompatible approaches across disciplines or organizations. Students will graduate with a sound understanding of the principles and techniques of searching, assimilating, and using resources in research. Teachers will be able to generate interest in science and engineering in students through innovative and exciting learning programs that present the disciplines in new and challenging ways. Researchers will access resources efficiently and effectively, keep up-to-date with work in their fields, communicate with colleagues and peers, and make their published results available widely. Disciplines that currently have little knowledge of or interest in grid computing, cyberinfrastructure, or digital libraries will discover unforeseen opportunities and be able to use new techniques in novel and innovative ways.

Many technologies, such as the Internet and the Web, evolve from university research projects and are later adopted commercially and then by the general public. Research in digital libraries will likely take this same path, leading to a common environment for managing information, with a potential impact comparable to the widespread adoption of the Internet. The principal threat to this potential is the absence of a single, universal model, as with the Internet. Such a model can only emerge, however, through engagement of all parties that have a significant stake in this future, on a global basis. Universities have demonstrated a unique capacity to lead such endeavors through collaborative research leading to demonstrations of operational capability that sets an example for others to follow.

A ubiquitous information environment has the potential to benefit more than just universities, primarily in the industrialized nations. Broader access to information resources; participation in e-science, e-learning, and related programs; and the ability to contribute to and benefit from the research and experience of others will be a significant accelerator for developing nations moving into the digital age. The ability to transfer a uniform, consistent information environment into those areas of the world that can least afford their own development programs could be a major long-term benefit of digital libraries research.

A decade of research in digital libraries has transformed large segments of research and scholarship and is currently making innovative inroads in education. NSF's support to this research has resulted in U.S. leadership in digital information, a position that would be easy to lose but is not expensive to retain. NSF support at a level of at least $20 million annually for new research in systems and technologies supporting the creation, collection, organization, use, and long-term preservation of digital resources in a rapidly evolving global information infrastructure is critical to sustaining the U.S. leadership position. But NSF should also fund a program devoted to incorporating new research findings and valuable information resources and services into the information infrastructure. A significant component of this work would include identifying and maintaining digital resources to support research and education. The construction of infrastructure is more expensive. An adequate digital library infrastructure is estimated to require an investment of $40 million annually. The Cyberinfrastructure initiative needs both of these components, and experience during the past decade demonstrates the high return on investment from both fundamental research on and infrastructure development for digital libraries.

It is clear that digital libraries and the knowledge made available therein have immense potential to contribute to issues of national priority. But the realization of this potential requires commitment to a long-term, systematic research program, buttressed by sustainable infrastructure. Outcomes from this research will have untold social and economic payoffs. The evolving infrastructure will undergird these new capabilities and provide structure and rigor to the processes of creating, managing, and utilizing collections of information resources.

The participants were unanimous in recommending that NSF initiate two new permanent programs: a *research* program that explores the many opportunities and challenges to achieve a goal of doubling overall research productivity across disciplines within the next decade, and an *infrastructure* program in digital libraries devoted to identifying and maintaining digital resources to support research and education.

Acronyms

ACM	Association for Computing Machinery
ACP	Advanced Cyberinfrastructure Program
CBRF	Content-Based Relevance Feedback
CISE	NSF Directorate for Computer and Information Science and Engineering
CT	Computed Tomography
CV	Computer Vision
DARPA	Defense Advanced Research Projects Agency
DC	Dublin Core
DFG	Deutsche Forschungsgemeinschaft
DL	Digital Library
DLI	Digital Libraries Initiative
DWIM	Do What I Mean
EHR	NSF Directorate for Education and Human Resources
GEO	NSF Directorate for Geosciences
GPS	Global Positioning System
HITS	Hypertext Induced Topic Selection
HPCC	High Performance Computing and Communications
IITA	Information Infrastructure Technology and Applications
IP	Intellectual Property
IT	Information Technology
ITR	Information Technology Research
JCDL	Joint Conference on Digital Libraries
JISC	Joint Information Systems Committee
LAN	Local Area Network
MPS	Mathematical and Physical Sciences
MRI	Magnetic Resonance Imaging
NARA	National Archives and Records Administration
NASA	National Aeronautics and Space Administration
NEH	National Endowment for the Humanities
NIH	National Institutes of Health
NLM	National Library of Medicine
NLP	Natural Language Processing
NSDL	National Science Digital Library
NSF	National Science Foundation
OAI-PMH	Open Archives Initiative—Protocol for Metadata Harvesting
PDA	Personal Digital Assistant
PI	Principal Investigator
PITAC	President's Information Technology Advisory Committee
RF	Radio Frequency
RIACS	Research Institute for Advanced Computer Science
SGER	Small Grants for Exploratory Research
TREC	Text Retrieval Conference
UK	United Kingdom
UKE	Ubiquitous Knowledge Environment
VRML	Virtual Reality Modeling Language

Appendix 1: Workshop Participants

Conveners: Ronald Larsen and Howard Wactlar

NSF Program Director: Stephen Griffin

Report Editor: Amy Friedlander

Staff Support: Carolyn Loether

Christine Borgman
University of California
Los Angeles

Ching-chih Chen
Simmons College

Hsinchun Chen
University of Arizona

Yi-Tzuu (Y.T.) Chien
World Technology Evaluation Center

Gregory Crane
Tufts University

J. Stephen Downie
University of Illinois
Urbana-Champaign

Robert Englund
University of California
Los Angeles

Leigh Estabrook
University of Illinois
Urbana-Champaign

Ed Fox
Virginia Polytechnic Institute and State University

James French
National Science Foundation

Amy Friedlander
Council on Library and
Information Resources

David Fulker
University Corporation for Atmospheric Research

Jerry Goldman
Northwestern University

Stephen Griffin
National Science Foundation

José-Marie Griffiths
University of Pittsburgh

Margaret Hedstrom
University of Michigan

Stephen Hirtle
University of Pittsburgh

Kevin Kiernan
University of Kentucky

Judith Klavans
Columbia University

Carl Lagoze
Cornell University

Ronald Larsen
University of Pittsburgh

Michael Lesk
Rutgers, The State University of New Jersey

Clifford Lynch
Coalition for Networked Information

Gary Marchionini
University of North Carolina
Chapel Hill

Eric Miller
World Wide Web Consortium

Reagan Moore
San Diego Supercomputer Center

Douglas Oard
University of Maryland

Ronald Overmann
National Science Foundation, retired

Joyce Ray
Institute of Museum and Library Services

Bruce Schatz
University of Illinois
Urbana-Champaign

Michael Spring
University of Pittsburgh

Shigeo Sugimoto
University of Tsukuba
Japan

Terence Smith
University of California
Santa Barbara

Herbert Van de Sompel
Los Alamos National Laboratory

Howard Wactlar
Carnegie Mellon University

Donald Waters
A.W. Mellon Foundation

Gio Wiederhold
Stanford University

Robert Wilensky
University of California
Berkeley

Norman Wiseman
University of Nottingham
United Kingdom

Appendix 2: An Exercise in Futurism

Workshop participants structured their projections about potential societal impacts into two phases: a realistic assessment of technological developments over the next decade, followed by assessments of potential societal impacts.

Technological Projections

Participants in this workshop speculated on technological conditions as they might evolve over the next decade. The following list does not represent any formal conclusion, or even consensus, among workshop participants, but is presented to illustrate the range of possibilities that many found possible, if not likely:

- Personal terabyte storage will be commonplace, and communities will routinely have petabytes of data.
- Computation and communication will be plentiful (by current standards).
- The primary information formats and communication channels will be digital (phone, TV, video cameras).
- Widespread sensor networks will be commonly deployed.
- Ubiquitous computing will have arrived, at least in the developed world.
- Technology for encryption will be readily available to ordinary users.
- The network will loom large for entertainment and interaction.
- Substantial progress will be made on long-term grand challenges like natural language processing (NLP) and computer vision (CV), but it is unlikely the central problems of these disciplines will be solved.
- Good progress will be made in selective areas such as speech recognition; speech will emerge as a more important format.
- Important breakthroughs will be made in niche areas, such as quantum computing, but these will not replace the current computing model.
- Society and individuals will (still) be overwhelmed by the quantity of available, diverse, information, and there will (still) be a lot of noise.

Related Social Projections

Similarly, workshop participants speculated on the nontechnological aspects of the future we might expect in a decade. The pessimistic view apparent here was considered all too likely unless the United States takes substantial proactive steps:

- A major unanticipated IT-related threat could be experienced.
- Surveillance and information pollution could result in a significant social revolt.
- Society could express disappointment and frustration about the failure of technology, in general, and information technology, in particular, to solve an array of social problems. This failure could, ironically, further increase expectations of a technology-based, quick-fix solution.
- The intellectual property (IP) regime will likely be recognized as fundamentally flawed, with no solution readily in sight. As a result, U.S. advantage in IP-dependent segments of the economy could be compromised.
- The United States risks falling behind other nations that have more aggressive e-nation initiatives.

On a more optimistic note, workshop participants also viewed it as very likely that scientists will rely significantly on primary, publicly available data sets as a research substrate, and that this will result in an important scientific finding, discovered through the fusion of data sets.

Digital libraries, like their analog predecessors, promise to become an essential part of an enabling intellectual infrastructure. As scientific investigation and discourse are increasingly expressed in the digital medium, the importance of organizing, archiving, and preserving that medium grows proportionately. Progress in science and technology have rested on stable means of expression and communication, which historically meant that results could be transmitted reliably and authentically across generations, enabling findings to be vetted, tested, extended, and sometimes rejected. Consequently, libraries and archives, whether formally embodied in bricks and mortar on research campuses or privately organized in the back runs of journals and reports, have been intrinsic to progress. Digital libraries that organize, store, preserve, and provide access to reliable information are therefore crucial to future inquiry. And all the more so as more and more instrumentation as well as investigation, analysis, and communication of findings are computationally intensive and dependent.[37]

Within that general observation, it is difficult to predict how the research based on information in digital libraries will play out. Some of the promising areas and implications include the following:

- Digital libraries can become a ubiquitous, global knowledge resource for education, training, and (international) collaboration. Impacts could be felt in all aspects of human activities, from industries to governments and from education to research.
- Availability of shared resources as well as cross-language resources can promote cross-cultural communication, facilitate cross-cultural understanding/exchange, expand socio-economic connections, create common agendas, and thus reduce the impulse to violence.
- Shared international research programs can focus the intellect, experience, and goodwill of scientists for peaceful ends who might, for purely economic and material reasons, otherwise be forced to sell their services to those who wish us ill. Digital libraries are an essential element of such programs as well as venues that invite broad, distributed participation among those who seek knowledge and learning amid global peace and prosperity.

References

[1] Atkins, Daniel E., et. al., "Revolutionizing Science and Engineering through Cyber-infrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure," January 2003, available online at www.communitytechnology.org/nsf_ci_report.

[2] Graves, Sara, Craig A. Knoblock, and Larry Lannom, "Technology Requirements for Information Management," RIACS Technical Report 2.07, November 2002, available online at www.riacs.edu/trs.

[3] Dertouzos, Michael, *The Unfinished Revolution*, Harper Collins, 2001.

[4] Dertouzos, Michael, *What Will Be*, HarperEdge, 1997.

[5] Bush, Vannevar, "As We May Think," *The Atlantic Monthly*, July 1945, Volume 176, No. 1; pp. 101–108.

[6] Lehnert, Wendy, *The Process of Question Answering: A Computer Simulation of Cognition*, Lawrence Erlbaum Associates, January 1978.

[7] Maybury, Mark T. (Editor), *New Directions in Question Answering: Papers from 2003 AAAI Spring Symposium*, April 2003.

[8] Newell, Allen, *Unified Theories of Cognition*, Harvard University Press, December 1990.

[9] Mani, Inderjeet, *Advances in Automatic Text Summarization*, MIT Press, July 1999.

[10] Kalinichenko, L. A., and N. A. Skvortsov, D. O. Briukhov, D. V. Kravchenko, and I.A. Chaban, "Designing Personalized Digital Libraries," *Programming and Computer Software*, vol. 26, 2000, pp. 123–133.

[11] Castelli, D. and P. Pagano, "OpenDLib: A Digital Library Service System," ECDL 2002, Rome, Italy, 2002.

[12] Digital Library Panel, President's Information Technology Advisory Committee (PITAC), "Digital Libraries: Universal Access to Human Knowledge," February 2001, www.hpcc.gov/pubs/pitac/pitac-dl-9feb01.pdf.

[13] Kleinberg, J.M., "Authoritative sources in a hyperlinked environment," Proceedings of ACM-SIAM Symposium on Discrete Algorithms, pp. 668–677, San Francisco, 1998.

[14] Schatz, B., "The Interspace: Concept Navigation across Distributed Communities," *Computer* 35(1): 54–62, January 2002.

[15] Simon, Herbert, "Designing Organizations for an Information-rich World" in "Computers, Communications and the Public Interest," pp 40–41, Martin Greenberger, ed., The Johns Hopkins Press, 1971.

[16] Chen, Ching-chih, and Kevin Kiernan, "Report of the DELOS-NSF Working Group on Digital Imagery for Significant Cultural and Historical Materials," DELOS-NSF Working Group on Digital Imagery for Significant Cultural and Historical Materials, 2002, available online at http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalimaging/DigitalImaging.pdf.

[17] Wildemuth, Barbara M., et. al., "How Fast Is Too Fast? Evaluating Fast Forward Surrogates for Digital Video," *2003 Joint Conference on Digital Libraries* (JCDL'03), Houston, Texas, May 27–31, 2003, pp. 221–230.

[18] Borgman, C.L., et. al., "Children's Searching Behavior On Browsing And Keyword Online Catalogs: The Science Library Catalog Project." *Journal of the American Society for Information Science*, 46(9), 663–684.

[19] Salton, G., Automatic Information Organization and Retrieval. New York: McGraw-Hill, 1968.

[20] Downie, J. S., and S. J. Cunningham, Toward a theory of music information retrieval queries: System design implications. In *3rd International Conference on Music Information Retrieval*: 299–300, 2002.

[21] Heidorn, P. B., Image retrieval as linguistic and nonlinguistic visual model matching. *Library Trends* 48 (2):303–325, 1999.

[22] Rui, Y., T. S. Huang, S. Mehrotra, and M. Ortega, A relevance feedback architecture in content-based multimedia information retrieval systems. In *IEEE Workshop on Content-based Access of Image and Video Libraries*: 82–89, 1997.

[23] Rui, Y, T. S. Huang, M. Ortega, and S. Mehrotra, Relevance feedback: A power tool in interactive content-based image retrieval. *IEEE Trans. on Circuits and Systems for Video Technology 8* (5): 644–655, 1998.

[24] Squire, W. Muller, and H. Muller, Relevance feedback and term weighting schemes for content-based image retrieval. *Visual Information and Information Systems*: 549–556, 1999.

[25] Svenonius, E., *The intellectual foundation of information organization*. Cambridge, MA: MIT Press, 2000.

[26] Sumner, Tamara, Michael Khoo, Mimi Recker, and Mary Marlino, "Understanding Educator Perceptions of 'Quality' in Digital Libraries," *Proceedings of the Third ACM+IEEE Joint Conference on Digital Libraries*, (JCDL 2003).

[27] Jonassen, David H. and Thos. C. Reeves, "Learning with Technology: Using Computers as Cognitive Tools," *In Handbook of Research for Educational Communications and Technology*, New York: Macmillan Library Reference USA, 1996, pp. 693–719.

[28] Keller, Michael A., Victoria A. Reich and Andrew C. Herkovic, "What is a library anymore, anyway?" *First Monday*, Vol. 8, No. 5, 2003.

[29] Crow, Raym, "The Case for Institutional Repositories: A SPARC Position Paper," 2002, available online at www.arl.org/sparc/IR/ir.html.

[30] Khoo, M., "Community design of DLESE's collections review policy: A technological frames analysis," *Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries*, p. 157–164, 2001.

[31] Besser, op. cit.

[32] Ramaley, Judith, drawn from a private meeting with David Fulker about the potential value of the NSDL in understanding how technology affects the way people learn, 2003.

[33] Frew, James, drawn from private communication with David Fulker on a mission statement for the NSDL, 2003.

[34] Graves, et al., op. cit., pp. 18–19.

[35] Besser, Howard, "The Next Stage: Moving from Isolated Digital Collections to Interoperable Digital Libraries," *First Monday*, Vol. 7 No. 6, 2002.

[36] Boyle, James, "The second enclosure movement and the construction of the public domain," 2003, available online at www.law.duke.edu/pd/papers/boyle.pdf.

[37] Friedlander, Amy, "A Hybrid Environment by Choice; The Digital Medium in Higher Education," Learning Times Online Library Conference, October 22, 2003, available online at www.learningtimes.org.

University of Pittsburgh

School of Information Sciences
135 North Bellefield Avenue
Pittsburgh, PA 15260