REPORT TO THE PRESIDENT

Digital Libraries: Universal Access to Human Knowledge

President's Information Technology Advisory Committee

Panel on Digital Libraries

February 2001

REPORT TO THE PRESIDENT

Digital Libraries: Universal Access To Human Knowledge

PRESIDENT'S INFORMATION TECHNOLOGY ADVISORY COMMITTEE

Panel on Digital Libraries

February 2001

President's Information Technology Advisory Committee

Co-Chairs:

Raj Reddy
Irving Wladawsky-Berger

Members:

Eric A. Benhamou
Vinton Cerf
Ching-chih Chen
David Cooper
Steven D. Dorfman
David W. Dorman
Robert Ewald
Sherrilynne S. Fuller
Hector Garcia-Molina
Susan L. Graham
James N. Gray
W. Daniel Hillis
Robert E. Kahn
Ken Kennedy
John P. Miller
David C. Nagel
Edward H. Shortliffe
Larry Smarr
Joe F. Thompson
Leslie Vadasz
Steven J. Wallach

February 9, 2001

The Honorable George W. Bush
President of the United States
The White House
Washington, DC 20500

Dear Mr. President:

Given the high priority focus of your new Administration on the importance of education and educational technology funding to strengthen us as a Nation, we believe that you will find the enclosed report on digital libraries of special interest.

Over the past year, the President's Information Technology Advisory Committee (PITAC) has focused much of its attention on providing a vision for information technology's role in driving progress in the 21st century, particularly progress in education and human development for all citizens. One of two PITAC panels focusing on educational issues examined the status of digital libraries – the networked collections of digital text, documents, images, sounds, scientific data, and software that are the core of today's Internet and tomorrow's universally accessible digital repositories of all human knowledge. We are especially pleased to forward to you this report, ***Digital Libraries: Universal Access to Human Knowledge,*** because of the profound relevance of this technology to advancing quality education in every school, learning center, and home in the country.

Realizing the full potential of digital libraries will take strong leadership and a steady commitment of resources to tackle the substantial technical and policy barriers that inhibit our rapid progress toward universally accessible knowledge libraries. The PITAC offers four key recommendations that will make digital libraries more pervasive and usable by all citizens:

- Expand research in new systems for organizing online content, and address issues related to system scalability, interoperability, archival storage and preservation, intellectual property rights, privacy and security, and human use;
- Create several Federally funded large-scale digital library testbeds;
- Provide Federal funding to make all public Federal content persistently available in digital form on the Internet; and
- Have the Federal government play a leadership role in evolving policy to fairly address intellectual property rights in the digital age.

c/o National Coordination Office for Information Technology Research & Development
4201 Wilson Boulevard . Suite II-405 . Arlington, VA 22230
703-292-4873 (ITRD) . FAX (703) 292-9097 . nco@itrd.gov . www.itrd.gov

The PITAC strongly urges the Federal government to continue its leadership role in the research and development efforts needed to extend the capabilities of digital libraries. We believe that the recommendations in our report can help move the Nation toward realizing the enormously powerful vision of anytime, anywhere access to the best of human thought and culture, so that no classroom or individual is isolated from knowledge resources. Digital libraries can and should be an essential resource for human learning and development in the new century.

We look forward to working with you, your Administration, and members of Congress to dramatically improve our education system through the use of information technology. As PITAC strives to provide sound, well-researched advice, we hope that you and members of your Administration will feel free at any time to discuss these and other important issues with the Committee.

Sincerely,

Raj Reddy
Co-chair

Irving Wladawsky-Berger
Co-chair

Enclosure

President's Information Technology Advisory Committee

Co-Chairs

Raj Reddy, Ph.D.
Herbert A. Simon University
Professor of Computer Science and Robotics
Carnegie Mellon University

Irving Wladawsky-Berger, Ph.D.
Vice President for Technology and Strategy, Enterprise Systems Group
IBM Corporation

Members

Eric A. Benhamou
Chairman
3Com Corporation

Vinton Cerf, Ph.D.
Senior Vice President for Internet Architecture and Engineering
WorldCom

Ching-chih Chen, Ph.D.
Professor, Graduate School of Library and Information Science
Simmons College

David M. Cooper, Ph.D.
Associate Director of Computation
Lawrence Livermore National Laboratory

Steven D. Dorfman
Retired Vice Chairman
Hughes Electronics Corporation

David W. Dorman
President
AT&T

Robert Ewald
President and CEO
E-Stamp Corporation

Sherrilynne S. Fuller, Ph.D.
Head, Division of Biomedical Informatics, Department of Medical Education
University of Washington School of Medicine

Hector Garcia-Molina, Ph.D.
Leonard Bosack and Sandra Lerner Professor, Departments of Computer Science and Electrical Engineering
Stanford University

Susan L. Graham, Ph.D.
Chancellor's Professor of Computer Science, Department of Electrical Engineering and Computer Science
University of California at Berkeley

James N. Gray, Ph.D.
Senior Researcher, Scalable Servers Research Group, and Manager, Bay Area Research Center
Microsoft Corporation

W. Daniel Hillis, Ph.D.
Chairman and Chief Technology Officer
Applied Minds, Inc.

Robert E. Kahn, Ph.D.
President
Corporation for National Research Initiatives (CNRI)

Ken Kennedy, Ph.D.
Director, Center for Research on Parallel Computation, and Ann and John Doerr Professor of Computer Science
Rice University

John P. Miller, Ph.D.
Director, Center for Computational Biology, and Professor of Biology
Montana State University

David C. Nagel, Ph.D.
President
AT&T Labs

Edward H. Shortliffe, M.D., Ph.D.
Professor and Chair, Department of Medical Informatics
College of Physicians and Surgeons, Columbia University

Larry Smarr, Ph.D.
Director
California Institute for Telecommunications and Information Technology

Joe F. Thompson, Ph.D.
William L. Giles Distinguished Professor of Aerospace Engineering, Department of Aerospace Engineering
Mississippi State University

Leslie Vadasz
Executive Vice President
Intel Corporation, and President
Intel Capital

Andrew J. Viterbi, Ph.D.
President
The Viterbi Group

Steven J. Wallach
Vice President
Chiaro Networks

Table of contents

Panel on Digital Libraries

Chair

David C. Nagel

Members

Ching-chih Chen
Hector Garcia-Molina
James N. Gray
Robert E. Kahn
Raj Reddy

About This Report

"Digital Libraries: Universal Access to Human Knowledge" is one in a series of reports to the President and Congress developed by the President's Information Technology Advisory Committee (PITAC) on key contemporary issues in information technology. These focused reports examine specific aspects of the near- and long-term research and development and policies we need to capture the potential of information technology to help grow our economy and address important problems facing the nation.

The 24-member PITAC, comprising corporate and academic leaders, was established by Executive Order of the President in 1997 and renewed for a two-year term in 1999. Its charge is to provide the Federal government with expert independent guidance on maintaining America's preeminence in high performance computing and communications, information technology, and Next Generation Internet R&D.

In February 1999, the PITAC issued an overview and analysis of the current state of Federal information technology research and development in a report entitled "Information Technology Research: Investing in Our Future." That report set forth a vision of how information technology can transform the way we live, learn, work, and play, with resulting benefits for all Americans. But the report warned that Federal information technology research and development is seriously inadequate, given its economic, strategic, and societal importance. The Committee concluded that the government is funding only a fraction of the research needed to maintain U.S. preeminence in information technology and propel the positive transformations it enables.

The Committee identified 10 information technology "National Challenge Transformations" that are critical to America's future. To meet these transformation challenges, the PITAC recommended a strategic Federal initiative in long-term information technology R&D and outlined the research priorities that will drive the necessary advances in the new century.

The PITAC subsequently convened a group of panels led by Committee members and including invited outside participants with relevant expertise to examine some of the transforming applications of information technology in greater detail. Three panels focused on information technology national challenges: Transforming Government, Transforming Health Care, and Transforming Learning.

Several other panels examined critical technology issues that span the transformations, including Digital Divide Issues, Digital Libraries, International Issues, and Open Source Software for High End Computing. Over the past year, each of the panels has analyzed relevant research data and documents; held workshop discussions and conducted interviews with experts in their fields; and studied the fiscal, organizational, and economic implications of strategies to generate necessary information technology research and development advances in these key areas of our national life. The Committee plans to convene additional panels in the months ahead.

"Digital Libraries: Universal Access to Human Knowledge" and the other reports in this series present targeted findings and recommendations to the President and Congress designed to help the nation realize the vision of these positive transformations. Their benefits for our future can be extraordinary, but they are not guaranteed. To make the vision a reality, we need the results of aggressive, well-funded, and well-managed Federal research programs.

Acknowledgements

The Panel on Digital Libraries met with representatives from 17 organizations to gain a broad understanding of the state of Federal digital library research and development and deployment, to explore a number of digital library topics in depth, and to hear the views of digital library visionaries. Many of their briefings are on the www.itrd.gov/ac Web site. The Panel extends its appreciation to all of these individuals for their enormously helpful contributions to the preparation of this report:

Alexa Internet
Brewster Kahle, CEO and Director, Internet Archive

AT&T Labs-Research
Yann LeCun, Head, Image Processing Research

Coalition for Networked Information
Clifford Lynch, Director

Computer Science and Telecommunications Board, National Research Council
Alan S. Inouye, Study Director

Department of Energy
Walter L. Warnick, Director, Office of Scientific and Technical Information

Library of Congress
Caroline Arms, National Digital Library Program Coordinator

National Aeronautics and Space Administration
Milton Halem, Assistant Director for Information Science/CIO, Goddard Space Flight Center

National Endowment for the Humanities
George F. Farr, Jr., Director, Division of Preservation and Access

National Institutes of Health
Milton Corn, Associate Director, Extramural Programs, National Library of Medicine
Alexa T. McCray, Director, Lister Hill Center for Biomedical Communications, National Library of Medicine

National Oceanic and Atmospheric Administration
William T. Turnbull, Deputy Director for HPCC

National Science Foundation
Michael Lesk, Director, Information and Intelligent Systems (IIS) Division
Stephen M. Griffin, Program Director, Digital Libraries Initiative, IIS Division
James Lightbourne, Science Advisor, Division of Undergraduate Education

Office of Management and Budget
Katherine Wallman, Chief Statistician

United States Geological Survey
Barbara J. Ryan, Associate Director for Geography

The Panel is grateful to Pamela Kirkbride, District Manager-Intellectual Capital at AT&T Labs, for her substantial contributions to the drafting of our report.

The Panel would also like to acknowledge the work of the National Coordination Office for Information Technology Research and Development in supporting our efforts. The Panel thanks Sally Howe, who coordinated Panel activities, provided insights and much useful information about current initiatives, developed early drafts of the report, and carefully reviewed all of the drafts. We appreciated the constant support of Yolanda Comedy, who kept us on track, and the efforts of Martha Matzke, who edited and formatted the final document. And we are grateful to Cita Furlani, Director, and the entire staff of the National Coordination Office. Our meetings went smoothly because of their careful preparation.

Digital Libraries: A vision for universally accessible collections of human knowledge

All citizens anywhere anytime can use any Internet-connected digital device to search all of human knowledge. Via the Internet, they can access knowledge in digital collections created by traditional libraries, museums, archives, universities, government agencies, specialized organizations, and even individuals around the world. These new libraries offer digital versions of traditional library, museum, and archive holdings, including text, documents, video, sound, and images. But they also provide powerful new technological capabilities that enable users to refine their inquiries, analyze the results, and change the form of the information to interact with it, such as by turning statistical data into a graph and comparing it with other graphs, creating animated maps of wind currents over time, or exploring the shapes of molecules.

Very-high-speed networks enable groups of digital library users to work collaboratively, communicate with each other about their findings, and use simulation environments, remote scientific instruments, and streaming audio and video. No matter where the digital information resides physically, sophisticated search software can find it and present it to the user. In this vision, no classroom, group, or person is ever isolated from the world's greatest knowledge resources.

Overview

Throughout recorded history, libraries have played a significant societal role in the preservation and diffusion of human knowledge. In 1800, only a few years after the birth of the Nation, the U.S. Congress created its own library, which is now the largest conventional library in the world. The Library of Congress contains nearly 120 million items, filling more than 500 miles of shelves. Its holdings include some 18 million books, 12 million photographs, 4 million maps, 2 million recordings, and more than 50 million manuscripts, and these collections

grow by 10,000 items each day. Almost half of the library's book and serial collections are in languages other than English – more than 460 languages at last count.

As vast as the Library of Congress is, however, it contains only a small fraction of the total amount of information available today worldwide. According to a recent report by Peter Lyman and Hal Varian at the University of California at Berkeley, the world produces between one and two exabytes (a billion billion 8-bit bytes) of information each year – or roughly 250 megabytes (a million bytes) for every man, woman, and child on the Earth. Most of this information is in the form of images, sound, and numeric data; printed documents account for only 0.003 percent of the total. An increasing proportion of the information being produced is created, stored, and can be retrieved in digital form; more than 90 percent of this enormous annual output is now stored digitally. Just to keep pace with the flood of new information, the library of the future will require the means to collect and make it available digitally; and because digital information is being produced so much more rapidly than other forms, libraries of the future will perforce increasingly be libraries of digital content.

Digital libraries promise new societal benefits. One is elimination of the time and space constraints of traditional bricks-and-mortar libraries. Unlike libraries that occupy buildings accessible only to those who walk through their doors, digital libraries reside on inter-networked data storage and computing systems that can be accessed by people located anywhere in the world. When the full potential of digital libraries is realized, any citizen will for the first time be able to access all human knowledge immediately from any location. A preview of this potential can be found in the explosively popular Internet. The World Wide Web has spawned an enormous amount of information on the Internet that can be considered a rudimentary digital library. Using search engines and software programs known as browsers, millions of people access digital libraries every day without knowing they are doing so. When a computer user browses the Web and stops at a specific site – whether it is a health care site, an online newspaper, a stock performance database, or scientific information – that user is visiting a digital library.

The new digital libraries will have features not possible in traditional libraries, thereby extending the concept of library far beyond physical boundaries. They will provide innovative resources and services. One example is the ability to interact with information: rather than presenting a reader with a table of numbers, digital libraries allow users to choose from a variety of ways to view and work with the numbers, including graphical representations that they can explore.

With the extensive use of hypertext links to interconnect information, digital libraries enable users to find related digital materials on a particular topic. Human genome research has already created enormous digital libraries of the coding of life. New interactive software will use these growing digital genomic databases to discover new clues about human developmental processes, diseases, and new medicines to control them. Today's improvements in the understanding and treatment of disease owe much to the efforts in the 1980s by computer scientists to develop information technologies such as database management systems and in the 1990s by molecular biologists to create, update, and provide access to their databases. Digital libraries that are accessible over the Internet provide opportunities to advance science and technology and to dramatically improve the quality of life for our citizens.

As the next-generation Internet is deployed, its significantly higher speeds and new services will enable broader access to new forms of digital content such as streaming audio and video, and its high-speed mobile networks will allow this information to be accessed from locations not connected by wires.

All of these capabilities are made possible by past research and development (R&D) investments in the foundations of information technology, including new forms of digital storage such as magnetic and optical systems, advanced networking technologies, and key inventions such as hypertext markup languages and browsers. However, today's Internet and the immense store of information on the World Wide Web only hint at the future of digital libraries. Additional study and invention are required to develop the tools to use digital libraries effectively, in order to deliver this vision and continue the growth the Internet has spurred in both the U.S. and the global economy.

We need, for example, more efficient tools to find particular nuggets of useful information in the mountain of digital content. Virtually all of today's widely used search engines are based on fundamental algorithms invented more than two decades ago. None of these popular engines can be used to find audio or image information by content description. Searching is only one area where we are exhausting our current understanding and our current technology foundations. A number of other areas that also deserve greater study will be described below.

The President's Information Technology Advisory Committee (PITAC), chartered in 1997, established a Panel on Digital Libraries to examine R&D issues that need to be addressed so that digital libraries can reach their potential in serving human needs. This report presents the Panel's findings and recommendations about existing and emerging R&D problems facing the digital libraries community.

Why is PITAC interested in digital library R&D issues?

The PITAC is interested in digital libraries because we believe they support all of the "National Challenge Transformations" that the Committee described in its February 1999 report,"Information Technology Research: Investing in our Future." The Committee believes that these 10 transformations, listed below, are the essential prerequisites to enabling all citizens to participate fully in our society and to benefit fully from the Information Age. Digital libraries will play a central role in the transformations, each of which assumes or requires digital library capabilities to bring it to fruition.

PITAC National Challenge Transformations

- The Way We Communicate
- The Way We Deal With Information
- The Way We Learn
- The Practice of Health Care
- The Nature of Commerce
- The Nature of Work
- How We Design and Build Things
- How We Conduct Research
- Our Understanding of the Environment
- Government Services and Information

Findings

Finding 1. Although the popularity and utility of the Internet, particularly the World Wide Web, has begun to reveal the power of digital libraries, the full potential of today's digital libraries to support the national challenge transformations has not yet been realized.

Barely 10 percent of public information in print has been digitized and made available on the Internet. According to "Accessibility of information on the web" (Steve Lawrence and Steve Giles, Nature, Volume 400, 1999, http://www.metrics.com), no search engine indexes more than about 16 percent of the publicly indexable World Wide Web, and this indexing is done selectively since search engines generally index sites that have more links to them.

A variety of new technical capabilities are needed in order to improve these statistics and more generally to improve the retrieval of information and knowledge from stored digital content. For example, advances are needed in metadata – the formal ways in which content is described and made accessible – in order to improve the efficiency of search and retrieval.

Making digital libraries easier to use will further help realize their power. We need a better understanding of the requirements for specific tasks and classes of users, and we need to apply that understanding along with new technical capabilities to advance the state of the art in user interfaces.A fifth grade teacher needs a different interface when preparing for class than for classroom use, and an office worker with limited sight, a reporter on deadline, a researcher, and a retiree learning at leisure may each have unique user interface needs and preferences.

Finally, specific content collections – actual digital libraries – must be developed. Only through creating and using a variety of digital libraries, particularly large-scale libraries, will improvements be possible. In summary, we believe that a focused and accelerated research program to improve digital library technologies and coordinated efforts to

incorporate these technologies in real digital libraries will help us realize their potential.

Finding 2. The Federal government has exercised early and significant leadership in developing digital library technologies with modest investment (e.g., in multi-agency Digital Library Initiatives and in building and providing access to libraries of medical literature and scientific data). However, the Government can and should do much more to further the science, technology, and creation of digital libraries.

This is an area where the Federal government has a unique role and has already made substantial contributions. For example, the National Library of Medicine has been a leader since the 1980s through its creation of the computer-based Medline bibliographic database of medical journal literature, metathesauruses that facilitate medical literature searches, and the public-domain Internet Grateful Med for searching Medline and 14 other medical databases. In 1993 NSF funded Mosaic, the first Web browser to incorporate all the best features of earlier prototype browsers, and the first to be available on the major operating systems of the day (DOS, Macintosh, and Unix), opening the floodgates to the growth in content and use of the Internet and the Web.

Many of today's digital library accomplishments can be directly traced to early Digital Libraries Initiative (DLI) funding. Phase I of the Federal multi-agency DLI, supported from Fiscal Year (FY) 1994 to FY 1998, involved NSF, DARPA, and NASA, and was funded at $6 million per year. In it, six university-led consortia conducted R&D in applying advanced computing and networking capabilities to make large distributed electronic collections accessible, interoperable, and usable.

Phase II, begun in FY 1998 and funded at about $11 million per year, is led by NSF, with participation by DARPA, NIH/NLM, the Library of Congress, NASA, the National Endowment for the Humanities, and the FBI, in partnership with the National Archives and Records Administration, the Smithsonian Institution, and the Institute of Museum and Library Services.

The Panel is impressed by the breadth of research accomplished in DLI Phase I for such a small investment and supports its continuation in Phase II.

We also recognize the contributions of other Federal organizations to the science and creation of digital libraries. NASA and NOAA, for example, have compiled what are now some of the largest publicly accessible digital databases. These agencies make their data available to a wide range of user communities. For example, NASA is collecting huge databases of Earth-observing satellite data for use by researchers, educators, and students. NOAA provides its weather data and official weather watches and warnings to local radio and television stations and to the agriculture and transportation industries. NOAA stores more than 188 gigabytes of climate, ocean temperature, shoreline erosion, and fisheries data in its National Climate Data Center in Asheville, North Carolina. Those data are growing exponentially and are forecast to reach more than 1,500 gigabytes by 2010 – a growth of nearly 800 percent.

Despite this progress, the Panel believes that the Federal government can do much more by creating digital libraries faster, improving the access to digital content by the many people who today cannot avail themselves of it, and adopting the aggressive and visionary goal of providing digital content to every citizen. We believe further that the Federal government can use digital library technologies and content to transform the way it serves its citizens.

Finding 3. Libraries, museums, archives, and holders of other digital library collections face significant technical and operational challenges as they migrate to and maintain their holdings in digital form.

Digital libraries will help ensure the preservation of our collective history and cultural memory. But digital technologies bring their own preservation challenges. It is ironic that, as we invent media that can store ever-greater amounts of information, each new medium appears to deteriorate more rapidly than its predecessor. We can still read illuminated manuscripts created in the 12th century, but within a few decades a file stored several years ago on an eight-inch floppy disk may

be lost forever. Carefully developed and maintained digital collections may someday be rendered completely unreadable due to technological obsolescence resulting from rapid innovation in storage technologies. Just as we lose some parts of our past through crumbling books and fading photographs, we are now also in danger of losing our digital past.

Agreement on standards will help us give some digital collections a longer usable life, but we also need archival processes for periodically transferring and transforming digital resources to newer hardware and software platforms.

Digitized histories are also fragile in that many resources found on the Internet are ephemeral. We must identify not only what digital content should be preserved, but also at which stages of the life cycle of changing content. Alexa Internet (www.alexa.com) estimates that the average Web site is altered, archived, or removed every 75 days. When a Web site is altered, there is no record of its past. When it disappears, a bit of history disappears. The history of a political campaign, a company, or an event may have value to future historians. Without preservation, our digital history will be beyond recovery.

As the largest library in the world, the Library of Congress is on the frontier of dealing with the problems of digital transformation and preservation. The Library commissioned the National Research Council's (NRC) Computer Science and Telecommunications Board (CSTB) to explore the research challenges the Library faces in developing and managing digital libraries. In its July 2000 report,"LC21: A Digital Strategy for the Library of Congress," the CSTB encouraged the Library to take a leadership role in the development of digital preservation technologies and in creating relevant metadata standards and practices. The PITAC lauds those goals and recommends that Congress provide adequate personnel and funding to accomplish them. (This and other CSTB reports can be accessed at http://www4.nas.edu/cpsma/cstb.nsf/web/published?OpenDocument.)

Finding 4. Intellectual property rights need to be addressed in order to facilitate the creation of and access to digital libraries.

A variety of legal issues confront the future of digital libraries. Many are described in the January 2000 CSTB report, "The Digital Dilemma: Intellectual Property in the Information Age." An example is the recent struggle between the holders of the intellectual property rights to digital music and companies such as Napster that facilitate the exchange of copyrighted content among individual users. We now are able to cheaply disseminate seemingly infinite numbers of perfect replicas of digitized music and art among users, but such transfers may not be authorized by the rights holder and may not include royalty or copyright payment. Areas that require further study and clarification include:

- Access to information subject to copyright
- Treatment of abandoned material (material whose provenance and/or ownership cannot be ascertained)
- Policies about Federally funded content, including:
    - What Federally funded research should not be publicly available on the Internet
    - Incentives for making Federally funded data publicly available (some Federal agencies have arrangements by which private-sector organizations sell Federally funded data to which they have added value)
    - Sharing Federal data across agencies
    - Handling copyrighted information embedded in Federal digital libraries
    - Pricing policies for copyrighted material (for example, by word, file, or other measure)
- The role of the private sector – for example, what constitutes fair use (without charge) for educational purposes

We expect that digital librarians and the information technology R&D communities will need to work with copyright experts, attorneys, legislators, and regulators to address such questions so that our legal system makes both fair and prudent decisions about these complex, far-reaching issues.

Recommendations

Recommendation 1. Support expanded digital library research in metadata and metadata use,scalability, interoperability, archival storage and preservation, intellectual property rights, privacy and security, and human use.

Metadata and metadata use

The term metadata refers to systems for describing and organizing library collections. The Dewey Decimal System is such a system for conventional libraries, and the Dublin Core (http://www.dlib.org/dlib/december00/weibel/12weibel.html) is one for digital collections of textual objects. Multimedia digital libraries present metadata challenges that aren't encountered with text; for example, we need to be able to describe and search for a melody that can take the form of a soundtrack, a video, or a musical score. New methods are also needed for describing digital representations of the contents of museums (for example, paintings and sculptures) and archives. Standard methods for capturing the provenance of content – when, where, why, and how an object was created, and its history since it was created – also require refinement. Metadata that convey information about the reputation of or trust one might place in a digital object are also needed; for example, it can be useful to know that the accuracy of a scientific data set has been verified by an independent organization or that material comes from a Federal Web site. Metadata and metadata issues are embedded in each of the other research topics described below. We recommend that Federal organizations create metadata for material in their collections according to existing and evolving standards and practices, and that they develop automated tools for creating metadata

from content collections. These organizations should include the Library of Congress, the repository for much of our Nation's unique store of historical, political, and social knowledge, and the National Archives and Records Administration (NARA). Universities and other research libraries should participate in these efforts, which can provide beacons for their fields.

Scalability

Humanity generates terabytes (10^{12} or trillions of bytes) of material each day. Some organizations will soon store petabytes (10^{15} bytes) of digital content. The process of moving from the manual collection management of the past to automated software-based methods has begun, but more work is needed, given the anticipated size and diversity of content and use of digital libraries. For example, these software systems will need to understand and classify large collections of complex images in ways that are meaningful to humans. Several research initiatives are working on developing a scalable universal standard for a networked catalog of digital content that is also interoperable with established systems, but this has proved to be a difficult problem, and further effort is needed.

Interoperability

Interoperability among digital libraries is the ability to store and retrieve material across diverse content collections administered independently. The interoperability problems that reference librarians solve routinely for conventional collections are more difficult in the world of software-enabled digital libraries. Early but ambitious projects such as the simple digital library interchange protocol (SDLIP) at Stanford University and digital library interoperability work jointly conducted by Cornell University and the Corporation for National Research Initiatives illustrate both the promise and the difficulties of creating automated interoperable digital libraries.

Archival storage and preservation

Storage and preservation have already become challenges for many digital collections. Long-term storage technologies and efficient procedures for transferring ephemeral content into long-lived storage have not yet been developed. In addition, cost and space constraints are

driving librarians to share collections and services; the cost of maintaining a research collection continues to grow as much as 15 percent a year. The Internet has facilitated this practice since networked collections can be made available instantly to anyone anywhere. Research on cost-effective long-term digital storage technologies and efficient archiving and preservation processes are pressing needs.

Intellectual property rights

Managing intellectual property rights may well be one of the most complex and challenging problems digital libraries will face. Both legal and cost issues are involved. For example, a user may need to know who owns a right and be able to negotiate specific permissions. Libraries need flexible licenses that enable them to legally create archival collections and to transfer content to newer storage technologies for preservation. Licensing arrangements become complex as libraries purchase and share access to both print and electronic versions of materials. The legal liability of providing access to licensed materials can be far greater with digital information products than with their print counterparts, given the ease with which unscrupulous users can gain unauthorized access. In addition, libraries may need to charge users for access to materials. For all these functions, libraries need cost-effective, flexible, and easy-to-use software to manage access rights and copyright expenses, as well as to handle billing and payment. We recognize that technical solutions may be inadequate to address all of the intellectual property issues that exist in the context of digital libraries. Other solutions, even including Congressional legislation, may ultimately be needed to balance the interests in individuals, institutions, and copyright holders. The U.S. Government, including the Library of Congress, should lead a forum that includes publishers and public interest groups to address these matters. Continued research is required to give digital libraries the ability to manage intellectual property rights and to protect those rights while not inhibiting users' legitimate access to materials consistent with those rights.

Privacy and security

A related set of challenges for digital libraries is the ability, or in some cases the requirement, to protect digital content from unauthorized access or from unauthorized or uncontrolled use of that content such as

replicating and transmitting it to others. In some cases (for example, digital music) privacy and security technologies are required to protect the rights both of the content owner and the consumer and to enable vibrant electronic commerce. In other cases (for example, health care), privacy and security are even more critical. These challenges, which are at the core of many of today's digital intellectual property issues, have yet to be solved in broadly applicable ways.

Human use

Usability, always a difficult problem in networked software systems, is a key issue in the design and operations of global-scale networked digital libraries.A number of difficult problems are yet to be solved. How can content, both spoken and written, be made accessible to speakers of diverse languages? How do we make digital libraries equally usable to persons of diverse skill levels (for example, teachers and students)? How do we make our input and output devices (for example, keyboard, mouse, monitor, and spoken word) more flexible? How can we facilitate efficient searching for images, for example, in collections of millions or even billions of such digital objects? How do we mitigate the problems, which occur especially with large multimedia files, associated with varying access bandwidths?

The Federal agencies that should be responsible for this work include (but need not be limited to):

Defense Advanced Research Projects Agency
Department of Energy
Federal statistics agencies (such as the Bureau of the Census and the Bureau of Labor Statistics)
Institute of Museum and Library Services
Library of Congress
National Aeronautics and Space Administration
National Archives and Records Administration
National Endowment for the Humanities
National Institutes of Health
National Oceanic and Atmospheric Administration
National Science Foundation
Smithsonian Institution
United States Geological Survey

Recommendation 2. Establish large-scale digital library testbeds.

The Federal government should establish one or more large-scale testbeds for research and development in new, scalable digital libraries that address the R&D issues described above. Many of these issues (especially metadata, scalability, interoperability, and storage and preservation) will be addressed only through attempts to assemble and make useful large digital collections with diverse content.

The Government could, for example, have a digital library component of a project applying information technology to crisis management. This includes information about weather, terrain, buildings, transportation systems, real-time data on the location and status of critical resources such as ambulances, and population data. In the testbed, technologies could be developed to facilitate interoperability among the digital libraries that contain this information. The testbed could also help accelerate the development of useful applications that leverage these libraries and promote partnerships with state and local governments. We are not the first to recommend such a testbed. The NRC/CSTB, for example, identified such a requirement in its December 1999 report, "Summary of a Workshop on Information Technology Research for Crisis Management."

Recommendation 3. The Federal government should provide the necessary resources to make all public Federal material persistently available in digital form on the Internet.

Such a policy could reap significant benefits for our Nation. This country spends considerable resources collecting data (for example, population and economic data) and creating data (for example, output from weather models), and many agencies make substantial amounts of resulting information available on the Web. But these data sets are not widely known, and greater standardization and interoperability are desirable. The Government could – and we believe should – take the lead in creating and organizing such a library. It should include: all Government-collected public domain data; results of Government-sponsored R&D; the growing digital collections at Federal organizations

such as the Library of Congress and the Smithsonian Institution; and public data about the operation of the executive, judicial, and legislative branches of Government. Similar challenges exist at the state and local levels, and the Federal government can and should work with other governmental organizations on issues of common interest.

Recommendation 4. The Federal government should play a leadership role in evolving policy to deal fairly with intellectual property rights in the digital age.

We believe that the Government has a legitimate role to play in providing leadership in developing technologies and practices germane to managing intellectual property rights. For example, the Government could deploy the infrastructure to authenticate and verify Government information and to support the use of governmental digital library materials (for example, for micropayments). Government organizations such as the Library of Congress and the National Archives and Records Administration should take the lead in exploring "safe harbor" policies to support research and scholarship in such areas as archival storage and preservation. Finally, the Government may be best positioned to develop practical and fair policies for managing ambiguous or unknown property rights (for example, those associated with digital content of lost or unknown provenance).

Publications of The President's Information Technology Advisory Committee

Information Technology Research: Investing in Our Future, February 1999, 80 pages.

Resolving the Digital Divide: Information, Access, and Opportunity, February 2000, 24 pages.

Transforming Access to Government Through Information Technology, September 2000, 32 pages.

Developing Open Source Software To Advance High End Computing, October 2000, 28 pages.

Transforming Health Care Through Information Technology, February 2001, 32 pages.

Using Information Technology To Transform the Way We Learn, February 2001, 48 pages.

Digital Libraries: Universal Access to Human Knowledge, February 2001, 32 pages.

Ordering Copies of PITAC Reports

This report is published by the National Coordination Office for Information Technology Research and Development. To request additional copies or copies of other PITAC reports, please contact:

National Coordination Office
for Information Technology Research and Development
4201 Wilson Blvd., Suite II-405
Arlington, VA 22230
(703) 292-4873
Fax: (703) 292-9097
E-mail: nco@itrd.gov

PITAC documents are also available on the NCO Web site:

http://www.itrd.gov

Illustration notes

Thanks to James J. Caras, National Science Foundation designer-illustrator, for the cover illustration. The image of the Declaration of Independence on the back cover and the typography of the inside graphics are from an engraving made by printer William J. Stone in 1823. This engraving is the most frequently reproduced version of the document. The original Declaration, now exhibited in the Rotunda of the National Archives Building in Washington, D.C., has faded badly – largely because of poor preservation techniques during the 19th century. Today, this priceless document is maintained under the most exacting archival conditions possible. Images of both the engraving and the original can be accessed at http://www.nara.gov, the Web site of the National Archives and Records Administration.


PITAC
July 4, 1776.
The unanimous Declaration of the thirteen united States of America,
President's Information Technology Advisory Committee