PRESERVING OUR DIGITAL HERITAGE

Plan for the National Digital Information Infrastructure and Preservation Program

A Collaborative Initiative of the Library of Congress

First Printing, October 2002

Contents

Preserving Our Digital Heritage—Appendices

Preface

The Library of Congress and the research community owe a great debt of thanks to the U.S. Congress for so generously funding this initial plan for preserving digital materials. This support continues Congress's long tradition of conserving through its library one of America's greatest resources: the extraordinary creative record of our citizens.

Our people have long benefited from Congressional legislation that has enabled the Library of Congress to acquire, preserve, and make available the memory of this great nation. Through the Copyright Act of 1870, the television, film, and sound recording preservation acts, and, most recently, the Veterans History Project, Congress has anticipated and met the challenge of preserving America's cultural heritage in its many forms.

Now Congress has asked the Library to develop a plan to make sure that digital materials can be preserved for our national information reserve. The new digital technology offers great promise, but it also creates an unprecedented surfeit of data in an unstable and ephemeral environment. Forty-four percent of the sites available on the Internet in 1998 were no longer in existence a year later, and the average life of a Web site is now only 44 days.

Congress has understood that systematically capturing and preserving digital materials that are important for our nation and that represent our culture is a critical task that their Library is uniquely suited to lead.

During this fact-finding and initial planning period, the Library has met with hundreds of potential partners and stakeholders in America's private and public sectors. We have listened, learned, and developed a collaborative plan for action, which we present to you herewith.

James H. Billington
The Librarian of Congress

Executive Summary

Digital technology is radically transforming the ways that we create and disseminate information. This new technology has spawned a surfeit of information that is extremely fragile, inherently impermanent, and difficult to assess for long-term value. The technology has enabled and encouraged many creators: It is possible for everyone to be his or her own publisher on the Web, in large part because it is not filtered for content or quality, as traditional modes of publishing have been. Digital formats are no sooner created than they are superseded by others. As a result, it is increasingly difficult for libraries to identify what is of value, to acquire it, and to ensure its longevity over time.

Never has access to information that is authentic, reliable, and complete been more important, and never has the capacity of libraries and other heritage institutions to guarantee that access been in greater jeopardy. Recognizing the value that the preservation of past knowledge has played in the creativity and innovation of the nation, the U.S. Congress seeks, through the Library of Congress, to find solutions to the challenges posed by capturing and preserving digital information of cultural and social significance.

The NDIIPP Legislation

In December 2000, the United States Congress passed legislation establishing the National Digital Information Infrastructure and Preservation Program (NDIIPP). It charges the Librarian of Congress to lead a nationwide planning effort for the long-term preservation of digital content, as well as to capture current digital content that is at risk of disappearing. The conference report for the legislation urges the Library to work jointly with key government agencies—the Department of Commerce, the White House Office of Science and Technology Policy, and the National Archives and Records Administration—and with those entities with expertise in the collection and maintenance of archives of digital materials—the National Library of Medicine, the National Agricultural Library, the National Institute of Standards and Technology, the Research Libraries Group, the OCLC Online Computer Library Center, and the Council on Library and Information Resources—as well as with the wide group of private sector institutions working in digital formats.

Fact-Finding and Initial Planning

The goal of the plan for digital preservation is to encourage shared responsibility for digital content and to seek national solutions for:

- the continuing collection, selection, and organization of the most historically significant cultural materials and of important information resources, regardless of evolving formats,
- the long-term storage, preservation, and authenticity of those collections, and
- persistent, rights-protected access for the public to the digital heritage of the American people.

In carrying out the NDIIPP mandate, the Library established a learning process in which each stage is informed and shaped by those that come before. The process began with a yearlong, nationwide, fact-finding effort and initial planning. Next, with Congressional approval, NDIIPP will invest in a set of activities proposed under the plan that include: practical applications and modeling of key components of the digital preservation infrastructure; developing core capacities for the preservation network; building a digital preservation architecture; and conducting targeted basic research needed for the management of digital content and of the systems that support it. These investments will both leverage the knowledge gained by a range of preservation stakeholders and broaden their participation in network building.

This document is the culmination of the initial research and planning phase. It represents the fruits of intensive consultations with a wide range of American and international innovators, creators, and high-level managers of digital information in the private and public sectors. This document reports on the planning approach and on what has been learned from a variety of activities. It proposes a strategy that, as it moves forward, will be continually scrutinized and refined to keep current with the rapid and unpredictable developments of technology, of the legal and rights regime governing digital content, and of the economic and security uncertainties of our time.

In order to respond to this NDIIPP mandate, the Library of Congress embarked on planning through four major activities:

- consultation with stakeholders,
- background research,
- scenario planning, and
- defining components of the digital preservation infrastructure.

In addition to these activities, the Library has begun to capture culturally and historically significant digital information before it can disappear, as also mandated by Congress. This includes capture of Web-based information that documents contemporary events, as well as multi-media materials, sites that exemplify the extraordinary range of creativity—both of new content and of new modes of distribution—that the Web has spawned.

The *consultation with the stakeholders* began with the establishment of a 27-member high-level advisory group, the National Digital Strategy Advisory Board. This was followed by a series of national stakeholder meetings that brought together people representing: professional associations; entertainment, film, music, radio, commercial and noncommercial broadcasting; higher education; libraries, museums, nonprofit organizations, foundations and cultural institutions; newspaper, magazine, book and textbook publishing, scholarly journals; and software, Web design, and development. There were also interviews with a variety of experts in a number of relevant fields designed to elicit opinions and advice on what a national digital preservation infrastructure should accomplish and how it could be designed, built, and maintained.

These sessions brought together communities that seldom, if ever, meet—in large part because they lack a neutral forum in which to discuss common concerns. The meetings established some *baseline areas of consensus* on:

- the need for the national preservation initiative, NDIIPP,
- the need for a distributed or decentralized solution,
- the need for more research into the technologies for digital preservation, and
- the recognition that technology is an important part of the solution within the broad context of social, legal, and economic issues.

The sessions also surfaced *priorities for action or research* on:

- intellectual property and liability issues,
- scope of collecting, that is, what is preserved by whom, for whom, and in what forms,
- understanding better who the users are and will be,
- developing sustainable economic models for preservation, and
- balancing the interests of preservation and access.

All agreed that these problems are urgent; that action is needed now, not some time in the future; and that everyone—from creators to custodians—must contribute to the solution and learn to operate fluently in a world of constant and unpredictable change. The stakeholders identified the Library of Congress as an important organization in bringing resolution to a number of these issues in playing a central role as convener and facilitator of collaborative solutions.

Background research included in-depth surveys of American and international libraries and their digital preservation programs; a review of how federal agencies are responding to the challenge of preserving digital information and records; six in-depth studies on new media and the challenges they present to digital preservation; an overview of the impact of current copyright legislation on the right of a library or other collecting institution to preserve digital content; and defining the roles, responsibilities, functions, and services that comprise key elements of the preservation network. In partnership with the National Science Foundation, the Library also convened experts in digital libraries, systems development, and computer science to articulate a research agenda for digital preservation.

This research revealed that a significant number of public and private institutions agree not only on the problems of preserving digital content, but also on the best approaches to developing common solutions through collaborative actions. It also revealed the consensus need for the Library to act as convener and honest broker to bring together the many participants in preservation to clarify their respective roles and responsibilities, coordinate activities, and address such issues as selection, intellectual property, and technical standards, among many.

Scenario planning brought together a number of stakeholders, from creators, publishers and distributors, digital librarians, computer scientists, archivists and librarians, to consider the impact that key driving forces may have in the future development of the digital preservation infrastructure. The resulting views into possible futures informed later thinking about how to develop the network of partners and technology components to enable digital preservation.

Defining components of the digital preservation infrastructure began with describing the types of roles, responsibilities, functions, and services that may be present in such an infrastructure and identifying all the key elements necessary for it to operate. The critical actors are new and existing services and technologies that shape the context of digital preservation and that must be considered when designing a reliable and scalable infrastructure.

What most distinguishes the digital preservation context from the analog one now in place for libraries, archives, museums, and other heritage institutions is the sheer scale of it. It comprehends vastly larger amounts of information created in a greater variety of formats and distributed in new venues to a broader and more heterogeneous user base. This complex environment demands an infrastructure that will:

- support the needs of multiple communities over long periods of time,
- respond to rapidly changing technologies and innovative behaviors, and
- be transparent and trustworthy.

5 Strategic Direction and Plans for Action

The vision of NDIIPP is to ensure the access over time to a rich body of digital content through the establishment of a national network of committed partners, collaborating in a digital preservation architecture with defined roles and responsibilities.

The creation of such a system will occur incrementally, because of the complexity of the challenge and the number and diversity of actors involved. To realize this vision, the Library of Congress will take actions that are:

- *catalytic:* investing in existing strengths, leveraging public and private investments, and stimulating research and development where needed,
- *collaborative:* engaging willing partners and key stakeholders in areas of mutual interest and expertise,
- *iterative:* learning from the initial planning and fact-finding to inform subsequent actions and investments, and continuing to feed results back into the chain of research, testing, and development, and
- *strategic:* addressing a broad spectrum of issues in technology, collection development, infrastructure and organization, intellectual property, technical standards, and other key components of the preservation network through a balance of early short-term and long-term actions and investments.

To begin building the preservation infrastructure, the Library proposes a strategy for working on the two key components that support it: developing a network of participants and building the technical framework.

Working with partners in the public and private sectors, including the National Science Foundation, the Library will invest in projects that develop core capacities of the infrastructure in the following areas:

Selection and Collection Development

Among the collaborative actions in the area of selection and collection development to be undertaken are:

- developing cooperative agreements between national libraries about the scope of collecting national materials, both Web-based and licensed,
- developing cooperative collecting agreements with libraries, archives, and other collecting institutions in the public and private sectors,
- convening experts to develop guidelines for assessing content for enduring value,
- convening experts to examine curatorial best practices for selecting dynamic objects, such as Web-based journals, GIS materials, interactive objects, and other genres,
- defining the boundaries of Web-based content for preservation purposes, and

- reviewing collection development policies, including those for best edition necessary for mandatory copyright deposit, in order to bring them up to date for digital materials.

Intellectual Property

Recommended actions include:

- investigation of the options and authorities necessary for the Library of Congress to preserve digital content captured on the Internet,
- development of acceptable methods of access to digital content for educational purposes within a library setting,
- investigation of the implications of mandatory deposit for digital content,
- investigation of the implications of various security and protection devices for preservation, and
- development of a better understanding of the international context of copyright, jurisdiction, responsibility, and reach of applicable law, possibly in cooperation with other national libraries and multinational publishing and media industries.

Business Models

Among actions in the area of business models to be undertaken are:

- identification of incentives for institutions to undertake preservation,
- identification of incentives for creators to deposit content,
- development of metrics for costs and benefits of digital preservation,
- development of metrics for appraising digital works for purposes of insurance and tax, and
- development of model safe-harbor agreements for those materials that are preserved by commercial entities or others that may not be best positioned to ensure longevity.

Standards and Best Practices

Activities that the Library will continue or initiate are:

- coordinating and documenting standards that support key preservation services, such as metadata and persistent identifier schemes,
- fostering research and best practice recommendations for formats and encoding schemes,
- fostering research and development of strategies, such as migration and emulation, that will ensure sustainability of digital content, and
- developing a communication strategy to track technology changes and their impact on preservation.

Communication and Outreach

Outreach activities targeting professional and public audiences include:

- maintaining the NDIIPP Web site *(www.digitalpreservation.gov)*, featuring current information on the program's status,
- outreach to professional groups through participation in professional meetings and contributions to professional literature, and
- outreach to the public through print and Web-based general interest publications and through the broadcast media.

Digital Preservation Architecture

To address the next steps in building the digital preservation architecture, the Library will work with a variety of public and private institutions as well as the National Science Foundation to:

- convene a design group to further develop the components of the preservation architecture,
- solicit proposals to test and model components of the system, and
- evaluate project outcomes to inform a next generation of implementations.

Expected Outcomes

Through the execution of the NDIIPP initiative, the Library expects to have achieved:

- a clearer definition of the roles and responsibilities of partners in the preservation network,
- establishment of the relationships among key partners in the private and public sectors, including international institutions,
- clarification of intellectual property issues that impede preservation, together with recommendations to address them,
- creation of an advanced design for the digital architecture,
- identification of the next set of investments to advance NDIIPP goals,
- an advanced research agenda for preservation of digital content, and
- raised awareness among key stakeholders and the general public about the challenges and opportunities of digital preservation.

The Library's plans for action over the next three to five years of NDIIPP will comprise investments and activities that will preserve digital content, build a resilient network of digital preservation partnerships, and begin development of the digital preservation architecture to support and enable these goals.

With the foresight and support of the U.S. Congress, together with a team of dedicated collaborators, the Library of Congress will lead a national initiative to preserve the nation's cultural heritage—in all its forms—for generations to come.

Introduction

What Is at Stake

The Library of Congress occupies a unique place in American civilization. Established as a legislative library in 1800, it grew into a national library in the 19th century as the unique depository of all copyrighted materials. Since World War II, it has become an international resource of unparalleled dimensions. It embodies the belief of the Founders that self-government depends vitally on free and open access to knowledge and the unhampered pursuit of truth by an informed and involved citizenry.

Democracy works through a knowledge-based society, and the system of libraries that undergirds our national knowledge network has been generously and steadfastly supported by the U.S. Congress, both directly through the Library of Congress and through the budgetary and legislative actions that support the country's library and information infrastructure. Together with the National Archives and Records Administration, the Library of Congress preserves a record of our national experience. In addition, the Library, through its special relationship with the Copyright Office, has built a record of the creativity and innovation of the nation.

This record of information is jeopardized by the transformation that digital technology is forging. This new technology has spawned a surfeit of information that is extremely fragile, inherently impermanent, and difficult to assess for long-term value. America's economic and political strength has relied for generations on the innovation and productivity of its people, and the "progress of science and useful arts" (stipulated in Article 1, Section 8 of the U.S. Constitution) depends on the reliable preservation of knowledge and information for generations to come.

Americans look to libraries to facilitate research in complete, authentic, original, undistorted sources. But we do not yet know how to preserve digital content, or even which content to preserve. Building a digital preservation infrastructure that will work alongside the one already in place for print and audiovisual materials poses great technical challenges. But to an even greater degree, it requires forging the legal, economic, and social agreements that will ensure that important digital data are deposited in their original form into a trusted repository for safe custody.

The NDIIPP Legislation

In December 2000, Congress passed PL106-554 establishing the National Digital Information Infrastructure and Preservation Program (NDIIPP). It charges the Librarian of Congress to lead a nationwide planning effort for the long-term preservation of digital content. The conference report urges the Library to work jointly with key government agencies—the Department of Commerce, the White House Office of Science and Technology Policy, and the National Archives and Records Administration—and with those entities with expertise in the collection and maintenance of archives of digital materials—the National Library of Medicine, the National Agricultural Library, the National Institute of Standards and Technology, the Research Libraries Group, the OCLC Online Computer Library Center, and the Council on Library and Information Resources—as well as with the wide group of private sector institutions working in digital formats.

For almost a decade, since the Internet became readily accessible through the World Wide Web, digital technology has been radically transforming the ways that we create and disseminate information. This continuing transformation of the information landscape is having profound effects on our society, our economy, our national security, and our lives as citizens of a democratic republic. It is also transforming the institutions that collect, preserve, and provide access to digital content in ways that cannot be overstated yet are still little understood. Now everyone can be his or her own publisher and digital formats are no sooner created than they are superseded by others. As a result, it is increasingly difficult for libraries to identify what is of value, to acquire it, and to ensure its longevity over time. Never has access to information that is authentic, reliable, and complete been more important, and never has the capacity of libraries to guarantee that access been in greater jeopardy.

By establishing the National Digital Information Infrastructure and Preservation Program, the United States Congress recognized that the Library of Congress is uniquely positioned to bring together all the stakeholders in this new digital landscape—creators, distributors, and users—to address the problem of sorting and preserving significant content in the burgeoning world of digital information. With its core mission to make information available and useful, and to sustain and preserve a universal collection of knowledge and creativity regardless of format for current and future generations of Congress and the American people, the Library of Congress has a long history as a trusted convener able to facilitate the development of standards and best practices in librarianship and cultural stewardship across the country and internationally.

In carrying out the NDIIPP mandate, the Library established a learning process in which each stage is informed and shaped by those that come before. The process began with a yearlong, nationwide, fact-finding effort and initial planning. Next, with Congressional approval, NDIIPP will invest in a set of activities proposed under the plan that include: practical applications and modeling of key components of the digital preservation infrastructure; developing core capacities for the preservation network; building a digital preservation architecture; and conducting targeted basic research needed for the management of digital content and of the systems that support it. These investments will both leverage the knowledge gained by a range of preservation stakeholders and broaden their participation in network building.

This document provides a plan for preserving digital information of national significance. It is the culmination of the initial planning phase. It represents the fruits of intensive consultations with a wide scope of American and international innovators, creators, and high-level managers of digital information in the private and public sectors. This document reports on the planning approach and on what has been learned from a variety of activities. (More details will be found in the accompanying appendices, and readers will be referred to these documents for additional information.) And it proposes a further set of actions and investments to begin practical applications and modeling approaches to implementation of NDIIPP. Even as NDIIPP moves forward, this plan will be continually scrutinized and refined, given the need to keep current with the rapid and unpredictable developments of technology, of the legal and rights regime governing digital content, of local and national economies, and of the grave uncertainties of our time.

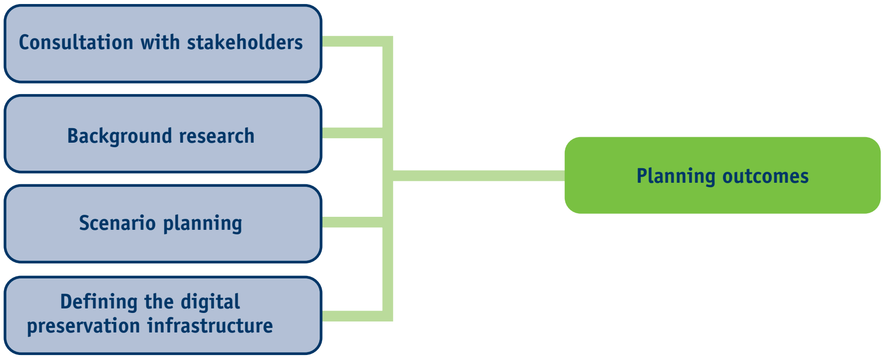
In the digital realm, the Library of Congress has developed model programs of sharing collections through digitization and has extended those programs through public-private partnerships to libraries in the United States and indeed abroad. (Among such programs are American Memory, Meeting of Frontiers, and the Global Legal Information Network [GLIN]). The Library must now move quickly to make similar progress with so-called born-digital materials and to develop a strategic vision for its role in this new information landscape. The Library recognizes that the critical next step in meeting the demands of stewardship in the digital age begins with casting its net wide to include other libraries, other federal agencies, and producers and distributors in the private sector.

Fact-Finding and Initial Planning

In order to respond to this NDIIPP mandate, the Library of Congress embarked on planning through four major activities:

- consultation with stakeholders,
- background research,
- scenario planning, and
- defining components of the digital preservation infrastructure.

Figure 1. Planning approach


Consultation with stakeholders
Background research
Scenario planning
Defining the digital preservation infrastructure
Planning outcomes

In addition to these activities, the Library has worked aggressively to begin to capture culturally and historically significant digital information before it can disappear, as also mandated by Congress. This includes Web-based information that falls within the Library's collecting scope of documenting the history and creativity of the American people. These collecting activities include documenting the elections of 2000, the events of September 11, the 2002 Winter Olympics, as well as the upcoming elections of 2002 and the complete archives of eight of the journals of the American Physical Society. Other pilot projects have explored the selection, capture, and preservation challenges of Web-based moving-image materials, sites that exemplify the extraordinary range of creativity—both of new content and of new modes of distribution—that the Web has spawned.

The goal of this plan for digital preservation is to encourage shared responsibility for digital content and to seek national solutions for:

- the continued collection, selection, and organization of the most historically significant cultural materials and the most important information resources, regardless of evolving formats,
- the long-term storage, preservation, and authenticity of those materials, and
- persistent, rights-protected access for the public to the digital heritage of the American people.

The national solutions we seek must be arrived at by a collaboration among key stakeholders and in an atmosphere of trust and of willingness to learn from experience and share knowledge. The Library is uniquely positioned to play the role of convener and honest broker in forging these new relationships and operating agreements. We designed a planning process to achieve the following outcomes:

- defining the problem of digital preservation and its scope,
- identifying and engaging preservation stakeholders, broadly defined,
- beginning the conversation among concerned parties,
- identifying a network of libraries and private-sector partners for action,
- defining their respective roles and that of the Library of Congress,
- developing a preservation research and development agenda,
- creating a framework for developing sustainable models of preservation, and
- raising awareness of digital preservation.

These goals called for the following planning steps, all accomplished in the past 18 months:

- establishing a 27-member National Digital Strategy Advisory Board,
- coordinating with other federal agencies,
- convening sessions of stakeholders to listen and learn,
- commissioning reviews of emerging digital content,
- commissioning a background study on copyright’s role in preservation,
- surveying national and international initiatives in this area,
- defining the types of roles, responsibilities, functions, and services for digital preservation,
- developing a series of possible future scenarios and contingencies, and
- developing a digital preservation architecture that establishes critical consensus on technical approaches.

As the planning process unfolded over the course of the year, much was learned from those the Library engaged in dialogue, and modifications were made in the process as needed. Unexpected outcomes early in the planning process enabled the Library to move more quickly than anticipated. For example, a resounding consensus was achieved rapidly on the part of all stakeholders that preservation is a serious problem that demands swift action. Even those who are chiefly concerned about access—such as publishers and the entertainment media—see the lack of preservation as a major threat to their core missions. There was also consensus that the most important role the Library can play in this stage of development is as a convener and facilitator of collaborative solutions.

Consultation with Stakeholders

Consultation with stakeholder communities began in the spring of 2001 with the formation of a high-level advisory board and was followed by several stakeholder meetings that also included international participation. This was accompanied by continuing interviews and consultation with a wide group of other experts during the year (see Appendix 1 for a list of participants).

National Digital Strategy Advisory Board

Early in the consultative period, the Librarian of Congress established the National Digital Strategy Advisory Board (NDSAB), comprising representatives from other federal agencies, industry, research libraries, and foundations (see Appendix 1 for membership). The Advisory Board is charged, among other things, to provide advice on national strategies for the long-term preservation of digital materials; advice on identifying and prioritizing national issues such as intellectual property and rights management, requirements for archiving and repositories, planning for life-cycle management of digital content, and promoting collaboration among stakeholders; and advice on practical applications and modeling of preservation strategies. The NDSAB has convened three times in the past year.

Stakeholder Meetings

Stakeholder meetings were a key part of the Library's systematic approach to consultation and were designed to:

- identify barriers to and opportunities for building a distributed preservation infrastructure,
- engage the actors and explore the new roles and responsibilities they might be willing and able to fulfill,
- seek advice on issues that included technology, rights to archive, and permission to access digital content, and
- explore models of sustainability for the costly enterprise of archiving.

In November 2001, the Library convened in Washington, D.C. approximately 70 people representing: professional associations; entertainment, film, music, radio, commercial and noncommercial broadcasting; higher education; libraries, museums, nonprofit organizations, foundations and cultural institutions; newspaper, magazine, book and textbook publishing; scholarly journals; and software, Web design, and development. The Library brought together vice presidents for technology and chief technology officers; senior librarians and archivists; senior executives and editors in commercial and noncommercial broadcasting, and in scholarly and commercial publishing and journalism; university professors and academic administrators; program directors; a present and a past CEO of major pioneering corporations; and others in leadership and decision-making positions. Many of these individuals have diverse career portfolios and were able to speak to several issues across multiple industries.

Other federal agencies and library organizations also participated in the meetings, including representatives from the Department of Commerce, the National Archives and Records Administration, the National Institute of Standards and Technology, the National Library of Medicine, the National Agricultural Library, the OCLC Online Computer Library Center, the Research Libraries Group, and the Council on Library and Information Resources.

These sessions brought together communities that seldom, if ever, meet, in large part because they lack a neutral forum in which to discuss common concerns. The meetings established some baseline areas of consensus: on the need for the national preservation initiative, NDIIPP; on the need for a distributed or decentralized solution; on the need for more research into the technologies for digital preservation; and a recognition that technology is an important part of the solution but not the most important.

The sessions also surfaced priorities for action or research on: intellectual property and liability issues; the scope of collecting, that is, what is preserved by whom, for whom, and in what forms; understanding better who the users are and will be; and balancing the interests of preservation and access. There were additional concerns about developing sustainable economic models for the expensive enterprise of preservation.

All agreed that the problem is urgent; that action is needed now, not some time in the future; and that everyone—from creators to custodians—must contribute to the solution and learn to operate fluently in a world of constant and unpredictable change.

In addition to these sessions, the Library continually sought out interviews and meetings, often on a confidential basis, with leaders in industry, technology, librarianship and archiving, and scholarship. The interviewees were asked to identify their enterprises' needs for preservation; to identify what roles their organizations could (and could not) play in a distributed preservation infrastructure; what role they would like to see the Library of Congress play; and what specific actions can and should be taken in the near term. These conversations, along with the public sessions, have deeply informed the planning process, and the fruits of those conversations will be found throughout the following section on the current state of digital preservation.

Background Research

Surveying the Landscape

Surveys, studies, and other information gathering continued throughout the year. As background for the stakeholder meetings in the fall, the Library worked with the Council on Library and Information Resources (CLIR), a not-for-profit organization devoted to issues facing libraries and archives, to commission a series of six reviews of media types, conducted by experts of national repute. This work focused on areas of digital collection development that present new challenges to preservation: e-journals; e-books; digital sound recordings; digital video; digital television; and Web archiving (see Appendix 2).

In addition, the Library commissioned through CLIR surveys and assessments of the current digital preservation activities of the Association of Research Libraries, the Digital Library Federation, and major national libraries abroad (see Appendices 3, 4, and 5), and inventoried digital initiatives in the United States. In response to concerns raised at the November stakeholder meetings, a leading expert in copyright was called upon to write a report detailing the relationship between preservation and copyright, and to clarify specifically the effect that copyright legislation has on a library's right to preserve materials and in what format (see Appendix 6).

The Library consulted systematically with other federal agencies and libraries that have expertise in managing digital information.

Defining Components of the Digital Preservation Infrastructure

A concurrent effort defined the scope of the digital preservation infrastructure, building on the knowledge yielded through stakeholder meetings, independent research, and further consultation (see Box 1, page 18). The purpose of defining this context is to: identify the scope of activities that transpire in this landscape; identify all the potential actors and institutions—from writers and filmmakers to libraries and archives; and to define future (existing and new) coordinating bodies, research and development activities, and enabling agreements that will ensure the preservation of and access to digital content over time. This broad context also accounts for the functions and services required for the preservation and use of digital assets and related information. Together, these actors, coordinating bodies, enabling agreements about roles and responsibilities, research, policies, and practices all constitute the *digital preservation network.* Finally, the infrastructure includes the technology components and technical standards necessary to build the *digital preservation infrastructure.*

What most distinguishes the digital preservation context from the analog one now in place for libraries, archives, museums, and other heritage institutions, is the sheer scale of it. It comprehends vastly larger amounts of information, created in a greater variety of formats, and distributed in new venues to a larger and more heterogeneous user base. The digital preservation infrastructure will be characterized by a complex network of relationships and dependencies that are simply unknown in the world of print and analog sound and image resources.

Box 1. Key NDIIPP Terms

The ***digital preservation infrastructure*** comprises two key components: the *digital preservation network* of partners collaborating to preserve and provide long-term access to digital content; and the *digital preservation architecture,* the technical components that enable digital preservation. The *digital preservation infrastructure* is designed to support the needs of multiple communities over long periods of time; to respond to rapidly changing technologies and innovative behaviors; and to be transparent and trustworthy.

The ***digital preservation network*** is the actors—creators and producers, owners, collectors, distributors, preservers, users, and others—who collaborate to preserve digital content. The network also comprises the coordinating bodies, enabling agreements about roles and responsibilities, research, policies, and practices that make collaboration possible.

The ***digital preservation architecture*** is the technical framework that specifies the structures, logical components, and logical interrelationships of a system that enables digital preservation through a network of partners.

Leveraging Federal Research Investments

In order to mobilize funding for work on the critical issues of digital preservation, the Library co-sponsored with the National Science Foundation's Digital Government and Digital Libraries Initiative programs an effort to define significant and feasible research challenges that will engage researchers from academia, industry, and government. Through initial consultations with numerous professional consortia and 15 federal agencies, a preliminary agenda of investigation was agreed upon, focusing on:

- core technical infrastructure,
- preservation technologies and tools,
- access to and use of preserved digital content, and
- basic computer science issues for large-scale storage systems.

In partnership with the Library of Congress, the National Science Foundation (NSF) convened a workshop in April 2002 that brought together 51 specialists from government agencies, academia, and industry with expertise in computer science, mass storage systems, archival science, digital libraries, and information management to develop a research agenda. NSF and the Library of Congress are planning to put out a call for proposals in fiscal year 2003. While research results from this investment will typically not be completed for three to five years, or even longer, there are also a number of areas of more narrowly defined applied research—into cost models of various acquisition technologies, of life-cycle management, and so forth—that are likely to find funding during this period from foundations, other libraries, and the Library of Congress (see Appendix 7).

Scenario Planning

As demonstrated by preliminary research and consultation, there is no clear solution or set of solutions to meet the challenges of digital preservation. The unpredictability of technological development (who could have predicted the explosive appearance of the World Wide Web in 1985, 10 years before it appeared?); the business climate (who could have predicted the Internet boom and bust of the late 1990s?); and, of course, the global political environment that changed forever on September 11, 2001 —all contribute to the challenge of plotting a course in the face of a wide range of possible futures.

To that end, the Library undertook a process familiar to corporations, that of scenario planning. It engaged Global Business Network (GBN) to conduct the planning, and it in turn enlisted a number of experts from libraries, media, content creation, and technology companies to identify collectively the key driving forces and variables in the foreseeable future—in this case, 2017—in order to prepare possible futures for the Library and for digital preservation actors broadly defined. The resulting views into possible futures informed later thinking about how to develop the network of partners and technology components to enable digital preservation (see Appendix 8).

Digital Preservation Architecture

The Library developed a high-level model of a technology infrastructure that would allow all the actors in the digital preservation network—from a manager of a digital repository to a television production company seeking to comply with the mandatory deposit requirements of the copyright law—to envisage how a distributed preservation infrastructure might work to benefit their stated interests in preservation. This has been defined as the *digital preservation architecture* (see Box 1, page 18, and Appendix 9).

Planning Outcomes

The Library has clearly learned, through extensive consultation and convening of diverse communities, that many new actors in the digital information landscape share Congress's concerns about the fate of digital content. They expressed readiness to partner with the Library to find appropriate ways to be stewards of our collective digital present and future. The Library identified shared understandings about digital preservation on the following:

- the problem is widespread among stakeholders,
- the problem is urgent,
- research is needed into new technologies for archiving,
- the problem cannot be solved exclusively through technology,
- the solutions are many,

- the solutions must be collaborative,
- the solutions will change and evolve with time,
- standards for information description and sharing are needed, and
- the Library of Congress can play a unique role as trusted third party to convene and facilitate.

The Library also identified a significant number of external factors that bear greatly on any outcome:

- the rapid and unpredictable growth of technology,
- the changing economic and political climate, and
- the legal and rights management regimes that control digital content.

All stakeholders recognize the danger in letting preservation issues be driven by those of access. Many came to the stakeholder sessions with the fear that issues of copyright and asset management would cut off conversations that need to focus on preservation. Stakeholders—from creators and distributors to scholars and librarians—all started from the same point: fear of losses. Business fears loss of current and future revenue; scientists fear loss of data crucial for the progress of science and engineering; scholars and librarians fear loss of the cultural and historical record.

Digital information technology is not evolutionary, but revolutionary: It transforms how we create, acquire, disseminate, and preserve information. The scope of the problem it creates is incommensurate with present experience and resources. The Library therefore advocates an approach to the problem that builds on the collective experiences, expertise, and will of digital creators and libraries. The approach is:

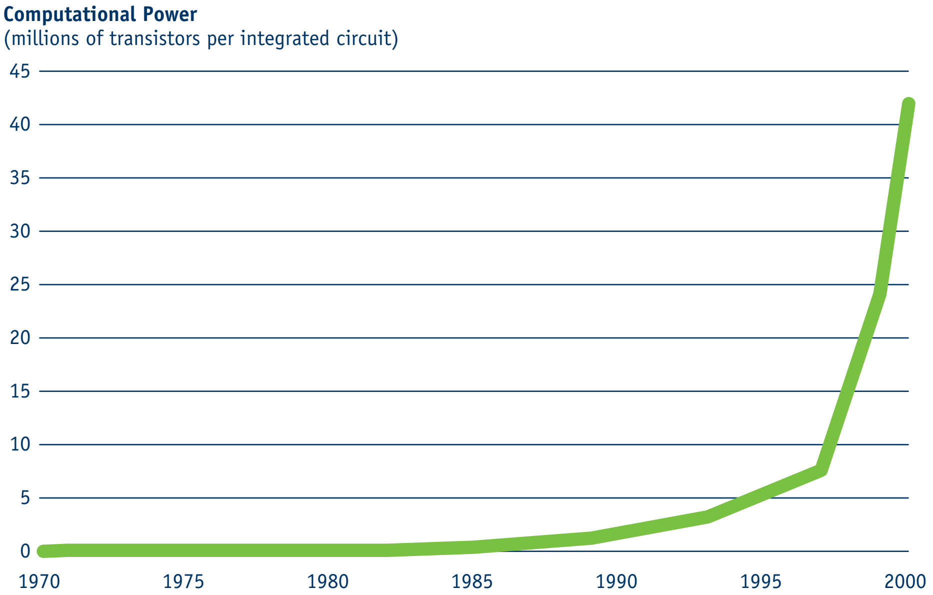
- *catalytic:* invests in existing strengths, leverages public and private investments, and stimulates research and development where needed,
- *collaborative:* engages willing partners and key stakeholders in areas of mutual interest and expertise,
- *iterative*: learns from the initial planning and fact-finding to inform subsequent actions and investments, and
- *strategic:* addresses a broad spectrum of issues in technology, collection development, infrastructure and organization, standards, and other key components of the digital preservation infrastructure.

Stakeholders in preservation must remain poised to adapt quickly to externalities that may prove critical but over which they have little control: economical, technological, legal, regulatory, and those relating to national security.

The State of Digital Preservation

It is often said, citing Moore's Law that the number of transistors (hence, computing power) held on a chip doubles every 18 months, that the costs of storage are going down, and therefore "preservation will not be a problem." But preservation is not merely storage. The goal of preservation is to maintain an information asset so that is it is readily accessible for use, no matter what format it was originally in, and ensuring that it is authentic and reliable by preventing such things as tampering, accidental corruption of files, media degradation, and losses through software and hardware obsolescence. This mandates active, not passive, management of content and thus involves a large number of actors to work collaboratively toward the common goal of preserving digital heritage.

Figure 2. Moore's Law: Computer Processing Power Doubles Every 18 Months



Source: Taken from information presented in "Moore's Law—Overview," Intel Corporation, *http://www.intel.com/research/silicon/mooreslaw.htm;* downloaded September 10, 2002.

Challenges of Collecting and Preserving Digital Content

Current Environment

One of the chief tasks of NDIIPP is to identify and provide for all the barriers to progress in digital preservation. The most salient are those caused by the rapid changes in technology. Frustrations are shared by industry and collecting institutions alike over the multiplicity of formats, rapid technological changes, and hardware and software obsolescence that plague the new information technologies. Even storage media that promise stability, such as CDs, are subject to unpredictable degradation and have not demonstrated that they are of archival quality. Producers of content and preservationists also report serious problems arising from the formal and informal standards-setting processes that result in both too many and too few technical standards, and, for film, television and sound, as well as early computer files, problems associated with playback.

Representatives of the film, television, and music industries also agreed that there are substantial technical differences among the formats, despite commonalities. There is some disagreement among experts about whether to focus on the issues that arise within format types (musical versus image-based versus text, for example) and in which consensus might be reached among fewer stakeholders, or whether it is better to address problems faced by all, where more resources might be mustered but agreements among disparate communities might be hard to cement.

But these technological problems are not the only ones people are concerned about. No one consulted in this planning process asserted that the problems are solely or even chiefly technical. Indeed, some even suggested that technology will address some problems sooner rather than later and that other major problems—legal, social, and economic—will remain. Moreover, it is clear that the new technologies are resulting in shifts in institutional roles and functions that are not well understood. However, before exploring some of the more important implications of those changing roles, it is important to clarify what content is being created with new technologies and what problems arise from the nature of these digital objects.

Transforming Content

Investigations into six new media types, commissioned for NDIIPP (see Appendix 2), provide a baseline for understanding emerging issues that will mark the digital landscape into the future. The formats explored—e-books, e-journals, digital music, digital television, digital video, and Web sites—are complex and present enormous technical challenges in their creation, distribution, and preservation. These difficulties have profound implications for the right to preserve as well as access. These new formats redefine genres: even such things as books and journals that seem so stable in the print regime are redefined online. The distribution models used with particular formats, above all, the Web, blur the line between published and unpublished. They facilitate new user behaviors, create new user expectations, and in fact draw new users to old content. They demand new approaches to selecting and cataloging. They use new means of distribution to reach new audiences. Some genres break traditional ties between ownership and preservation, such as e-journals that license their content when their print counterparts were bought and sold as physical artifacts in which a library had certain rights. Nearly all require careful thought be given to what constitutes the so-called best edition of a given work, the copy of a copyrighted work published in the United States that is deposited with the Copyright Office to comply with the mandatory deposit requirement of the copyright law. Above all, these formats show how different each new genre can be and that no one solution for preservation works for all, any more than book conservation techniques work for restoring nitrate film.

The summaries below, based on the environmental scans, illustrate the range of format and genre complexities that must be addressed in NDIIPP. They are followed by a short consideration of even more complex digital media types, such as Geographic Information Systems (GIS), that will take libraries, archives, and museums well beyond the formats that they have expertise in preserving. Taken together, these digital content types also begin to define who are the actors in preservation and access, what are their roles and responsibilities, and what agreements among them must be renewed or redefined to ensure future access to our digital heritage.

E-Books and E-Journals When books "go digital," the conversion of texts into digital formats may seem fairly straightforward from a technical point of view. But in fact it reveals just how complex an object a book really is—a form that took centuries of development after the invention of printing to evolve into something familiar to us today, with standards for spelling and grammar, typeface, production and distribution, and preservation.

Because e-books are read on handheld devices that do not approximate the size of a book, the correlation between a printed page and an e-book screen varies—and it varies in turn from device to device—so even elementary orienting devices, such as pages and page numbers, need to be reconceived for the e-book. Industry is devising new standards for online books that allow for some conformity among different proprietary software approaches. But, given the commercial nature of the enterprise, so far the problem is not a lack of standards, but a proliferation of competing ones (at least 26 e-book standards initiatives are cited by Frank Romano, the expert who wrote the environmental scan on e-books). Perhaps out of fear induced by peer-to-peer music file sharing programs, publishers and distributors of e-books are turning to elaborate security precautions in a market that is, as yet, underdeveloped.

But peer-to-peer sharing of text has been the norm in print, protected by the doctrine of first sale, and has supported the public access provided by libraries and undergirds the market in used books. And it has other important commercial benefits: advertising rates for journals, for example, are calibrated to the estimate of how many times one copy will be shared and thus how many eyes can be captured with one copy. The future market for e-books remains a subject of some debate in the publishing world. At present it looks as if the needs of the commercial sector to protect and promote its digital books through proprietary software and hardware devices throws up additional barriers to cost-effective and scalable preservation approaches. Much work has been done by publishers and libraries with e-book content, and much remains to be done. NDIIPP will leverage the experience and expertise that exists in this field to resolve some of the issues that have been raised above.

Journals—serial publications that aggregate articles by different authors—present even greater problems than books when they move online. Journals usually comprise a great variety of information, from articles and short features, to editorial board listings, graphics and photographs, and advertising. Online they are frequently in different formats and often provided solely through links to other providers. With articles replete with citations to secondary online resources that may themselves be under rights protection, or may link to a source (such as a database) that is not preserved or itself changes, what is a library to preserve? What are implications for science and the need to test and replicate results if data referred to in one article are linked to a site that is later unavailable? To preserve an article, must one preserve all the links? Does one have a right to? What should a publisher do about correcting errata online? Does the corrected version supersede the first one? Are they both necessary for the historical record? And the issues of online advertisements is most perplexing: usually ads are targeted for specific audiences, often created "on the fly" (dynamically), and frequently updated.

In print newspapers and magazines, advertisements have come to be seen by researchers as rich original resources that provide social, economic, artistic, and other context for the contemporary content they accompany. They are highly valued by historians, theater and film set designers, genealogists, and any number of users. How should libraries preserve the technically complex advertising that supports so many digital periodicals? Similarly complex are the growing variety of supplemental materials that can be conveyed online that tend to exploit the possibilities of the technology more than the texts do: spreadsheets, visualizations, computer simulations, executable files that project weather patterns, to name but a few. These all constitute primary evidence of the innovation spawned by the digital revolution and, as such, should be prime collecting targets for contemporary libraries. There are a number of fruitful collaborations between libraries and e-journal publishers under way that can be leveraged by NDIIPP to address such issues as how to handle links to outside documents, what content should be preserved as part of a best edition, and which business models would support both sustainability and long-term preservation.

Digital Sound Recordings The issues surrounding the preservation of digital recordings of music are exponentially more complex than those for print materials, for example, books or sheet music. As with digital text, the problems are technical, legal, and economic in scope, but because the technology and media used in sound recording are more complicated and fragile than print on paper, and the rights regime surrounding uses such as performances and reproductions more layered than for print publications, the way forward for preservation is even less clear. There was no copyright protection for the sounds and the contributions of recording artists until

1972. So with respect to pre-1972 U.S. sound recordings, it is difficult to determine who owns rights in the recording itself (as opposed to the music) because there is no central registry of such information before this date. This has direct impact on preservation in the digital realm, because the future preservation technique for analog recordings (including all those created before 1972) will be largely digital.

The importance of preserving the music and recorded sound created since the invention of recording techniques just over 100 years ago was recently confirmed by Congress in the National Recording Preservation Act of 2000. The legacy of recorded sound in peril includes the voices of native Americans speaking in tongues now nearing extinction; the songs of birds and whales; the oral traditions of folk artists; and, of course, music both commercial and noncommercial, classical and popular. Because of the fragility of analog tape, wax cylinders, acetate discs, and other media on which sounds have been recorded, reformatting is necessary to secure access to all forms of analog recordings into the future.

Indeed, as Samuel Brylawski notes in his paper on sound recordings, "ultimately, preservation reformatting will be required for all media upon which sound has been recorded, since preservationists acknowledge that there is no permanent format," either analog or digital (see Appendix 2, page 76). For a variety of technical and economic reasons, there is an overwhelming consensus that all preservation reformatting should and will be digital. Solving the digital preservation problem will be the only way to secure our aural heritage, both analog and digital, and there is little time to lose.

Under the National Recording Preservation Act, Congress has charged the Library of Congress with the hard work of developing standards for reformatting and preserving recorded sound, and because the future of audio preservation, even of the heritage of analog recordings, is digital, there will be a concerted effort in the near term to develop a system that will ensure the continued access to at least a portion of the aural wealth that abounds in libraries, archives, and recording studio vaults, in collectors' basements and attics.

As summed up by Brylawski, "The future of audio preservation is reformatting audio tapes and discs to computer files and systematically managing those files in a repository." Such audiovisual archives, also known as digital mass-storage systems, exist in Europe. "The well-planned repository presumes media obsolescence, plans for it, and, according to its supporters, frees the archive community of the futile search for an affordable permanent medium." But, he continues, "whether for lack of foresight or funding, libraries are not creating mass-storage systems for audiovisual works…. We face an extraordinary dilemma: at a time when a greater range of audio is available to more people than ever before, and the means are finally at hand to preserve those sounds for posterity, we stand the greatest risk of losing them" (see Appendix 2, page 80).

Digital Television and Video No longer are there questions about the primacy of broadcast television as a record of contemporary history and culture. Local, national, and global in scope and reach, television and video have transformed the ways that society views and understands itself. One need only recollect the role television played in the Gulf War or, more recently, the unfolding of the September 11 tragedy, to understand that contemporary history cannot be told without a full record of television.

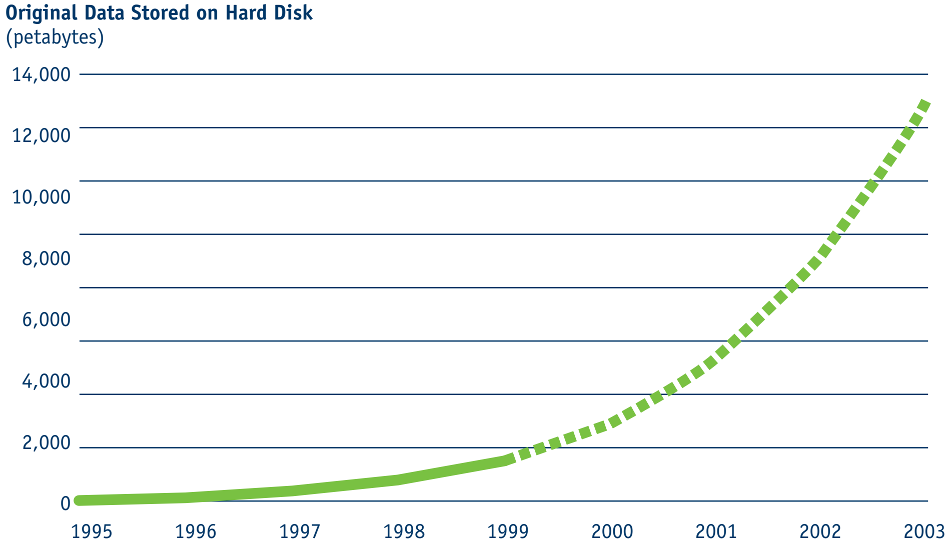
As with digitally recorded music, many of the salient difficulties of acquiring and preserving digital television and video are related to machine and media dependencies that affect analog as well as digital broadcasts. Technical experts such as Mary Ide, Dave MacCarn, Thom Shepard, and Leah Weisse on television, and Howard D. Wactlar and Michael G. Christel on video, see the move to digital as offering many solutions to problems inherent in both analog and digital television and video formats. Again, as in music, moving image archivists are familiar with a rapid and expensive pattern of technical innovation and obsolescence, with the constant need to refresh and reformat from one medium and machine dependency to another, with the need for massive storage systems, and with the dizzying succession of formats that demand new standards and are quickly superseded by others.

Digital television and video nearly always comprise a complex mix of elements—text, image, and audio, each with their own technical and metadata requirements for preservation—that demand very large scale storage systems. What one thinks of as shows are made of audio and visual elements that are stored and managed separately. In addition, a good deal of additional information needs to be preserved with the files, and there are digital rights management systems that need to be integrated into the large digital asset management infrastructure in an archives.

Digital moving image does promise to solve one serious problem in the television and film preservation communities, and that is of having to deal with the increasingly fragile medium of tape. “The notion of an ‘artifact-free’ method of distribution,” write Wactlar and Christel, “will have a great impact on preservation. Instead of moving digital information to tapes for distribution, data will simply consist of a file transfer to some temporary storage device, which might periodically be wiped clean.” However, they note, “failure to assign clear responsibility for preserving these broadcast materials may result in tremendous losses” (Appendix 2, page 93). Once again, the key to preserving the rich record of digital creativity and history will depend on a strong network of institutions willing to claim responsibility for preservation and the maintenance of a transparent system of tracking and accountability.

Web Sites The Web is best understood as a medium of information exchange that uses the delivery mechanism of the Internet. The Web functions as the most accessible bulletin board imaginable: Anyone can create a publicly available Web page, no prior authorization needed. It is much easier to “publish to the Web” than through traditional means, and the Web has consequently been populated by millions of creators who would not normally have access to other publishing outlets. As of January 2002, the Web comprised more than 550 billion public pages and linked documents. While it is not even a decade old, the Web is enormous and grows by 7 million pages

Figure 3. The Creation of Born-Digital Content Estimated to Nearly Double Every Year



Source: Lyman, Peter, and Hal R. Varian, "How Much Information," 2000. *www.sims.berkeley.edu/how-much-info/magnetic.html*

a day. At the same time, the mortality rate of Web sites is equally impressive: 44 percent of the sites available in 1998 were gone by 1999 (see Appendix 2, page 53).

"Saving the Web," then, is no more feasible nor desirable than saving the contents of everything that has ever been put to paper, to film, and to recorded sound disc across the globe. Nonetheless, it is very important for libraries to collect and preserve the content on the Web that is appropriate for the institution and for cultural memory, and this presents challenges on a scale as large as the Web itself. There are formidable technical challenges, common to all digital documents, of course. But beyond those, the problems start with capture of the Web. While Web harvesting, an approach used to create a "snapshot record" of the Web, can capture static HTML pages (the so-called surface Web), the deep Web, where much of the complex and culturally rich materials reside, is normally inaccessible to harvesting technologies. Even the surface Web is closed to harvesters in many cases because the materials require a license or other authorization to enter. And the average Web page contains 15 links. How does one define the boundaries of a Web site? This is among many questions that are emerging from early experiments in capturing content from the Web, and NDIIPP will engage these issues in its next phase.

Libraries have centuries of experience in selecting content that has long-term cultural value among an abundance of compelling material on paper. In this sense, the Web is familiar to librarians as a medium that contains text, numbers, and images, and that indifferently carries content as diverse as Shakespeare and screenplays, Ansel Adams photographs and family snapshots, manuscript maps of the Lewis and Clark expedition and AAA Trip Tiks for a road trip from St. Louis to Portland, nervous doodles and laundry lists, tax returns and top secret memoranda. Libraries have built collections of great merit, all the while having to make difficult choices of selection among such materials on paper. The challenge of selecting from the Web may turn out to have similar conceptual complexities, but the scope of materials is vaster. Most experts agree that identifying and capturing content of enduring value on the Web is the most formidable challenge of preserving it. It may be advisable to start with "published" content, such as online journals and other items that have known value in the analog realm for acquisition and preservation, and negotiate the deposit of the content (see above on e-books and e-journals).

Another area of promise is to capture government information that is on the Web where the online versions have superseded print versions. It is also important for American libraries to capture foreign sites (both official and unofficial or dissident sites) in Latin America, the Middle East, Southeast Asia, and other places where national libraries may not be positioned to do so at present. Libraries will need to revisit their collecting policies and develop shared agreements to reduce undesirable redundancies of Web capture.

The Library of Congress has been collaborating with the several entities that capture and preserve endangered and short-term sites, and this Web-based capture should be continued and extended broadly to other libraries; at the same time the legal issues surrounding copyright need to be clarified. While we can safely say that content on the Web is protected by copyright, can we determine whether or not a document on the Web is published or unpublished? The answer will have significant impact on a library's ability to capture, preserve, and provide access to that site.

Other Media These six formats are by no means the only ones that are new in the information landscape. There are others yet emerging that exploit more fully the interactive and customizable features of the technology. Chief among them are those that produce documents "on the fly" as a result of a query to a database. A prime example is the Geographic Information System (GIS), which "maps" a response to a query by matching data to spatial coordinates. It is not far-fetched to assume that within a decade or two, GIS will supersede the mass production of maps on paper. Pocket maps may be replaced by handheld GPS (Global Positioning Systems) and census maps replaced by massive datasets with a number of query interfaces that produce maps on the fly to answer questions about any number of demographic queries.

In this case, digital cartography is really best understood as an access tool, rather than something that produces a document, such as a map, or a specific set of data. What happens then to the map collections in research libraries? How will the Library of Congress and other collecting institutions "collect and preserve" maps? Will they instead acquire and preserve massive datasets and the software that people use to query the data? What kind of technical infrastructure, collection development and access policies, and user services must be in place to ensure future access to such geospatial materials?

Similarly, there are other formats that currently fill the stacks and storage shelves of libraries that are disappearing. The way that the telegraph came and went, so too are manuscripts going as people adopt word processing technologies. Correspondence is being replaced by e-mail. Libraries will need to partner with important actors and agencies to ensure that they are keeping correspondence documents in forms that can be accessioned. We do not know the full implications for libraries of these radical transformations in formats and presentation modes for digital information. All we can do is acquire some of these new formats and track closely what their custodial needs are and how users interact with them. A big focus of NDIIPP into the future will be not just a technology watch, but also a format and genre watch and a careful watch of users.

To develop rich and culturally significant digital collections in this environment, librarians and others must examine the implications of these changes to information and what they mean for such things as:

- Definitions of genres: What is a digital object and what are its boundaries?
- Dynamism of data: How does one select and curate digital objects built of dynamic data?
- Assessment of value: How does one identify enduring value?
- Intellectual property rights: How does one comply with the terms and conditions of use and payment when necessary?
- Mode of acquisition: What are the advantages and disadvantages of Web harvesting versus deposit of source file?
- Best editions: What should be deposited for copyright and in which format(s)?
- User studies: Who is using digital content and how? In what formats do they prefer it?

Transforming Roles

Who Is Responsible? No longer can preservation be seen as a "just in case" activity that takes place after distribution, in anticipation of some future use at an unknown time, as it has been with libraries and archives for centuries. Rather, it calls for active management of files from the beginning, and therefore a decision about preservation almost at the time of creation—a huge shift in the relationship between preservation and access that is little recognized and less understood.

Key technical factors affect the longevity of digital objects. A digital object must be created in sufficiently standard formats (that is, not idiosyncratic or proprietary) that it can move from one hardware/software configuration to another over the course of its life cycle. The digital object must be accompanied by metadata—the digital cataloging that describes the object, gives provenance information, specifies its file formats, and so forth. These requirements place a new burden on the creator and the publisher, distributor, or aggregator, who do not presently have to think about preservation when they shoot a film or send a manuscript to the printer.

As one participant said, much of what is most innovative and worthy of preservation is created by men and women without the resources to store and manage their output over time. They are too busy creating to become their own archivists. What incentives can we provide to musicians, writers, scientists, videographers, choreographers, architects, photographers, and others to assume the burden of preservation?

There are some new models of preservation that are emerging, however (see Appendix 4). There are academic disciplines that are managing their own pre-print papers (for example, the arXiv [*www.arxiv.org*] that serves the physics and mathematics communities) or their own datasets (Inter-university Consortium for Social and Political Research [ICPSR, *www.icpsr.umich.edu*] and PubMed Central [*www.pubmed central.nih.gov*]), and for-profit and nonprofit academic publishers (Elsevier Science, the American Physical Society) that are partnering with libraries in pilots to preserve electronic content. Film and television firms are developing in-house digital asset management systems designed to preserve content for reuse ("repurposing") for at least some limited periods of time that in many ways model longer-term preservation interventions. There are archiving services such as JSTOR for scholarly journals (*www.jstor.org*) and library service providers such as the Online Computer Library Center (*www.oclc.org*) and the Research Libraries Group (*www.rlg.org*) that are undertaking digital preservation services for their members. There are even individuals who represent a new breed of collector in the digital realm who are capturing readily accessible parts of the Web. The Library has identified a number of these new participants as potential partners going forward. Clearly, more are needed.

WHO PAYS? Preservation is expensive, and heretofore few institutions and businesses other than libraries and archives have undertaken it on behalf of present and future generations. We have not had to pay upfront costs for preservation before. We have always benefited from the actions taken by previous generations. But things have changed. Distribution of costs among the stakeholders, from creators and archivers to the users who benefit from both of their activities, will be a crucial sticking point in the development of any sustainable preservation infrastructure. As noted by Dale Flecker (see Appendix 2, page 31), some of the costs that must be considered include:

- notification or identification of content to be collected and preserved,
- creation of an archival version of the content if the access version is not sufficiently robust,
- creation of preservation metadata,
- storage, monitoring, and management of data in the repository,
- preservation actions taken on the content, and
- services to users and owners/rights holders.

The costs of many of these actions are unknown, and a major area of research in the near term must be cost modeling and the development and testing of business models. Many industry representatives consulted by the Library declared themselves "eager to align business and cultural needs" through cooperation on preservation. The key for them is to ensure protection of their current and future revenue streams.

Conversations at the stakeholder meetings and elsewhere explored issues related to valuing donated content and the use of possible tax incentives, a strategy that has been employed for cultural preservation projects in the analog realm. What is an appropriate method for appraising the value of digital assets that an owner may wish to donate to a library or museum, as donors do now with rare books, maps, photographs, artworks, manuscripts, and other cultural treasures? How do we develop the financial infrastructure that would encourage digital philanthropy?

There remains substantial ambiguity surrounding key economic issues. Since the development of tape- and disc-based digital storage systems in the 1960s, more than 200 storage formats alone have been deployed, with none lasting more than 10 years, necessitating massive migration of data from system to system (see Appendix 2, page 40). This puts content not aggressively managed every year at risk of sudden death. Moore's Law notwithstanding, a representative of a major research university library lamented, the cost of storage remains a line item in the annual budget. And costs are not going down.

The digital realm is one of change and uncertainty, and it is likely to remain so for the foreseeable future. Even the most astute businesspeople cannot forecast anything comfortably because change is so rapid that it is too difficult to develop viable business models. Moreover, as one adviser remarked, not only is the content ephemeral, so are the relationships between the information and the creators. How does one secure revenue? Attempts to "lock down assets" through encryption and other high-tech means may end up removing the asset from the marketplace, choking off revenue, and, ultimately, making the technical challenges of preservation overwhelming. It will be important to strike a balance between too much security and too little.

The Library was advised over and over, in this atmosphere of uncertainty and lack of tested business models and enduring relationships, that participants in digital creation and preservation need trusted third parties who will address issues of longevity when others have abandoned digital material. There should be fail-safe mechanisms to rescue digital content of high cultural and historical value that is in peril. The Library is experimenting with safe-harbor agreements with some rights holders that would ensure the safe transfer of digital content to the Library in case of a business failure. There may be so-called trigger events, such as the expiration of copyright protection, the imminent demise of a collection, the bankruptcy of the copyright holder, or other things that might activate a preservation intervention. This type of arrangement between owner and repository must be explored and expanded in order to find scalable solutions, and the Library can play a leading role in developing best practices that stakeholder communities endorse.

Time and again participants in the planning process expressed concern and some confusion about what libraries and repositories can and cannot do legally to preserve digital content, and particularly what the Library of Congress, as the recipient of copyright deposit material, could do. Therefore, the Library sought to clarify the impact of copyright management on preservation (see Appendix 6). One of the gravest implications of the present copyright regime for long-term access to digital content is the sheer fragility of the information. Given how short the life span of digital content is and the length of copyright protection (to life plus 70 years), we face the very real prospect that the nation's most valuable intellectual and cultural content—that protected by copyright—will not pass into the public domain for more than a century.

At present, that places the burden of preserving that content on the rights holders, who may be unaware of the implied cultural mandate or may be ill-positioned to guarantee the responsibility over the long term. This is an important matter that must be addressed before Congress can be assured that its charge—to develop an effective infrastructure for the preservation of significant digital content—can be effected. While the specifics of the rights regime lie outside the scope of the Library's NDIIPP mandate, the responsibility to elucidate the effects of rights protection on the integrity of the historical record is clearly an important component of its charge.

Current Digital Preservation Efforts

To understand further the very challenging issues of scope, collection development, and roles of various players in the preservation landscape, it will be helpful to review what was learned about the activities of other organizations in the field of digital creation and preservation.

U.S. Government

The National Archives and Records Administration (NARA), as the archives of the U.S. government, is responsible for safeguarding the records of all three branches of the federal government. In recognition of this mandate and significant changes in the federal records management environment, in which most records are now created digitally, NARA has embarked on a three-pronged approach to further the management and preservation of electronic records. NARA is in the second year of a multiyear project to redesign its records management program to meet the challenges of today's government records. By creating mutually supporting relationships with agencies whereby NARA's records management program adds value to agency business processes, records will be managed effectively for as long as they are needed, and records of continuing value, particularly electronic records, will be preserved and made available for future generations.

A key effort of the Archives is the Electronic Records Management E-Government Initiative, one of 24 e-government initiatives sponsored by the Office of Management and Budget. This initiative will provide the tools that agencies will need to manage their records in electronic form, addressing specific areas of electronic records management where agencies are having major difficulties. This project will provide government-wide guidance on electronic records management and will enable agencies to transfer electronic records to NARA in a variety of data types and formats so that they may be preserved for future use by the government and citizens.

NARA, the managing partner for this initiative, is working with several other agencies to integrate electronic records management concepts and practices with comprehensive information management policies, processes, and objectives to assure the integrity of electronic records and information. It is also focused on employing electronic records management to support interoperability, timely and effective decision-making, and improved services to customers. Finally, NARA will provide the tools for agencies to access electronic records for as long as required and to transfer permanent electronic records to NARA for preservation and future use by government and citizens.

The Electronic Records Archives (ERA) Program is NARA's strategic response to the long-term preservation of electronic records. Its goal is to enable NARA to preserve and provide access to virtually any type of electronic record created by the federal government. ERA must be inherently responsive to known problems associated with electronic records and adapt to new and unpredictable challenges and opportunities that will arise as information technology and its application in government continue to evolve. The resulting system will have three primary characteristics: It must be persistent; it must preserve authentic records; and it must be scalable.

To develop and build ERA, NARA is collaborating with government, industry, academic, and international partners who are leading the way in developing the next-generation national information structure. The system will be scalable both upward to meet NARA's exponentially growing workload, and downward so it will be useful to smaller archives, libraries, universities, and businesses. The technology developed for ERA will provide a common framework for all agencies in managing their electronic records for as long as needed in conducting their business. ERA will maximize the use of government and commercial off-the-shelf components, be developed in a series of fully funded usable increments, and use performance-based contracting methods. Currently, the ERA Program is continuing research and development, documenting system requirements, completing an analysis of alternatives, and developing a business case analysis.

In addition to NARA, the Library also interviewed representatives of some of the other federal agencies concerned with preserving digital records, in particular the Department of Commerce, National Institute of Standards and Technology (NIST), the White House Office of Science and Technology Policy (OSTP), National Agricultural Library (NAL), and National Library of Medicine (NLM).

The interrelationships among mission, organization, and technology surfaced in the interviews and review of materials provided by the agencies. Representatives from other agencies emphasized the technological challenges and the importance of maintaining internal technological capability. This expertise is not required to build and maintain proprietary systems but is essential to identifying and evaluating the utility of systems and tools that may be available or customized to meet the specialized needs of libraries and archives. One of those interviewed advised that the Library should either ramp up its internal capabilities and technical infrastructure or find a reliable partner. Representatives were also emphatic on the importance of a clearly defined mission.

Mission shapes organizational relationships, and those relationships, in turn, affect how technologies and technological solutions may be evaluated and crafted. Several agencies are accumulating substantial hands-on experience with data accessioning and validation from multiple sources as well as with metadata, bibliographic utilities, and so on.

Other topics specific to NDIIPP that surfaced consistently in these discussions include: the importance of information security; the utility of a test bed in one of the nonprint formats (for example, music), which would be hard enough to push the technology but not so hard as to be intractable; the importance of meeting Congressional needs relating to information capture and the business of government (for example, archiving Thomas, a legislative information Web site [*thomas.loc.gov*]) and of addressing long-term preservation of copyrighted materials; and the importance of access as a feature of a digital preservation strategy. Finally, several agencies have programmatic commitments to the professional development of archivists and librarians, including training that is tailored to digital resources.

The Library of Congress will benefit from the examples that NARA, NLM, and NAL, among others, have set in developing organizational responses to the need for a distributed information architecture.

U.S. Libraries

The survey of members of the Association of Research Libraries (ARL), undertaken in October 2001 at the request of the Library and in support of the NDIIPP planning, found that most of the 67 libraries that responded are currently preserving or intending to preserve a mix of born-digital and digitized materials that were created by the libraries themselves. The libraries cope with a range of objects—dissertations, online serials, various “Web collections,” social science-economic data, electronic student records, and so forth—in which the data are in a mix of relatively common formats. A handful of libraries have policies and practices in place; others are developing them either in isolation or collaboratively. Many are still working out best practices and looking to other libraries, including the Library of Congress, to model those practices and develop standards.

These libraries identify as critical the need for staff training and development. The Library has been a resource for development and dissemination of best practices in preservation of analog collections, and it should be positioned to assume that role in the digital realm.

Along with the Library of Congress, there is a set of leading research libraries looked to for development of dissemination of best practices in the digital realm, the members of the Digital Library Federation (DLF). Twenty-four nonfederal members of DLF were also surveyed (see Appendix 4). Respondents indicated collection/preservation priorities that were consistent with the priorities adopted by ARL members, namely:

- institutional records,
- digital materials received as part of heterogeneous archival collections,
- locally hosted e-journals, and
- materials created or collected by local faculty (working papers, databases, converted textual documents) and other local materials (for example, dissertations, local Web sites, student portfolios, and learning or classroom objects).

The number of Digital Library Federation activities relevant to NDIIPP mentioned by respondents is striking. Four institutions are involved in an e-journal archiving program that partners with commercial and nonprofit publishers. There are, in addition, repository systems oriented toward digital preservation available or under development at the several major public and private universities, with one even pursuing the construction of a service to provide long-term archiving for external depositors. Finally, one research library is engaged in investigating ways of evaluating preservation risks for Web-based resources.

The survey indicates that there is a set of research libraries that would be natural partners to the Library of Congress in the creation of a national cooperative plan for digital preservation. These institutions have already identified sets of digital resources for which they expect to take responsibility, are creating the infrastructures to support digital preservation activities, and have expressed a willingness to work with the Library in this domain.

Private Sector

Yearlong consultations with representatives from the private sector added another dimension to the overall environmental review. This sector agreed upon the need for the Library to be a convener and facilitator in this new information space. This is especially critical in the arena of standards, for, as one publisher noted, it may be in industry's interest to develop a common standard, but he feared that others who felt left out would accuse the industry of collusion. If the Library were to endorse and adopt some standards, it would increase their durability and credibility. There is also a lack of common practice among publishers and media companies about what gets preserved in-house. Some save just text files, others also save graphics and illustrations, but do not save drafts or correspondence. Some recording companies save graphics as well as new output, but others do not. Under these circumstances, what is happening to the raw history of creativity? Publishers and studios do not behave like preservationists—and will not—but they understand the importance of some third party taking responsibility for saving the historical record.

Yet another reason for the general approval of NDIIPP among the creative industries is their fear that back-up systems might fail: They would welcome a disaster-recovery backup that would be one part of that preservation infrastructure. It was even noted that industry needs access to outmoded software just as libraries do, and one individual suggested that, were NDIIPP to provide for archiving of software for access formats, his firm would pay a fee to use them.

There was a clear sense among the private sector representatives that what gets preserved for posterity is something different from what they are best positioned to preserve within their own industries. Saying that they would like to align business and cultural interests means cooperating with the heritage sector, not taking on those responsibilities alone. And, finally, it was clear that innovation in technology will continue to come from the private sector, for, as someone from a technology company noted, libraries are seldom able to offer competitive compensation for highly skilled technologists.

International Libraries

A survey of the leading international developments in digital preservation commissioned for NDIIPP reveals that the problems discussed throughout this report are shared across the globe (see Appendix 5). No library environment abroad is comparable to that in the United States, in part because of the de facto, not de jure, nature of the Library of Congress as the national library; its different governance and funding; the unique role of the Copyright Office in the Library; and the Library's inclusion of essentially all major nongovernmental information formats, from maps to moving images. Nevertheless, there are many areas of common experience, concern, and potential collaboration among libraries across the globe. Chief findings of the survey include:

- the technology is revolutionary, not evolutionary,
- collecting patterns are shifting from purchase to license,
- digital preservation is poorly funded, often relying on soft funds,
- solutions require collaboration and coordination among many partners,
- collaboration is difficult,
- it is easier to collaborate on research than on policies (because external funding encourages collaboration on research),
- coordinated collection development is advisable but difficult to achieve and sustain,
- nowhere do comprehensive legal provisions for archiving of digital publications exist, and
- the scope of collecting is affected by the information explosion.

Finally, and crucially, "digital preservation relies substantially on the collaboration of key stakeholders outside the memory institutions and the professional sectors they represent. An important part of digital preservation activity as a public good is funded either from public funds by government or through private benefactors. However, awareness of digital preservation issues among the public, government, and other key stakeholders remains low" (see Appendix 5, page 133). There is nothing comparable to the Congressional action and funding taken on behalf of digital preservation abroad, and NDIIPP has generated great enthusiasm overseas for the attention it draws and resources it musters to this issue. Areas of potential initial collaboration with the United States include:

- technical research,
- standards development,
- collection development, and
- development of shared services needed by repositories.

National libraries indicated that there is significant scope for international collaboration and potential cost benefits in developing preservation services on a shared basis.

Implications for the Future

Future Environment

The extensive planning and consultation in this first phase of NDIIPP included: stakeholder meetings and interviews that addressed the issues affecting their enterprise's future; a parallel effort at conceptualizing the environment in which digital preservation occurs; and advice from expert technologists and legal scholars about possible and probable trends in their areas of knowledge. These activities yielded much rich information that then served as the foundation for developing future scenarios, in which participants identified the driving forces that may shape the future environment in which preservation will play out. Given the complexity of the digital environment—one in which preservation of and access to digital content involves a significant number of actors working in a network—it becomes especially important to understand how the nature of future events could affect the strength and flexibility of that infrastructure to withstand various disturbances. Individuals and institutions consulted by the Library were asked to identify what outside factors are most likely to affect the viability of the digital preservation infrastructure, and what kind of infrastructure would thrive under even the most adverse circumstances.

Among the driving forces in the near-term future that will affect whatever infrastructure is developed to preserve digital content are:

- technological change,
- national and global economics,
- national and global politics,
- national security,
- copyright, and
- regulatory changes (such as encryption, classification, privacy).

Several executives in the publishing and media industry pointed out that issues of trust were paramount in determining the future. The real barriers to their digital enterprises' success at the moment are the lack of business models that guarantee that an investment in the new technologies would pay off. Until the time that such models emerged and were tested in the marketplace, content creators and distributors would feel economically vulnerable, even when they were not. Competition could drive out the collaboration all acknowledge is necessary at this point.

As long as trust remains a scarce commodity in the digital landscape, each new technological innovation is likely to have unintended and often negative impacts on the techniques available for archiving. This was remarked in the National Academy of Sciences report on copyright, *The Digital Dilemma,* and it remains a serious driving force in any future scenario that must be considered.

There is no way to predict how the future will unfold, of course. The Library recognizes the need to track the evolving circumstances at home and abroad that can have decisive yet unanticipated effects on the preservation mission. What is described below is the best thinking at present about how these various influences may play out and affect the digital preservation infrastructure and the role the Library can and should play in it.

Plausible Scenarios

Scenario planning is a process that enables organizations to learn and adapt to change by creating several possible alternative futures. By generating divergent scenarios, organizations can think beyond any preconceived notions of the future instead of trying to find a single "right answer" for today that risks being obsolete by the time the future arrives. The Library, as a cultural heritage institution that is conservative by nature and has succeeded for generations by creating and adhering to standards that work well in the analog realm, is particularly vulnerable to the temptation to think in one direction only about the future. The scenario planning exercise pushed the Library and its partners to consider the effects of driving forces over which the preservation community has no control, such as technology advances, the global economy, international instability, the digital rights regime, and others.

Scenario planning was staged over several meetings. The first session required participants to think most broadly, expand the parameters under consideration, and identify three different scenarios to investigate. Subsequent sessions drilled down to what the consequences of the scenarios would be like on the ground. The meetings brought together a mix of experts from within the preservation community, broadly defined, and digital librarians, computer scientists, media and publishing executives, and engineers and technologists.

Preliminary findings of the first scenario planning session in February 2002 further confirmed the sense taken away from the stakeholder meetings the previous November: that is, the problem of digital preservation is urgent; can be addressed only by a distributed, decentralized, and networked infrastructure; and collaboration among all stakeholders to share the responsibility for the fate of digital culture can and should be catalyzed by the Library of Congress as representative of the public interest.

There was consensus that the Library has a critical role to play in all three scenarios addressing the future of digital preservation, but they vary significantly in scope and influence. They are:

- the Library focusing on preserving the most critically endangered materials, acting in an environment without many partners,
- the Library playing a clearinghouse role in a landscape populated by several distributed repositories with well-defined preservation responsibilities and in need of close coordination,
- the Library, operating in an environment with maximum participation of stakeholders in preservation, facilitating the development of technical standards, intellectual property agreements, and business models that are robust, diverse, and serve to undergird a "peer-to-peer" system of comprehensive preservation.

In no scenario does the Library itself "save everything" or collect universally. The issues involved in replicating in the digital arena the comprehensiveness of the Library's physical heritage assets are far too complex to envision in the foreseeable future. Significant differences in key investments and competencies are required of the Library in each scenario. As the universe of preserved collections increases, moving from the first scenario of circumscribed activities to the third of large-scale preservation by many communities, the centrality of the Library as a repository decreases, and its role as facilitator, honest broker, and strong focusing actor increases.

In many ways, it is best to look at the three scenarios as a progression of evolving roles and responsibilities within changing environments. The first scenario envisions a bad economic and/or regulatory climate when resources flowing in for the support of the public good are scarce and when the intellectual property rights regime or national security regulations also restrict the flow of information resources in the marketplace of ideas. In this scenario, the Library is one of only a few key institutions in the network of preservation actors.

The third scenario, which envisions the best of economic times, occurs when confidence in both the public and the private sectors is high, trust among players in the information landscape is also high, and when people are motivated to cooperate out of their enlightened self-interest. In these circumstances, the Library will continue to be responsible for the collection and curation of the digital objects that fall under its collecting policies. But it will also play an expanded leadership role because there will be a wider group of preservation parties coming from increasingly disparate communities. These groups will need an honest broker who defines best practices for archiving different formats, who works closely with industry in test-driving new preservation technologies and end-user services, and who keeps the institutions and third parties in the digital preservation infrastructure connected and informed. One Internet entrepreneur optimistically speculated that under these circumstances, the Library could even help industry by ensuring the privacy and integrity of their information assets and thereby helping to generate revenue, which could in turn be used to support the digital preservation infrastructure through tax incentives or other mechanisms.

With great consistency the Library heard again and again that in this new information environment, trust is the necessary force to build a resilient digital preservation network and keep it together, and that the development of trust among players is a key role for the Library. As the preservation safety net grows and strengthens, the Library will need to find other libraries, repositories, and proven service providers that can play the same role of trusted third party, since centralization of deposit and access at one site, even the Library of Congress, is neither feasible nor desirable. Especially in the early development phases, trust must be there for the participants to feel comfortable trying out their new roles. This will be a crucial role for the Library in the following phases of implementation of NDIIPP. Experts asserted unanimously that trust cannot be built or guaranteed through technological warrants alone, though they have their roles in authorizing access and other matters. Trust adheres in people and in organizations, and the Library is widely viewed by stakeholders in its traditional role as third party that would act for the public good, as an extension of Congress. In some ways the Library's role as a trusted third party means it is scripted into certain roles under all conditions, but that role becomes especially important in periods of uncertainty.

In each possible future, the Library will collect and preserve critically important digital content; will be a portal to material in high demand but not take responsibility for preserving it; and will be a key enabler of the work of others. Again, this result echoed the discussions held in November among the stakeholders (see Appendix 2, pages 17–24). In the areas of technology, intellectual property rights and access, payment and business models, the Library was seen as being one of a number of key players, but not the most powerful. In the fourth area of concern—roles, scope, and collection development—the Library was cast in the leading role in everybody's vision of the future. These issues seemed particularly hard to sort through for many because of the overwhelming scale of information production made possible by a technology that we do not yet fully understand.

Collection development—who collects what for whom in what format—is especially fraught with ambiguity and some frank anxiety. Unlike the contentious issue of copyright, for example, there are no laws that can be enacted or revoked to give parameters for good behavior. Yet because of the ephemeral nature of the data, if we make mistakes about collection development now, we are unlikely to get a second chance to collect in the future. A digital file cannot sit neglected on a bookshelf for 200 years before someone discovers its value. By then it will be corrupted or trapped in an obsolete software encoding. All who were consulted during this phase of planning declared that collection development is clearly in the portfolio of libraries, chiefly because we must rely on experience and judgment for making decisions about selection. Those who took part in the scenario planning were especially keen to see the third, most inclusive scenario, come into being, for they want to ensure access to the widest possible field of resources to those generations who follow. Collection development is clearly an area that needs to be addressed sooner rather than later, even if we cannot expect easy answers.

Again, in each possible future for preservation, individuals and companies expressed general willingness to try out new roles because there is a consensus that the problems are real. For example, entertainment media and publishers said that repositories do not need to be “dark” (that is, without access to anyone), but companies do need to control the terms of access for purposes beyond those served by preservation. At the second scenario planning workshop in late April 2002, many of the participants expressed interest and willingness to undertake some pilot projects under the auspices of NDIIPP in order to test out new technologies, new approaches, and new roles. They expect the Library to facilitate those pilots.

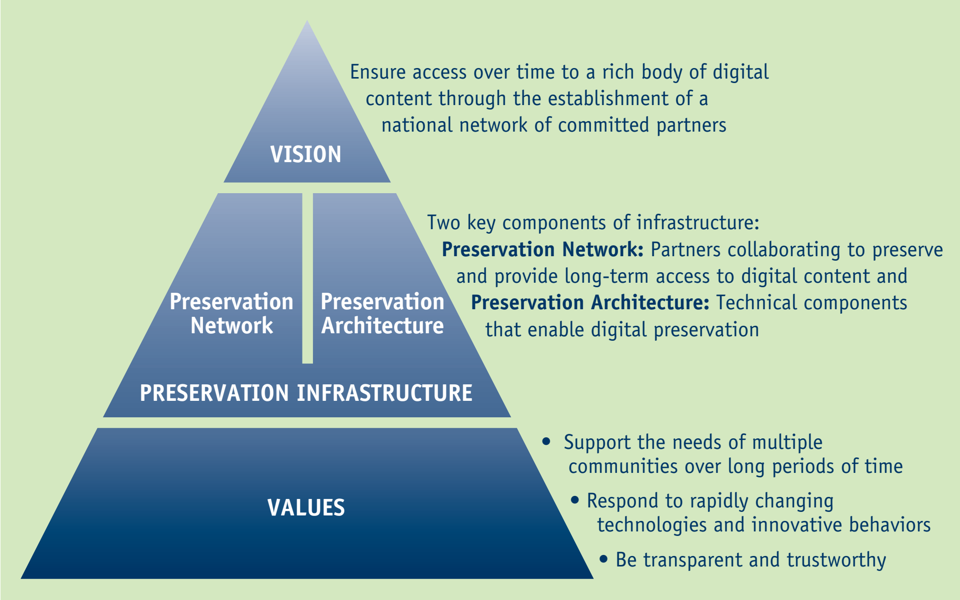
As noted at various stakeholder meetings and workshops, the Library should play the role of “stimulator of initiatives and a consumer of successful technologies.” The Library will need to leverage both federal funds, such as those at NSF, and private funds to build knowledge and bring those with expertise to bear on germane issues.

NDIIPP Strategic Direction and Plans for Action

A year of consultation and information gathering has confirmed widespread agreement on the urgency of addressing the preservation of digital heritage. There is consensus among stakeholders that the problems are complex and occur on many levels—technical, legal, social, economic, intellectual—and therefore the solutions must be equally nuanced and multivalent. The preservation infrastructure must be responsive to new findings that will be generated by the research community but developed and operationalized in other arenas.

It is also clear that there is a continuing need on the part of all partners in preservation to build and support a strong communication effort that helps raise awareness of what is at stake, that catalyzes actions among preservation partners, and encourages

Figure 4. Strategic Direction



VISION
Ensure access over time to a rich body of digital content through the establishment of a national network of committed partners
Preservation Network
Preservation Architecture
PRESERVATION INFRASTRUCTURE
Two key components of infrastructure:
Preservation Network: Partners collaborating to preserve and provide long-term access to digital content and
Preservation Architecture: Technical components that enable digital preservation
VALUES
• Support the needs of multiple communities over long periods of time
• Respond to rapidly changing technologies and innovative behaviors
• Be transparent and trustworthy

all who benefit from access to digital content to support preservation. For without preservation, there will be no access.

Strategic Direction

The vision of NDIIPP is to ensure access over time to a rich body of digital content through the establishment of a national network of committed partners, collaborating in a digital preservation architecture with defined roles and responsibilities.

The creation of such a system will occur incrementally, because of the complexity of the challenge and the number and diversity of actors involved. To realize this vision, the Library of Congress will take actions that are:

- *catalytic:* investing in existing strengths, leveraging public and private investments, and stimulating research and development where needed,
- *collaborative:* engaging willing partners and key stakeholders in areas of mutual interest and expertise,
- *iterative:* learning from the initial planning and fact-finding to inform subsequent actions and investments, and continuing to feed results back into the chain of research, testing, and development, and
- *strategic:* addressing a broad spectrum of issues in technology, collection development, infrastructure and organization, intellectual property, technical standards, and other key components of the preservation network through a balance of early short-term and long-term actions and investments.

This digital preservation infrastructure will neither be built quickly nor be completed at a foreseeable end date, as the infrastructure is by design a dynamic and ever-adapting system.

Box 2. Expected Outcomes

Through the execution of the NDIIPP initiative, the Library expects to have reached a number of outcomes. There will be:

- a clearer definition of the roles and responsibilities of partners in the preservation network,
- the establishment and deepening of relationships among key partners in the private and public sectors, including international institutions,
- the clarification of intellectual property issues that impede preservation, together with recommendations to address them,
- the creation of an advanced design for the digital architecture,
- the identification of a next set of investments to advance NDIIPP goals,
- the encouragement of an advanced research agenda for preservation of digital content, and
- raised awareness among key stakeholders and the general public about the challenges and opportunities of digital preservation.

The Library's plans for action over the next three to five years of NDIIPP will comprise investments and activities that will preserve digital content, build a resilient network of digital preservation partnerships, and begin development of the digital preservation architecture to support and enable these goals.

Development of the Digital Preservation Infrastructure

The digital preservation infrastructure must be flexible, responsive to innovation without becoming anarchic, and be accountable and transparent before all the stakeholders. Specifically, the national preservation infrastructure must:

- support the needs of multiple communities over long periods of time,
- respond to rapidly changing technologies and innovative behaviors, and
- be transparent and trustworthy.

These are the core values that define NDIIPP's plans for action. NDIIPP will start building a preservation network of committed partners around a collaborative preservation architecture with defined roles and responsibilities through investments and actions on two broad fronts:

- core capacities crucial for collaboration among institutions (the shared knowledge, expertise, skills, and consensus regarding essential areas of concern that support the digital preservation framework) and
- the digital preservation architecture needed to operate the network.

Core Capacities of the Digital Preservation Network

Major actions in developing a strategy for collecting, preserving, and ensuring rights-protected access to digital content will be a series of collaborative initiatives, catalyzed by the Library of Congress. These actions take advantage of opportunities and demand a high-level commitment of certain types of capital—primarily the time and expertise of groups with curatorial, legal, financial, economic, sociological, and other analytical skills. Among the issues identified through NDIIPP consultations and research that demand action are:

- selection and collection development,
- intellectual property,
- business models,
- standards and best practices, and
- communication and outreach.

Selection and Collection Development

Among collaborative actions in the area of selection and collection development to be undertaken are:

- developing cooperative agreements between national libraries about the scope of collecting national materials, both Web-based and licensed,
- developing cooperative collecting agreements with libraries, archives, and other collecting institutions in the public and private sectors,
- convening experts to develop guidelines for assessing content for enduring value,
- convening experts to examine curatorial best practices for selecting dynamic objects, such as Web-based journals, GIS materials, interactive objects, and other genres,
- defining the boundaries of Web-based content for preservation purposes, and
- reviewing collection development policies, including those for best edition necessary for mandatory copyright deposit, in order to bring them up to date for digital materials.

Selection of digital content begins with identifying material for accessioning, assessing its long-term value, ensuring its completeness and authenticity, determining the most appropriate formats for acquisition, and considering the impacts of various preservation strategies, such as migration or emulation, on the longevity of the digital object. Selection also includes the development of institutional collection policies (and making those commitments widely known), and then defining coordinated collecting and preserving responsibilities among key stakeholders.

The expertise of curators found in libraries, archives, museums, and other collecting institutions must be brought to bear on the variety of new forms of expression in the digital realm (streaming audio and visual content, for example), capturing them as quickly as possible and assessing their short-term use and long-term value, and calculating their preservation requirements.

In addition, the research agenda for selection identified by NSF and the Library includes the development of cost-benefit models to inform decisions about preservation formats and standards, choices of preservation strategies (normalization, migration, emulation), and the costs and benefits of various levels of description and metadata. These are all important criteria that will powerfully influence collecting decisions, especially in the realm of complex digital objects in which matching a preservation strategy to an object requires a keen understanding of the most valuable features of that object. Is a digital video clip important chiefly for its informational value, as it would be if it were a news report or a video of a choreography? Or are the aural and visual elements more important to preserve through (more expensive) high sampling rates and uncompressed storage, as would be the case in digital art? Each type of digital content must be defined for its long-term value and will have its own preservation requirements.

Finally, the Library of Congress will undertake, in consultation with national and international collaborators, a definition of what constitutes a best edition, that is, what should be deposited for copyright, in what format, and under what conditions. There is also a need to explore the deposit of copyright materials at authorized third-party agents of the Library, as is done in European countries, to handle the unprecedented volume of material that is copyrighted and worthy of long-term preservation.

Assessing the appropriate formats to accession, deciding how to preserve and provide access, determining which techniques are appropriate for preserving objects—these are crucially dependent on anticipated use. How users interact with a certain database, for example, depends on whether they are geographers, genealogists, biologists, or a seventh-grade history class. That said, there has been little applied research to date about the use of digital heritage content, in sharp contrast to such studies in the private sector. NDIIPP should engage these issues in the next phase as well.

Intellectual Property

An overview of the complex effects that copyright has on the fate of digital preservation can be found in Appendix 6. Among recommendations cited in the study that will be acted on, as appropriate, are:

- investigation of the options and authorities necessary for the Library of Congress to preserve digital content captured on the Internet,
- investigation of what are acceptable methods of access to digital content for educational purposes within a library setting,
- investigation of the implications of mandatory deposit for digital content,
- investigation of the implications of various security and protection devices for preservation, and
- development of a better understanding of the international context of copyright, jurisdiction, responsibility, and reach of applicable law, possibly in cooperation with other national libraries and multinational publishing and media industries.

Intellectual property issues were cited again and again in the consultations the Library conducted as among the most challenging of all the barriers that now impede the progress toward preserving digital heritage. One of the guiding principles of this initial fact-finding and planning phase of NDIIPP has been to bracket the concerns that owners and users have about access in order to give preservation priority in the discussions. But that separation between preservation and access cannot be maintained forever, in large part because crucial decisions about what to preserve and how depend entirely on *for whom* the digital assets are being preserved. The Library of Congress, home of the U.S. Copyright Office, is uniquely positioned to address the continuing need for rights-protected access to information.

Business Models

Among actions in the area of business models to be undertaken by the Library and its partners are:

- identification of incentives for institutions to undertake preservation,
- identification of incentives for creators to deposit content,

- development of metrics for costs and benefits of digital preservation,
- development of metrics for appraising digital works for purposes of insurance and tax, and
- development of model safe-harbor agreements for those materials that are preserved by commercial entities or others that may not be best positioned to ensure longevity.

Research and testing of economic models and policies is a crucial need for the development of a sustainable digital preservation infrastructure. Some of the key research areas include: the costs of acquisition and preservation; financial and other incentives for creators to deposit their work in a repository; the costs of standardization of formats to optimize preservation; and who should bear these costs at which stages of the digital object life cycle.

Because incentives for long-term preservation of digital information vary greatly among the communities creating and using digital assets, a rich menu of incentives that would encourage organizations to develop digital archiving capabilities, build repositories, provide archiving services, and to create content in ways that facilitate its long-term preservation should be modeled and tested. "A variety of mechanisms warrant investigation, including direct public subsidies, tax incentives for placing content in the pubic domain prior to the expiration of copyrights, philanthropic donations, and market mechanisms that provide for cost recovery or revenue streams to support the repository" (see Appendix 7, page 216).

The Library and its partners should identify a range of incentives that may encourage creators to deposit digital content in a repository as well. This is an arena in which we already know trust is a deciding factor, but we have little empirical evidence about how trust is engendered and maintained through past performance, up-to-date security systems, transparent behaviors in enforcing rules and agreements, and other factors possibly not yet identified. Depositors must have a very high level of trust in a repository, based on sophisticated use of security technologies, a track record of sound performance, and consistent application of rules and agreements.

The Library was advised by many stakeholders that it will be important to craft safe-harbor agreements in which proprietary digital assets can be maintained by the owner and yet dependably transferred to a trusted repository. Understanding what constitutes a trustworthy repository will be important to develop such agreements. Finally, much research is still needed in the area of metrics, since we presently lack reliable and objective ways to measure the costs, benefits, and value of digital content.

Standards and Best Practices

Activities that the Library will continue to lead or initiate are:

- coordinating and documenting standards that support key preservation services, such as metadata and persistent identifier schemes,
- fostering research and best practice recommendations for formats and encoding schemes,

- fostering research and development of strategies, such as migration and emulation, that will ensure sustainability of digital content, and
- developing a communication strategy to track technology changes and their impact on preservation.

Standards for data formats, data models, metadata, and other aspects of digital information are commonly cited as essential to collaborative partnerships engaged in digital preservation. Longevity of digital data and the ability to read those data in the future depend upon standards for encoding and describing, but standards change over time. Research on the evolution of standards is required to understand the impact of standards changes on long-term preservation methods and practices. The Library is the maintenance organization for descriptive metadata standards, a role it has played in the realm of print media for cataloging standards. As it leads a national digital preservation effort, the Library must monitor relevant standards not only concerning metadata, but also data formats, or the ways that digital information is encoded.

The topic of what types of metadata are necessary to support long-term preservation continues to be foremost in discussions of standards. Some metadata standards for digital materials are emerging. The Library should continue to facilitate consensus-building on digital preservation metadata among the diverse communities of creators, libraries, and archives.

During periods of emergent standards, best practices serve the critical function of guiding practitioners in their decision-making. The Library can continue to support and play an active role in the many existing research and communication relationships that help define best practices for preservation.

Communication and Outreach

Outreach activities targeting professional and public audiences include:

- maintaining the NDIIPP Web site *(www.digitalpreservation.gov),* featuring current information on the program's status,
- outreach to professional groups through participation in professional meetings and contributions to professional literature, and
- outreach to the public through print and Web-based general interest publications and through the broadcast media.

The aim of the communication and outreach program is to build a national constituency that will support NDIIPP and become active participants in preserving digital heritage.

The Library of Congress and its partners will engage the library and archival community, the business world, the creative community, and the general public in an effort to communicate the importance and urgency of preserving digital heritage. This outreach program will also encourage stakeholders to become active participants in the public conversation about this critical issue and to heighten awareness that preservation in the digital age must be considered at the time of creation. Preservation cannot be an activity relegated to the expertise of libraries and archives, but rather must be seen as intrinsic to the act of creation.

Digital Preservation Architecture

The areas of engagement discussed above are designed to start building a strong network of committed preservation partners and inform development of the digital preservation architecture. A parallel set of actions and investments will be directed toward developing the digital preservation architecture that will support that preservation network. In accordance with NDIIPP's strategic vision, the architecture's design principles support the needs of multiple communities, respond to changing technologies, and operate in an open and transparent manner.

The digital preservation architecture that supports the NDIIPP will specify the overall structure, logical components, and logical interrelationships of the system. It is especially important that the architecture design for the digital preservation infrastructure aid in simplifying the complexity of the environment, account for the broad capacities required to build a resilient and multidimensional infrastructure, and enable a wide range of stakeholders with different economic requirements and business models to participate in an integrated way.

In building a digital preservation infrastructure based on this architecture, the Library will:

- convene a design group to further develop the components of the preservation architecture,
- solicit proposals to test and model components of the system, and
- evaluate project outcomes to inform a next generation of implementations.

The next phase of NDIIPP will take the digital preservation architecture described below further, both by undertaking additional conceptual work with a newly convened architectural group and by applying the principles outlined here in applied experimentation with partner stakeholders during the next phase of NDIIPP. What follows is a summary of the proposed architecture and its design principles (see Appendix 9 for full details).

Box 3. Preservation Architecture Design Principles

The NDIIPP digital preservation architecture must:

- support relationships between institutions,
- allow questions of preservation to be handled separately from questions of access,
- be built modularly, using existing technology and efforts where possible,
- be able to be assembled over time, rather than needing to be built all at once,
- be upgradable in pieces, without disrupting the whole system, and
- be specified using broadly adoptable protocols.

The six design principles were developed in the initial planning phase by experts in digital libraries, computer science, and systems design. These principles collectively support the values of transparency, collaboration, incremental development, stability, flexibility, heterogeneity, and innovation. These principles suggest a flexible approach based on modular development in order to accommodate these seemingly disparate values while ensuring overall coherence by engaging many different stakeholders to work on different pieces of the architecture.

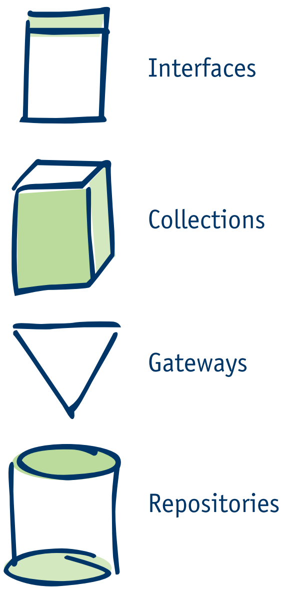
Consistent with the design principles described above, the group of technical experts have articulated a four-layer preservation architecture that assumes that NDIIPP will be built over time, and that its construction will involve both public and private institutions as well as the Library. It also assumes that the digital preservation infrastructure will never achieve stasis, but will instead evolve continually to integrate new forms of hardware and software, easily integrate new preservation partners, and preserve digital material of new formats and types.

In order to accomplish these ends, the architecture proposes four layers, with each layer comprising a different set of functions, governed by a related set of rules for use. These layers and their interconnections are designed to allow the preservation community to customize the architecture to its particular needs, and to make it possible to adjust the architecture as those needs change.

The proposed architecture has four layers:

- a *Repository* layer, for the long-term storage of digital data,
- a *Gateway* layer, which provides protection and control for the Repositories,
- a *Collection* layer, where agreements and decisions about the acquisition, access, and context of preserved digital materials are made, and
- an *Interface* layer, where those materials that patrons are allowed to access are made available.

Figure 5. Four Layers Between People and Bits



This architecture creates an unbundling of the functions currently associated with libraries and other organizations such as business archives and museums. The separation of people from the "bits" that make up digital objects illustrated in Figure 5 is analogous to a library patron seeking to use a rare book kept in a secure vault. Instead of turning to the card catalog to find the entry for the book (the bibliographical record at the

Collection layer), then asking the librarian (at the Gateway layer) to retrieve the book from the vault (the Repository layer), the user is now able to have access potentially to many digital objects in many different repositories, with the crucial cataloging, gateway, and repository functions enabled through the network. No one layer has the all the functions of the contemporary library, but each layer has some of those functions. In this system, preservation is a process that involves many stakeholders at various levels. The system as a whole provides the balance between access, control, and preservation.

Each of these four layers can be as open or closed as required, based on agreements made between the rights holders and the preserving institutions, so that anything from public-domain to commercially valuable material can be preserved, with terms and conditions of access being applied per item, rather than per archive or across the whole system.

Work can commence on individual pieces of the system without requiring that all the pieces interact with one another from the outset. The Collection layer, for example, will require significant work on principles of digital acquisition, including documenting and storing information about the format and playback software for collected digital materials. Likewise, the Repository layer will require a set of best practices for avoiding or recovering from hardware and media failures. However, these projects can be undertaken separately, by different institutions, and integrated into the system as a whole over time. A key part of NDIIPP's next phase will be to test the architecture's assumptions in practical implementations, as well as to convene a second group to create a more detailed iteration.

NDIIPP Management

Options for the long-term management of a national network of public-private partners must be carefully considered during the execution of the initial NDIIPP investments and actions. The Library proposes to employ the advice and counsel of its National Digital Strategy Advisory Board executive committee to help identify the options to be brought forward for Congressional review. Subsequent investments in building and developing a national strategy for collecting and preserving digital content may necessarily need oversight that includes representatives of the various stakeholder communities. It is the Library's intention to examine and recommend a management approach that optimizes the important investment being made in long-term preservation of our cultural heritage.

Management of the Initial NDIIPP Investment Portfolio

The Library of Congress proposes to invest Congressional and matching funds in a series of activities and programs that begin to realize the vision of securing our digital heritage. Public Law 106-554 appropriated $100 million to the Library of Congress for NDIIPP. (A rescission of $220,000 in another section of the same law reduced the amount appropriated to $99,780,000.) Of this amount, $5 million was made available immediately. The remaining $95 million is available following approval of this plan by the Library's Congressional oversight committees (authorizing and appropriations) and approval before expenditure by the House and Senate Appropriations committees. The law further stipulates that of the $95 million, $75 million may not be expended without matching contributions from nonfederal sources.

Concurrent with the submission and approval of this plan, the Library is seeking Congressional authorization to utilize the $95 million made available, but we are only seeking authority from the Appropriations committees to expend at this time $35 million to fund key practical experiments and work that will help provide a foundation for building the digital preservation infrastructure.

Initial investments will focus on four core capacities—selection and collection development, intellectual property, business models, and standards and best practices—as well as the preservation architecture. They will comprise a balanced portfolio of short-term and long-term projects that include practical applications and models together with applied and basic research. Near-term investments have anticipated time frames of one to two years, mid-term investments will be realized in two to five years, and long-term investments will exceed five years.

We expect the investments in practical applications and models to begin to both capture significant and at-risk content and to establish the core capacities of the preservation network based on actual experience. For example, this work will lead to priority policy decisions in such key areas as the desired digital formats for copyright

Figure 6. Investment in the Digital Preservation Infrastructure—Sample Projects and Anticipated Time Frames

SAMPLE PROJECTS WITHIN INFRASTRUCTURE COMPONENTS	YEAR ONE		YEAR TWO		YEAR THREE		YEAR FOUR		YEAR FIVE+	
BUILDING CORE CAPACITIES	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
1. Selection and Collection Development	▨	▨	▨	▨	▨	▨				
Prototype capture of "at-risk" and important materials by media types	🟦	🟦	🟦							
2. Intellectual Property	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Explore best edition of media types for content submitted through mandatory deposit	🟦	🟦	🟦	🟦						
Test options and authorities to preserve digital content captured on the Internet	🟦	🟦								
3. Business Models	▨	▨	▨	▨	▨	▨	▨	▨		
Test alternative agreements and rules among actors within one repository	🟦	🟦	🟦							
Test alternative agreements and rules among actors across several repositories	▨	▨	▨	▨	▨	▨	▨			
Model and test trusted repositories concepts	▨	▨	▨	▨	▨	▨	▨	▨	🟩	
Explore new collection and preservation service models	🟦	🟦	🟦	🟦						
4. Standards and Best Practices	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Test preservation strategies for different media types	▨	▨	▨	▨	▨	▨	▨	▨		
Institute a cooperative effort to monitor and document developing preservation standards and best practices	🟦	🟦	🟦							
BUILDING PRESERVATION ARCHITECTURE	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Compose current alternative architecture models	🟦	🟦	🟦							
Survey of available technologies, approaches and tools	🟦	🟦								
Define preservation architecture requirements			🟦	🟦						
Design components of the preservation architecture					▨	▨	▨	▨	🟩	
BASIC RESEARCH	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

KEY

🟦 Near Term (one to two years) ▨ Mid Term (two to four years) 🟩 Long Term (more than four years) deposit. Investments in the preservation architecture will yield, in the design phase, implemented protocols and system components in the mid-term. Research investments are envisioned as long-term projects that will support the strategic development of continuing digital preservation.

Initial Investment Portfolio Goals

The goals of the initial investment portfolio are to:

- target high-priority areas for action,
- initiate new projects or leverage existing work in key areas that identify and capture significant and at-risk born digital materials,
- explore organizational and functional models and their relationships to the evolving digital preservation architecture,
- research technical issues associated with long-term preservation of digital content, including a variety of formats and types of content,
- examine relevant copyright issues, and
- begin building a network of economically sustainable partnerships and collaborations for the long term.

Basic Investment Principles

The following principles represent the goals of the NDIIPP and are consistent with its underlying values:

- broad scope of coverage among diverse custodians of digital content, across multiple digital media types, business models, life-cycle management phases, and architectural layers,
- broadly implementable open standards and adoption of protocols rather than adherence to any single proprietary technical solution set,
- incorporation of existing efforts, tools, protocols, technical applications, and so forth, rather than customized development,
- iterative approach that will yield critical data, experience, and knowledge for developing subsequent strategies,
- “early-stage” or “seed capital” type investment in public-private collaborations that have good potential for longer-term economic sustainability beyond the public sector, thus incorporating a provision for in-kind private sector contributions from funding recipients whenever possible,
- periodic benchmarking built into the investment time frame to allow for portfolio assessment and evaluation; provisions made will allow for investment project realignments and consideration of additional fund infusions, and

- adoption of a National Science Foundation peer-review model during the project selection phase where practical and appropriate.

Collectively, these principles will enable investment in a suite of projects to address key questions emerging from the proposed preservation architecture in ways that are focused, practical, and easily incorporated into the further development of the digital information infrastructure.

In the initial period of the NDIIPP program, we anticipate investing approximately $35 million in a portfolio of projects. The following table shows the recommended distribution of funding investments across three broad categories (described below). Distribution of investments is expressed as relative percentages of the portfolio. The planned execution period of the portfolio is three to five years. Recommended investment features, such as in-kind contribution, peer review, and funding mechanism, are shown where applicable in the relevant investment category. The expected number and dollar size of projects are expressed as an anticipated range.

Box 4. Recommended Initial Investment Portfolio

Investment Category	**Approx. Percentage of Portfolio**	**Approx. Project Dollar Range**	**Approx. Number of Projects**	**Funding Mechanism and Features**
Category I Practical Applications and Models	70 percent	$2–5 million	5 to 8	•Cooperative agreements •Peer review •In-kind contributions •Contracts
Category II Digital Preservation Architecture	20 percent	$1–3 million	1 to 3	•Contracts •Cooperative agreements •In-kind contributions
Category III Basic Digital Preservation Research	10 percent	$2–3 million	1 to 2	•Interagency agreements •In-kind contributions

Investment Categories

Three categories of investments are: practical applications and models, digital preservation architecture, and basic digital preservation research. The three broad investment categories are defined as follows:

Practical Applications and Models

The largest portion of the initial investment portfolio focuses on the capture and preservation of “at risk” digital and significant materials while simultaneously testing the technological and organizational framework necessary for building the digital information infrastructure on a nationwide scale. These Category I projects will provide practical experience in acquiring and preserving such new media types as Web sites, e-books, e-journals, digital television, digitally recorded sound, and digital film in order to ensure that the lessons of these investments are broadly applicable to a wide group of archival institutions and communities. This set of projects will focus on organizational, managerial, and usability issues and include: defining and testing trusted repository environments; establishing agreements and rules among rights holders, collections curators, and archival services; and performing tests that stress the layers of the digital preservation architecture, both across the system (say, from repository to repository) and up and down the system (from repository to end-user services). This will also include the development or deployment of new collection and preservation services and functions that the architecture demands for ease of use. All these projects intend to demonstrate how the architecture works—or falters—in implementations carried out by a network of participating collaborators and partners drawn from a diverse group of stakeholder communities. These projects will also help identify policy issues that need priority attention.

This is the largest and most complex set of projects, and the results of these projects have ramifications across a complicated set of interlocking requirements, ranging from the organizational to the technical to the legal to the economic.

Digital Preservation Architecture

These investments focus on a more detailed iteration and further definition and design of the proposed digital preservation architecture. Activities may also include a parallel effort to conduct comparative surveys of the best architectural models that are being defined elsewhere, together with surveys of the available technologies, approaches, and tools that can support architectural components and protocols between architectural layers.

Recognizing that diverse stakeholder communities and industries face common problems of accessioning, annotating, and archiving of digital materials, the Library will identify the best thinking on these issues and use them when advisable. The Library will identify and bring together the relevant technical expertise, tapping resources in both the public and private sector, from government, to industry, to academia. For example, the Library may rely on the technical resources and network of Principal Investigators that have worked under the auspices of the NSF Digital Libraries Initiatives. It may also rely on industry experts in the archival, library, computer science, publishing, recording, and movie industries.

Basic Preservation Research

A third priority, which complements the focused digital preservation architecture, will be basic research intentionally focused on long-term and open-ended exploration of complex issues that may not necessarily have specified outcomes. The Library can leverage the well-vetted digital preservation research agendas of such research support organizations as the National Science Foundation and the San Diego Supercomputer Center to address large-scale research issues such as automatic metadata generation, alternative long-term migration strategies, and the handling of dynamic databases, among others. The NSF's program in digital government offers a recognized forum for arranging interagency collaborations on research problems of common concern.

Funding Mechanisms

The recommended mechanisms for providing financial resources to the project participants will balance the need for broad-based participation, fairness and equity, expediency, shared financial or in-kind match and contribution, assumption of joint responsibilities over the program duration, and reliance on preexisting core competencies. Funding mechanisms that the Library will consider using include:

- cooperative agreements,
- interagency agreements,
- contracts (with individuals, business entities, and public and private institutions), and
- peer-reviewed proposals.

We anticipate that many of the projects in the initial $35 million investment portfolio will include matching funding from the collaborating partners or other nonfederal sources. Therefore, concurrent with the submission and approval of this plan, the Library is seeking authority from the Appropriations committees to expend up to $15 million only upon receipt of matching funds or in-kind contributions. Approval to spend these funds now will encourage private sector participation in the program and the prompt initiation of collaborative projects.

Project Solicitation, Selection, and Evaluation

Solicitation

Given their scale and volume, and to build the broadest collaborative network, we anticipate that most of the Category I projects will be identified through a program solicitation process. We are exploring the possibility of employing, through an interagency agreement, the National Science Foundation to help us administer a peer-review process for selecting the most qualified projects.

Selection Criteria for Initial Investment Portfolio

A set of criteria has been developed to select the initial investment portfolio. Not every project will meet all 10 criteria, but collectively the portfolio of projects should meet these requirements. (For further details about the criteria, see Appendix 10.)

Criterion 1: Does it preserve diverse or at-risk media?

The content that is employed in the portfolio should either test the diversity of potential types, such as dynamic data, geospatial data, or executable files—or examine ways that vulnerable “at risk” materials—orphan collections, ephemera, and so forth—might be identified, captured, collected, and preserved.

Criterion 2: Does it test collaborative network models?

Given the importance of networked collaborations to the overall digital infrastructure building strategy, and given the range of forms in which that collaboration might be expressed, the Library proposes to examine different national and international collaborative network structures, relationships, and mechanisms through the various projects.

Criterion 3: Is there sufficient capacity to achieve satisfactory execution of the project?

The timeliness of the practical application and modeling projects require committed participants who are ready to engage and have identified the requirements, including resources, objectives, and goals of the project.

Criterion 4: Does it address pertinent copyright concerns?

Since rights management is a significant feature of certain collections and processes, proposed projects must recognize relevant rights issues and should begin to craft solutions to aspects such as best edition definition, deposit mechanisms, and acquisition functionality.

Criterion 5: Does it advance the development of standards and best practices?

The portfolio of projects should collectively test a range of approaches to different technical and organizational contexts, examine the conditions under which the different approaches should be employed, and develop sample representations of each, as appropriate.

Criterion 6: Does it help clarify collection selection issues?

Strategies will be required in the digital environment to ensure that the collective scope of the nation’s cultural heritage collections is sufficiently redundant to ensure safety and sufficiently broad and deep to satisfy information and research needs now and long into the future.

Criterion 7: Does it test the digital preservation architecture?

The four-layer digital preservation architecture shows the relationships among the technical layers (Repository, Gateway, Collection, Interface) and suggests the range of organizations that might undertake responsibility for or provide services to different parts of the overall architecture. Because it requires cooperation across organizational boundaries, it is critical to test the digital preservation architecture in a collaborative environment.

Criterion 8: Does it test scalability?

Projects that consider the implications of going from necessarily small-scale prototyping to realistically large and heterogeneous environments are essential.

Criterion 9: Does it test sustainability?

Projects that take into account best practices and other strategies for continuation over time will test the notion of sustainability.

Criterion 10: Does it leverage other efforts?

Existing yet uncoordinated efforts offer resources that may potentially be mobilized into a national system. This represents a prudent and effective investment of federal resources to catalyze a public-private system for the national good.

Evaluation

Before launching the initial investment portfolio, the Library will establish an independent evaluation process to assure that the projects, once selected and funded, are achieving their intended objectives and to confirm the intended results of the projects. The goal is to provide monitoring and feedback that will improve the projects' performance, to leverage learning across the broad spectrum of NDIIPP actions and investments; and to provide the objective validation of the process and outcomes that will assure that the investments are truly strategic and catalytic.

At the start of each project, the Library will ensure that the project objectives are clearly articulated and realistic, that the proposed methodology is appropriate, and that measurable success indicators, such as baseline information or projected benchmark outcomes, are defined.

At the completion of each project we will measure against a set of review criteria for the NDIIPP. This may include validation of the original assumptions and objectives or comparison of project results against previously established benchmarks.

Box 5. Sample Project Description

What might an NDIIPP project look like? Web sites, electronic books, digital periodicals, digital television, digitally recorded sound and digital film are content forms that present opportunities for investigations within the portfolio of practical experiments that may be proposed for the NDIIPP. Any project considered would be measured against the 10 portfolio criteria (see Appendix 10).

An example of a possible project would be the acquisition and preservation of electronic journals across several scholarly disciplines. Electronic journals have become a significant category of born-digital materials. An increasing amount of the nation's intellectual and cultural heritage is embodied in this critical vehicle for scholarly communication. Not surprisingly, many research institutions, including major university libraries, publishers of scholarly material, and the National Library of Medicine (NLM), are investigating the underlying organizational, technical, business, and legal issues involved in archiving electronic journals.

An electronic journals project would advance several of the portfolio criteria, including:

- testing the collaborative model by involving multiple archival institutions, publishers, technology businesses, and scholarly foundations in a network of organizations and services to support and implement the project,
- demonstrating sufficient capacity by proposing a focused action plan and a committed and informed team of participants,
- addressing pertinent copyright concerns by providing a forum and focus for discussions and agreements on access control policies that are key to trusted relationships that support long-term preservation,
- clarifying the collection and management of electronic journals over time, subscribed to by multiple institutions, including issues relating to the breadth, scope, and use of an individual institution's collections,
- testing the preservation architecture by modeling the requirements and functionality for three of the four layers: Repository, Gateway, and Collection. The project would also include defining mechanisms for building a trusted repository environment for e-journals,
- testing sustainability by examining and planning for long-term administrative, organizational, technological, business, and economic models for electronic journals, and
- leveraging other efforts by building upon current e-journal preservation efforts supported by scholarly and federal agencies and by identifying and employing existing suitable technologies.

This is an example of a conceptual project that complies with the portfolio criteria defined for practical experiments to test and model components of the NDIIPP preservation architecture. It does not meet all 10 criteria but could be included in a portfolio of 12 to 15 projects of varying size and scope that all together sufficiently address all the criteria.

Conclusion

The 21st century, barely two years old, is already marked by the expectation of great progress through technological breakthroughs, such as the decoding of the genome, broadened access to information through digital technology networks, and the promises of advances in materials science through nanotechnology. In this century, we have also seen how complex and fragile is the infrastructure that supports these technologies and how easily it can be abused or manipulated. This nation has relied on free and unfettered access to information and innovation, and it has been fostered, to a remarkable degree, by the library and archival network that Americans often take for granted. We are at a critical juncture as we witness the transition from the tested and trustworthy information infrastructure for analog resources to the promising, yet fragile, untested, and potentially insecure digital one. This transition will force us to make a series of important decisions about how to build a system that fosters creativity, protects the rights of individuals, and balances the claims both of creators and of users to access to information and the legacy of innovation.

The vision of the National Digital Information Infrastructure and Preservation Program is to ensure access over time to a rich body of digital content through the establishment of a national network of committed partners, collaborating in a digital preservation architecture with defined roles and responsibilities. It is a vision shared by many. The individuals and institutions consulted in the fact-finding and planning phase of NDIIPP have shown a sophisticated and subtle understanding of the potential of new information technologies to foster creativity and innovation, as well as a keen understanding of what is at stake if we do not secure our digital heritage for future generations. Many expressed great willingness to work toward collaborative solutions for digital preservation. They do not underestimate the complexities of the digital challenge and, in spite of their urgency to begin, they do not expect simple answers or quick solutions. They do, however, believe that little can be accomplished without the steadfast stewardship of this digital preservation infrastructure by a trusted third party such as the Library of Congress.

If we engage the issues, we will begin to transform the national information infrastructure in a way that will strengthen democracy, contribute to a more robust economic climate by providing incentives for innovation and creativity, and secure the unbroken record of the nation's achievements. This we can accomplish through a careful and consultative process of reaching out to an ever larger number of creators, conservators, publishers, and producers who each day make decisions about digital content—decisions that determine whether or not that information will survive into the future.

It will not be easy to engage these issues, nor will it be easy to reach all of the individuals, organizations, and nations that we must involve in forging the new network of digital preservation. The Library of Congress has a historical role to play in convening interested parties and in providing a neutral forum in which all participants can meet as equals. And it is committed to doing so.

But the Library alone cannot achieve any of the aspirations that we heard over the course of the past year. Its power lies in its ability to listen, to learn, and to leverage the efforts and will of others to create an environment in which collaboration is effective and the fruits of innovation continue to accumulate in the extraordinary creative record of our citizens.