The National Digital Information Infrastructure and Preservation Program (NDIIPP) and its Implications for a Research Agenda for Digital Preservation[1]

Laura E. Campbell

The author

Laura E. Campbell is Associate Librarian for Strategic Initiatives at the Library of Congress, where she is responsible for a broad range of digital services and programs. Ms. Campbell is responsible for overall strategic planning for the Library, which includes the National Digital Information Infrastructure and Preservation Program (NDIIPP).

Abstract

Legislation enacted by the U.S. Congress in December 2000 allocates funding to the Library of Congress to lead the National Digital Information Infrastructure and Preservation Program (NDIIPP). The collaborative initiative is focused on materials created primarily in digital form for which there are no analogue representations and which users experience as digital products, sometimes known as "born digital". The investigators have consulted with numerous parties in public, private and not-for profit entities and have defined the "infrastructure" as having two major components: a preservation network of individuals and a technical architecture that provides coherence to localized efforts to archive digital works but is able to accommodate change as technologies advance and organizational needs evolve. This article describes the progress of the initiative and its implications for near and long term research. A striking feature of the research is the integration of technology and organization. The program emphasizes collaboration among a wide range of partners, looking toward solutions that can accommodate multiple and disparate requirements, and communication and outreach to many communities and the public.

Introduction

In December 2000, the U.S. Congress enacted legislation (PL 106-554) directing the Library of Congress (LC) to facilitate development of an infrastructure to support long term preservation of digital content. The legislation allocates $100 million for the National Digital Information Infrastructure and Preservation Program (NDIIPP). The funds are to be released in stages: $5 million was immediately authorized, $20 million are to be made available after Congressional approval of an NDIIPP plan, and up to $75 million to be contingent upon raising $75 million in matching funds. The initiative is focused on materials created primarily in digital form for which there are no analogue representations and which users experience as digital products. Sometimes known as

1 I would like to acknowledge the assistance of Amy Friedlander in conceptualizing and preparing this paper.

"born digital" materials, they are voluminous, heterogeneous, frequently ephemeral and subject to change and updates. Increasingly, our cultural legacy -- from politics to genomics -- are embodied in digital media, and absent a coherent long term preservation strategy, much of our intellectual legacy is at risk.

Unlike preservation of analogue materials, preservation of digital materials must occur at creation, when it is cheaper to capture relevant information and when key information, such as the hardware and software configuration in which works are created, is known. Digital works are not necessarily self-contained. Many are created with implicit and explicit assumptions about the environment in which the work will be displayed; this is a serious problem for resources as diverse as databases of scientific information and dissertations in the performing arts that seek to take advantage of emerging technologies that are least standardized and most subject to change. Thus, digital preservation links the work to its context, and both work and context must be recognized as part of any strategy for long term preservation. Finally, digital materials are more secure when they are redundant. Unlike rare analogue works, which can be protected through restrictive physical controls, digital materials are more vulnerable when they are locked away. Bugs, viruses, damage, and mischief are detected when digital collections are *used*, so we must simultaneously use *and* protect digital materials. But the degree of use must also balance the legitimate rights of rights holders.

In keeping with the goals of the legislation, LC has initiated a range of collaborative activities and partnerships with a host of traditional and non-traditional communities, intended to educate ourselves and others in the library/archival professions on the issues that both impede and encourage concerned groups to engage in long preservation of culturally valuable content. In a nutshell, we have learned that although the technological challenges are great, technology is not the only or even necessarily the most severe obstacle to long term preservation. Technology does inform the solutions, however, and research into relevant technological tools and services must be conceptualised within the appropriate organizational framework. The challenge for the research, then, is to identify long and near-term projects that will enable us to begin to build an infrastructure now that captures vulnerable materials, creates trust among the many stakeholders and accommodates rapid advances as the technologies mature and change.

Lessons from Initial Fact-Finding{tc \l 1 "Lessons from Initial Fact-Finding"}

Our initial fact-finding activities took place between summer 2001 and summer 2002; they spanned the waterfront. We hosted meetings and scenario planning workshops organized by a well-known consulting firm, the Global Business Network (GBN), and conducted structured interviews. Through these strategies, LC eventually reached over 200 representatives from professional associations; entertainment, film, music, radio, commercial and non-commercial broadcasting; higher education; libraries, museums, nonprofits organizations, foundations and cultural institutions; newspaper, magazine, book, scholarly journals, and textbook publishing; and software, Web design and

development. Many of these people already meet in different, more specialized venues. For example, chief technical officers of major studios, entertainment companies, and commercial publishers participate in working groups hosted by the U.S. National Institute of Standards (NIST) to discuss technical standards and issues. Rarely do they talk to librarians and archivists. We have been heartened by the core belief in the cultural value of preservation and in the level of trust potentially vested in the Library even while we recognize that there are many challenges to forging cooperation.

In parallel with these consultations and workshops, we commissioned a series of studies. Six environmental scans on formats of particular relevance to LC's collections policies were prepared by leading experts; these covered archiving the Web (Lyman 2002), electronic journals (Flecker 2002), electronic books (Romano 2002), digitally recorded sound (Brylawski 2002), digital television (Ide et al. 2002) and digital moving images (Wactlar and Christel 2002). Both the Digital Library Federation (DLF), a consortium of about 29 research libraries that are on the forefront in the adoption of information technologies to extend their collections and services, and the Association of Research Libraries (ARL), which consists of more than 120 member institutions who represent the major research libraries in North America, surveyed their members concerning their activities in long term preservation of digital content. A report on international preservation activities in the United Kingdom, France, the Netherlands, and Australia was prepared, and the Council on Library and Information Resources (CLIR), which provided coordination for many aspects of the consultation and research process, completed a report that summarized the landscape of digital preservation. Participants in some of the meetings had expressed some confusion about the current copyright regime, and in response, we commissioned an introduction to copyright law and its implications for building a digital archive in the U.S., recognizing that the international implications of digital archiving represent a future topic for research. All of these studies have been summarized and included in the NDIIPP Master Plan, scheduled to be released this fall.

This fact-finding and survey effort taught us a number of lessons: There are a number of initiatives in the U.S. and worldwide in the public, not-for-profit, and commercial sectors that are addressing the challenges of preserving digital content but we are all learning together; the "problem," so to speak, has not been solved. Preservation poses technical, organizational, and legal issues and requires cooperation as well as research on several fronts. There is a reservoir of good will and a general belief that the threat to the collective digital memory is imminent and that major libraries, including the Library of Congress and other national libraries and memory institutions in the U.S. and worldwide, can exercise leadership that will balance the legitimate legal rights and economic interests of multiple stakeholders with the collective public interest in preserving our heritage as it is embodied in digital form. We intend to continue the listening and consultation process as NDIIPP moves from its planning phase toward research and testing of pilot projects over the next two-to-five years.

Despite tremendous advances in processor speed and storage capacity, preservation is simply more than putting 0s and 1s onto stable storage media, hard as that is. Rather, our

goal is to ensure persistent, rights-protected access to this material that will enable future generations to satisfy their information needs through an infrastructure of partnerships, agreed-upon practices, and standards. Much remains unknown and/or untested at a scale that is realistic for the needs of a modern major library or requirements of a network of cooperating institutions. Therefore, with the National Science Foundation (NSF)'s program in Digital Government, LC has facilitated definition of a long term research agenda that involves input from other federal agencies as well as from the universities, commercial laboratories and the non-profit foundations. The library and archives communities have evolved such an infrastructure through decades of agreements, participation in national and international standards bodies and professional organizations, and informal arrangements among cooperating institutions. One of our challenges is to collapse that time frame into a few years and to begin collecting and preserving ephemeral digital resources before they vanish. Coming to terms with the requirements of a new infrastructure, which may well function in parallel with the existing one, necessitates both long and short-term research.

The Two Faces of Infrastructure {tc \l 1 "The Two Faces of Infrastructure"}

At the same time that we undertook the background fact-finding and baseline studies described in the previous section, LC also commissioned an analysis of the roles and functions required in a national infrastructure for long term preservation. The results of this investigation described a context that identifies the scope of activities that transpire in this landscape; identifies many of the potential actors and institutions, and defines future (existing and new) coordinating bodies, research and development activities, and enabling agreements that will ensure preservation of and access to digital content over time. Together, these actors, coordinating bodies, enabling agreements about roles and responsibilities, research, policies, and common practices constitute the largely *intangible* digital preservation network of people and institutions.

The second face of the infrastructure concerns the technical architecture. In February 2001, we convened a small group of technical experts who sketched a four-part architecture that will support the values of transparency, collaboration, incremental development, stability, flexibility, heterogeneity, and innovation. We are aware of concurrent efforts into the architecture of digital preservation underway at the San Diego Super Computer Center (SDSC), Massachusetts Institute of Technology (MIT) and elsewhere. One of the advantages of the proposed four-part approach is that it allows multiple parties to work on different pieces while ensuring overall coherence. We see work that LC undertakes or initiates as contributing to other efforts, just as these projects will inform the work that we do.

The architecture itself consists of four layers: repository; gateway; collections; and interface. No single layer represents a digital version of a contemporary bricks-and-mortar library; taken as a whole, the architecture embodies many of the functions of a library or archive and supports, we hope, nearly all of the activities that future librarians

and archivists will undertake.

The bits -- 0s and 1s -- are stored at the repository level. They are associated with an identifier; we do not specify a syntax for that identifier -- only that an identifier exist. The "gateway" exercises control over which requests may access data stored in the repository. The third level, "collections", provides many of the functions and decisions associated with professional librarians and archivists, including the acquisition and appraisal, the technical information that provides context to the data from the repository, metadata, software needed to render the properly, and so on. Finally, an interface layer exists where patrons can access information. Protocols defining how each layer works and how information flows from layer to layer have yet to be defined. The intent is to specify a minimum requirement to enable different implementations to provide tools and services that are appropriate to their users and to their collections while permitting the systems to interoperate.

Different tools and services might be available at the different levels depending on the nature of the collections and the mission of the institution. For example, at the collection level, individual objects can be aggregated into different logical collections depending on local decisions and the terms of use associated with given objects. A collection that supports computer science research might consist of all items in a repository that are in the jpg format; this collection might be a temporary one to support a research project, or it might be permanent. Some of the same items be also be incorporated into a collection defined by content, say, World War II, where photojournalism from the battlefield was both an innovation in news gathering in its day as well as an important body of historical material. Also associated with this collection might be official reports, diaries, memoirs, published histories, and so on. Whereas the jpg collection could be created by running a program ("find all files or objects with the suffix 'jpg'"), the World War II collection would presumably be created by a scholar who might employ information retrieval tools to assemble it. In both cases, the "collection" has been logically defined; users can see and work with the "collection" but no bits have been physically moved from one repository to another.

Within the frameworks offered by this organizational network and technical infrastructure, LC has envisaged a three-tier research program: (1) a set of activities that is focused very directly on the technical infrastructure and building the preservation architecture; (2) a set of pilot projects and experiments that test that infrastructure and related organizational issues and are associated with building core capacity; and (3) a long term basic research program that will be implemented through NSF's program in digital government. The part of the research program focused on the technical infrastructure will consist of four major activities: (1) compare and evaluate alternative models of the technical architecture; (2) survey available technologies, approaches and tools; (3) define preservation architecture requirements; and (4) design components of the preservation architecture. All of these activities will require coordination with interested parties who are already engaged in building such systems or who are likely to need such systems. The results of our investigations will be published and posted to the NDIIPP

Web site (http://www.digitalpreservation.gov) for use by interested communities.

Building Core Capacities

Core capacities refer to the shared knowledge, expertise, skills and consensus regarding areas of concern essential to supporting the collaboration among organizations that comprise the preservation network. Sample projects that illustrate pilot projects that go toward building core capacities are shown in Table 1.

1. Selection and Collection Development
    a. Prototype capture of vulnerable and important materials by media types

2. Intellectual Property
    a. Explore best edition media types for content submitted through mandatory deposit
    b. Test options and authorities to preserve digital content captured on the Internet

3. Business Models
    a. Test alternative agreements and rules among separate actors (or institutions) cooperating within a single repository
    b. Test alternative agreements and rules among actors (or institutions) across several repositories
    c. Model and test concepts of "trusted" repositories
    d. Explore new collection and preservation service models

4. Standards and Best Practices
    a. Test preservation strategies for different media types
    b. Institute a cooperate effort to monitor and document developing preservation standards and best practices

Table 1. Building Core Capacities

The striking feature of these example projects is the integration of technology and organization. For example, prototyping capture of vulnerable and important materials by media requires a systematic evaluation of collection development policies by individual institutions to determine what is and is not relevant to their mission. Moreover, given the long term expense of preservation, the need for redundancy in digital collections as a means of enhancing security, and the historically broad missions of the major institutions, coordination in collecting policies becomes an important feature of capture requiring the major libraries to develop processes and agreements to support coordination. Possible tools include shared registries, model agreements, procedures for loans, and, perhaps, standing working groups through IFLA and other international organizations.

Other projects consider the implications of the technology but do not require technological research per se. Those associated with aspects of intellectual property, definition of "best editions", and developing model contracts and agreements among cooperating institutions fall into this category. Still, some of these efforts may result in demand for new and different tools or for further research into related topics.

One of the best examples of how the cycle of research may be iterated is the problem of selecting among preservation strategies. For many years, preservationists have debated the merits of different preservation strategies: emulation, encapsulation, and migration. The preservation strategy "emulation" preserves the original application program which enabled or supported a digital work; the goal is to preserve the look and feel of the work as well as its functionality (Lee et al. 2002, pp. 95-98). Obvious examples of the kinds of works that might rely on emulation are games and simulations. "Migration" refers to transfer of digital materials from one configuration to another to preserve use of the material despite evolution of the underlying systems. Migration can mean merely copying the material ("refreshing" it) from one storage medium to another, or it can mean transforming it so that the content is stable but the underlying bits may be changed. Databases of scientific information, for example, might be good candidates for migration; indeed, since the 1960s, the Library of Congress has migrated over 12 million bibliographic records and 4 million authority through three internal software formats, three types of tape formats and six types of hard disk drives. Finally, encapsulation bundles B or "wraps" -- information about the interpretation and display of the work into the work itself so that the work becomes self describing to the system on which it is to be displayed. As Lee and his colleagues at NIST argue in their review essay (Lee et al. 2002, p. 02), all three of these major techniques actually embrace a range of strategies and although different strategies appear to be appropriate for different types of works, both the strategies and the settings under which they can be deployed remain relatively unexplored.

A key component in deciding among preservation strategies is use: how frequently is material accessed? Does "look and feel" matter? We know from many historical examples that the value placed on works and their uses change. For example, US census data, which have long been the preserve of genealogists and local historians, took on new life as social scientists used computational and statistical techniques in the 1970s and 1980s. Continued employment of new technologies enabled a new kind of historical research, culminating in such wonderful modes of expression as the Web-based "Shadow of the Valley" project that mines multiple sources to present a unique interpretation of the American Civil War (1861-1864) (http://www.iath.virginia.edu/vshadow2/). Although we must begin the process of preservation at the creation of the digital work, we are also learning that strategies for preservation must take future as well as current use into account.

As a result, ongoing user studies are essential to preservation. As expectations and formats evolve, we may look to a day in which initial preservation strategies may be

modified. Perhaps a given instantiation of a dynamic database will become historically important, and thus, that "slice" of the database will be preserved as an independent object even as new information is continually added to the original. One tool that may enable this kind of evolution of preservation strategy is the computer science notion of a "wrapper", code that surrounds or wraps a digital object and contains information about that object. Thus, a given digital work can be "wrapped" multiple times on its own or as part of a larger object or collection. Metadata is another tool for characterizing objects and their relationships to other objects. Both are elements in a basic research agenda, which is discussed in the next section.

Basic Research

Unlike the kinds of projects described in Table 1, basic research is open-ended; its goal is creation of new knowledge rather than solution of an immediate problem. To coordinate this dimension, LC is working with Professor Margaret Hedstrom of the School of Information at the University of Michigan, who has been working on problems of digital preservation and archiving for many years. She and a team of experts drawn from LC, other federal agencies and the private sector organized a workshop sponsored by the National Science Foundation's program in digital government with assistance from the interagency Digital Libraries Initiative and with the cooperation of the Library of Congress in April 2002. About 50 people were invited and by the end of the two-day discussions, the organizers were able to articulate the framework for a research program in the long-term preservation of digital content.

According to the NSF program manager, Lawrence Brandt, the digital government program, in general, seeks to partner with other federal agencies to develop research in a domain of joint interest. The key, from his point of view, is to deepen relationships among agencies that might not otherwise recognize a common interest in a body of research. But, he cautions, it is important to hold expectations in check. Systems that can result from research are not likely to be the robust systems an agency may eventually require (As reported in Friedlander 2002). Consequently, there will be a need beyond the initial research phase for work that converts a promising experiment into a reliable, operating system. LC's strategy of supporting targeted pilot projects as well as collaborative basic research offers an opportunity for cross-fertilization of ideas from near- and long-term research projects.

Core ideas and research topics discussed at the April workshop included: (1) definition and attributes of digital collections; (2) tools and technology; and (3) policy and economic models. A full report on the workshop is in preparation and will be posted to the Web (http://www.si.umich/digarch/). The following paragraphs elaborate on some of the issues subsumed within these broad topics.

Definition and attributes of digital collections:

Core issues arise from the difficulties of establishing the boundaries of a "digital object", agreeing upon identifiers, and naming conventions, and developing hierarchies of significance that will enable may agencies and organizations in the commercial, not-for-profit and public sectors to participate in the complementary collection policies. Specifically:

- Selection and preservation of complex digital objects: Although methods exist to preserve relatively simple, static digital objects, it is increasingly difficult to establish the boundaries and necessary preservation characteristics of hyperlinked, nested, compound digital objects that represent works with relationships to other resources. For example, who preserves a portal site, which may be updated daily and is rich with links to many independent sites? If it changes all day long, like a newspaper site, which instantiation is the "version of record"?

- Aggregation of items and objects into collections: Collections represent aggregations of individual works; they have historically been arbitrarily defined B by tradition, circumstances, processes of collecting, mission of the agency, company or cultural institution, and so on. Preliminary work undertaken by the San Diego Super Computer Center as well as the experience of the Library of Congress (LC), the National Archives and Records Administration (NARA), the National Library of Medicine (NLM) and the National Agricultural Library (NAL) suggests that a more active approach to collection definition, including metadata and multiple definitions (e.g., by format, content, terms and conditions, etc.), may support effective management of the collections as well as efficient search and retrieval.

As the example of historic photography previously given in the text suggests, some collections can be thought of as "permanent" while others may be intentionally transitory. If a collection is deemed transitory, what happens to the collection level metadata that was used to manage the temporary collection? Is it maintained as part of the historical record? Or is there a policy of weeding such as libraries now practice and should some aspects of weeding be automated, based, perhaps, on use depending on the provenance of the items. Thus metadata created for purposes of management might be subject to weeding at the collections level but material deposited in the repository might be subject to more stringent controls and policies.

- Decision mode for selection: Long term preservation of digital content will continue to require decisions about what to preserve, for whom and for how long -- increases in storage capacity notwithstanding. Much research is subsumed into this topic -- from devising decision making matrices to selection of the appropriate preservation strategy to tools to support methods of coordinating collection development among cooperating entities.

- Resolution of naming hierarchies: Preservation requires consistent identification of individual works, collections and meta-collections. The structure, management and

long term persistence of identifier systems has been a long-standing computer science research problem. Several approaches exist; work in digital preservation is positioned to leverage this research within the systems of acquisition, ingest, management, and storage.

- Metadata: Defining metadata has challenged researchers for better than a decade. It is critical to search and retrieval, efficient management of collections and repositories and interoperability among collections as well as to the management of rights and terms and conditions of use. Various schemas for partitioning metadata by function have been proposed and metadata specific to preservation is also a lively topic. This is an area in which LC expects to contribute expertise as well as benefit from the ongoing research of others.

Tools and technologies

Long-term preservation of digital content will require tools and technologies to support work flow and digital collections management. For example,

- Acquisition, ingest, and collection management: Digital collections are vast and highly heterogeneous in format (e.g., images, text, motion pictures, data bases) and content (e.g., scientific measurements, geographic information, dissertations in the performing arts that are presented digitally via the Web, etc.). Appropriate tools and technologies are required to enable librarians and archivists to identify relevant information, catalogue it, store it for future use, and communicate with other institutions as necessary to support cooperative relationships and redundancy.

Although it may not be difficult to imagine this flow of information among five or six major libraries, the volume of digital information means that these cooperative relationships will have to occur over a hitherto unprecedented scale, perhaps requiring automation of many tasks. As the example of weeding suggests, the versatility of logically created collections offers a set of management challenges. The scale of the collections means that individuals alone cannot manage obsolete data that might potentially interfere with system performance. Tools to identify and manage use of materials will be needed.

- Data management and storage: Storage media degrade over time, requiring information managers to monitor and periodically copy the stored data from one medium to another; this is known as "refreshing the data". At a minimum, given the sizes of digital collections, automated systems for monitoring will be required as well as tools (e.g., algorithms) that can traverse the collections to identify vulnerable materials.

- Standards and interoperability: Disparate, stove piped collections of heterogeneous information will be of little value for future users and will be extremely expensive to

maintain. Research into standards and interoperability will support consistent management and use of data in multiple formats and created under different requirements.

- Metrics: Consistent and agreed-upon metrics are essential to consistent and cost-effective management of information. Although the need for metrics has been recognized, which metrics, what attributes they measure, how they are constructed, and measures of their effectiveness are all topics for future research.

Policy and economic models

Digital systems, technologies and tools exist in a framework of policies and procedures. Devising incentives that encourage cooperation across traditional competitors remains a challenging area of behavioural, economic, and social research. Possible topics include:

- Appraisal of digital information: Currently, there exists no agreed-upon model for appraising digital works and collections. This inhibits formulating reliable business models, proposing viable tax incentives, and developing cost-effective long term plans.

- Public support: Citizens rely on public access to information for a variety of business and personal reasons, from marketing plans to family genealogies, but are relatively unaware of the implied costs of creating, managing and preserving this information. An educational and outreach program, built into a popular Web site such as the U.S. National Park Service's Ellis Island Web site (http://www.ellisisland.org/), suggests that the public face of digital preservation can be used to deepen awareness of the value of these resources and the requirements they pose for the present and the future.

Conclusion{tc \l 1 "Conclusion"}

NDIIPP is embarking on a journey in which the landscape is changing rapidly. Experience teaches us that some of those changes will prove fleeting and some of the contemporary tools will become classics, almost without our knowing it. As stewards of the collective memory, we have set ourselves the task of coping with change even though we have long thought of our mission as inherently conservative. The stakes are great; the rewards are even greater. In our turbulent times, we choose to support the acquisition of new knowledge through a program of near- and long-term research in the cause of preservation for sake of the future.

References

Flecker, Dale (2002) Preserving digital periodicals, In: *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving,* Washington D.C., Council on

Library and Information Resources, pp. 10-23.

Friedlander, Amy (2002) The National Digital Information Infrastructure Preservation Program: expectations, realities, choices and progress to date, *D-Lib Magazine* 8(4) (April). URL: http://www.dlib.org/dlib/april02/friedlander/04friedlander.html.

Ide, Mary, MacCarn, Dave, Shepard, Thom and Weisse, Leah (2002) Understanding the preservation challenge of digital television, In: *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving,* Washington D.C., Council on Library and Information Resources, pp. 67-79.

Lee, Kyong-Ho, Slattery, Oliver, Lu , Richard, Tang , Xiao, and McCrary, Victor (2002) The State of the art and practice in digital preservation, *Journal of Research of the National Institute of Standards and Technology* 107(1) (January-February), 93-106.

Lyman, Peter (2002) Archiving the World Wide Web, In: *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving,* Washington D.C., Council on Library and Information Resources, pp. 38-51.

Romano, Frank (2002) E-books and the challenge of preservation, In: *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving,* Washington D.C., Council on Library and Information Resources, pp. 23-37.

Wactlar, Howard D. and Christel, Michael G. (2002) Digital video archives: managing through metadata, In: *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving,* Washington D.C., Council on Library and Information Resources, pp. 80-95.