

How Digital Technologies Have Changed the Library of Congress: Inside and Outside

By Laura E. Campbell, Associate Librarian for Strategic Initiatives and Chief Information Officer, Library of Congress

While the benefits of digital technology are obvious to the millions of constituents the Library of Congress has attracted with its Web site and other electronic services, the impact of technology on the organization itself has been no less significant.

The digital age has enabled the Library to exploit the benefits of technology for the benefit of our constituents. We now reach many more millions of users online than ever enter our doors on Capitol Hill. Approximately 2 million visitors are greeted by the Library each year; more than 50 times that number visit our Web site during the same period. We have established new constituencies, such as teachers and their students, and we are making an important difference in the education of this nation's future leaders through the more than 11 million primary source materials we offer online.

These are the changes that are obvious to those outside the institution. But there are other, I would argue, no less important changes that digital technologies have brought to the internal operations of this 207-year-old institution.

The Inside

The Library of Congress comprises six service units – Office of the Librarian of Congress, Congressional Research Service, Library Services, Law Library, U.S. Copyright Office and Office of Strategic Initiatives, which, among other programs, leads the National Digital Library Program and the National Digital Information Infrastructure and Preservation Program (NDIIPP). These service units, which have pursued a “silo” approach to achieving goals for most of their existence, are now having to work more closely together in new networks being forced upon the institution by the advent of the digital age. Various types of “social” investments are required in order to form these partnerships.

Practices that were developed primarily for published materials must be changed, and the changes that need to be implemented are requiring archival institutions to transform the way they have been doing business – in some cases for centuries, as in the case of the Library of Congress.

For example, the collections of the Library of Congress are built primarily through copyright deposit, which requires in most cases that two copies of every book be deposited with the Library as part of copyright registration. The typical journey of a book is that it is received by the U.S. Copyright Office, a curator selects it for inclusion in our collections, the book is processed and cataloged, placed in its proper place on a shelf and eventually served to a reader who requests it. Once it leaves the Copyright Office, it goes to our Library Services unit for the remainder of its lifecycle. No other area of the Library has a hand in the process.

Our Copyright Office now must work with all areas of the Library as it retools itself to handle digital media. For example, Copyright staff must work with our curators to determine what is considered to be a “best edition” of an electronic book for the purposes of deposit. The Office of Strategic Initiatives has a role to play in determining the formats required for digital deposits to ensure their preservation over the long term. The Congressional Research Service, which works solely for the U.S. legislature, must be able to serve these materials to members of Congress while still adhering to copyright law.

One auxiliary benefit of the collection and preservation of digital materials for the Library is that new relationships are forming among our staff and their respective service units. We are all learning to think about the lifecycle of materials from the moment of acquisition, and everyone involved needs at least a basic understanding of the entire lifecycle – the “anti-silo” approach. Digital content needs curation by

people working in teams who combine domain understanding (nature of stuff, expectations of specialist users) with understanding of the technology (formats, data structures, system design).

The Outside

For most of its history, the Library of Congress, like most centuries-old organizations, has been slow to change. Until about 20 years ago, in the so-called analog age, the Library was seen as a somewhat insular government agency with few ties to the broader community of content creator and collecting institutions. We have always served the U.S. Congress in its lawmaking duties through our Congressional Research Service, and researchers traveled (and still do) from across the nation and around the world to use our unparalleled collections. But we did not serve readers under age 18 nor did we have special programs for educators and, of course, you had to get here to use our materials.

That has all changed and so has our approach to working with outside organizations. When, in 1994, we started our flagship Web site, called American Memory, we said from the beginning that we would build this digital resource for the nation with other repositories nationwide. Although the majority of the site's content comprises digitized versions of unique materials from the Library of Congress, a substantial portion of the site, 23 of the 137 thematic presentations, are the result of collaborations with other institutions. This is significant because not only were we working with nearly two dozen institutions in a single program but also, for the first time in our history, our "collections" now included materials that were not housed at the Library.

The digitization of the Library's collections also had varying effects on our staff. Some curators and reference specialists resisted the idea of placing "copies" of original works online for scholarship purposes. Others sniffed at the idea of serving those who were not "serious researchers." But the head of our agency, Librarian James H. Billington, by force of his will and his political instincts, took the Library headlong into the digital age. He realized that if the Library was to remain relevant in the latter part of the 20th century and beyond, we had to make ourselves useful to the broader American public – Congress's constituents. That decision is responsible for the enormous success of our National Digital Library (NDL) Program and its auxiliary Educational Outreach Program, and has resulted in the Library's leading role in the dissemination of electronic information.

With the encouragement of Congress, we are now rolling out our educational program to all 50 states, the District and territories. Our leadership among federal agencies in making important materials accessible on the Web, through the NDL Program, directly influenced Congress's decision in 2000 to ask the Library to lead a digital preservation program for the nation. The expertise, trust and goodwill we gained from the NDL in the collaborative arena taught us many lessons about working with other major repositories as well as about a variety of technical issues.

The transition has not been easy, but digital media have forced us to look strategically outside our walls, and that strategic point of view is what guides our National Digital Information Infrastructure and Preservation Program (NDIIPP).

The goals of NDIIPP are to:

- X encourage shared responsibility nationwide among many institutions and organizations for the collection, storage and preservation of digital content
- X seek national solutions for the continuing collection, selection and organization of historically significant cultural materials regardless of evolving formats
- X ensure the long-term storage, preservation and authenticity of those collections and
- X work toward persistent, rights-protected access for the public to the digital heritage of the American people.

Project Highlights

In September 2004, NDIIPP made its first set of “investments” in building its so-called “digital preservation network.” In this case, “network” refers to the building of institutional partnerships with other organizations that would agree to collect specific types of born-digital content deemed to be essential to America’s intellectual heritage and at risk of loss if not preserved now.

At that time, NDIIPP made awards totaling nearly \$14 million to eight consortia comprising 36 institutions. Approximately 70 percent of NDIIPP funding has gone toward content-based projects to model and test sustainable approaches for a decentralized and distributed network of partners. Approximately 20 percent of NDIIPP resources are going toward exploring the technical architecture necessary to support these partners. Basic digital preservation research comprises the remaining 10 percent of NDIIPP funding.

In addition to these three investment areas (details below), the Library of Congress has formed an independent working group designed to examine an important portion of the U.S. copyright law that deals with libraries’ use of archival materials. We have learned that we would not be able to move forward with the digital preservation program until we resolved some of the intellectual property issues that hindered our work.

The Library is in a unique position because the U.S. Copyright Office is part of the institution. The newly formed working group, known as the Section 108 Study Group, convened in April 2005 under the sponsorship of the Library of Congress and the U.S. Copyright Office to re-examine the exceptions and limitations applicable to libraries and archives under the Copyright Act, specifically in light of the changes produced by the widespread use of digital technologies since the last significant study in 1988. The group will make recommendations this year for changes that result in draft legislation for Congress, addressing exceptions for libraries and archives to collect, preserve and serve digital materials.

Today we have 67 partners and hope to soon add another 60 partners to our expanding digital preservation network of committed institutions. These partners are devoted to building a distributed network of accessible content through the selection, collection and preservation of important at-risk materials.

Outcomes So Far: By Types of Partners

1. Collecting and Preserving Content Partners

Collaborative Collection Development

The 36 institutions working as collaborative partners within NDIIPP have formed both formal and informal networks since receiving their awards in September 2004. Their activities, however, have come to focus on four cross-cutting areas:

- Selection and collection of digital content
- Intellectual property issues
- Development of a secure technical architecture and
- Economic sustainability of the digital preservation work that they are now engaged in.

The networks built around these four areas have become so well-established that when the partners meet as a whole group twice yearly, they also break into these four so-called “affinity groups” to focus on these areas.

In general, we have learned so far through these eight project partnerships:

- It is better to separate preservation and access, at least conceptually. Although they may often go together, collaboration for access can be independent of collaboration for preservation.
- Construct your digital preservation system modularly.
- Assemble the system over time, not all at once.

- You need to be able to upgrade parts without disruption of the whole.
- The more broadly adoptable the standards and protocols used, the greater chance for success and sustainability of the preservation approach.

Archive Ingest and Handling Test

The Archive Ingest and Handling Test, which was completed in June 2005, serves as an example of how NDIIPP is catalyzing joint problem-solving to achieve programmatic goals. AIHT tested the ingest of a large archive into diverse systems. The digital archive was donated by George Mason University, and the Library conducted the test with Johns Hopkins, Harvard, Stanford and Old Dominion universities. The archive contained approximately 57,000 files totaling about 12 gigabytes. Although relatively small, it was complex in its mix of formats and metadata.

The archive test proved that different approaches to the same problem can coexist and work successfully and coincidentally. We learned which aspects of digital preservation are institution-specific and which aspects are more general. In fact, the Library believes that taking several approaches to the same problem is preferable to homogeneity, which risks data corruption or irretrievable loss should the single-system solution fail.

The test also taught us that a data-centric approach to the transfer of content is preferable to a tool-based strategy. Thus, this approach assumes that data will pass among institutions in its original context, to be interpreted by the ingest system of the receiving/preserving institution. Of course, heterogeneous approaches to the same problem can only be successfully guaranteed when networking among various institutions exists to the degree necessary to ensure interoperability.

Portico, LOCKSS, SCOLA

Other partners collecting important content are Portico, which is developing an archiving service for electronic journals; LOCKSS (Lots of Copies Keep Stuff Safe), which is a multi-site distributed archive of content; and SCOLA (Satellite Communications for Learning), which is saving high-interest foreign news broadcasts such as those from Al-Jazeera and from Pakistan, Russia and the Philippines so that they are available for future research.

2. Digital Preservation Research Partners

In May 2005, the Library and the National Science Foundation awarded 10 university teams a total of \$3 million to undertake pioneering research to support the long-term management of digital information. These awards are the outcome of a partnership between the two agencies to develop the first digital-preservation research grants program.

3. Technical Tools and Services Partners

The network of collecting and preserving partners has identified tools and services that are needed to preserve digital content. One of the most important services is storage for large volumes of files. Others include tools to work with metadata, tools to examine and verify file formats. Several partners within the network are testing and demonstrating these tools and services.

In May 2005, the Library of Congress began a one-year pilot project with the San Diego Supercomputer Center (SDSC) to assess the ability of a trusted partner to store digital data from the Library. The two main objectives of this project were for SDSC to:

- Reliably host Library of Congress digital content and guarantee data integrity and access.
- Enable the Library to remotely access, manage, process and analyze that content.

The Los Alamos National Laboratory Research Library is building mechanisms that will help address challenges related to collecting, storing and accessing digital materials.

Of particular interest are “complex objects,” such as a document with multiple separate pieces. Tools are under development for assigning metadata, transferring content between repositories and storing content within repositories. Los Alamos is using MPEG-21 as the underpinning of this work.

The Library has come to conclude that the relationships among stakeholders and their commitments to collaborate will vary from one type or body of content to another. For example:

- Are the future users from the same community as the content creators?
- Must an archive be relatively dark because of rights issues?
- Is a consortium a group of peers with mutual commitments or is it the building of a centralized service or set of services?
- Is the archive also the primary resource for continuing access and use (not for published literature, but likely for scientific data)?

The preceding conclusions prove that there will not be just one technical architecture for digital preservation. Each instance of technical architecture must support the nature and relationships of that organization and its stakeholders.

There is a need for a shared effort to develop certain centralized support elements, elements that will be used by a number of preservation organizations. One example pertains to digital formats, whereby collective efforts could construct registries for format information and develop tools that support automated identification, validation and characterization of digital files. This idea has received strong endorsement from our NDIIPP partners.

Systems will be built with a layered architecture.

- There is a growing consensus that we understand what is needed for bit preservation (the “bag and tag” idea). Bit preservation applies to any sort of digital content and can be a commodity service.
- Accepting a common general approach to bit preservation means that digital curators/custodians must take on responsibility for understanding the content and making decisions about dealing with technological obsolescence.
- The “higher” content management layers that provide the capabilities to normalize or migrate content, or to provide emulations of obsolete systems, are not as well understood as bit preservation. Since the application of migration and even emulation is likely to lead to changes in look, feel or behavior, curators must be able to identify and articulate what is essential to the meaning of various classes of content.

Various types of “social” investments are required in order to form partnerships across institutions, and trust is at the heart of these relationships. Archival institutions (as well as society) need to have trust in different aspects of the digital preservation landscape:

- Trust in selecting what to preserve. Will the necessary networks be formed so that, in the aggregate, the various collecting and preserving institutions are saving a “universal” collection?
- Trust in preserving the bits against threats: hardware, software, media failure; communication system errors; network failures; media and hardware obsolescence; operator error; natural disaster; internal attacks; external attacks; economic failure; or organizational failure. Will institutions work together sufficiently to assure that technical failure at one institution does not mean permanent loss of information?
- Trust in making wise choices and robust implementations for future format migration or emulation to deal with obsolescence of software to render the content even if the bits are safe. Will institutions’ need to collect and preserve for their individual needs and technical architectures override the overall needs of the entire network of preservation partners?

How does trust develop?:

- Trust in people or organizations comes through direct experience of working with them and through reputation or accreditation.
- Trust in organizations develops through understanding of governance and financial stability.
- Trust in systems can be ensured through auditability and an understanding of how they work.

Conclusions

Because the problems of digital preservation cannot be solved by a “silver bullet” or by a single institution, the success of the Library’s catalytic approach to its leadership of NDIIPP will be the key to the success of the entire digital preservation program. What is most important is the quality of the social networking that results from NDIIPP, which is formally set to conclude in 2010 (though digital preservation work will continue) with a report to the U.S. Congress on our achievements and suggestions for going forward.

Although collaboration through these social networks is the key assumption behind a successful digital-preservation program, the current state of digital preservation is such that there are many institutions developing systems to satisfy their own particular needs, and it is unrealistic to think this situation will change. In fact, the trend is for digital preservation efforts to continue using systems keyed to institutional goals. One size will not fit all.

There is now a deep concern among public institutions about the proper care of digital information. Collecting institutions are also interested in the development of practical models for digital preservation, given the fact that their resources are limited. We also have learned that there are expectations for the Library to be a catalyst and coordinator for various digital preservation efforts, without asserting “top-down” ownership.

Who will pay for digital preservation is not clear, and it is of paramount concern to all our partners. There is much educational work to be done. The case for digital preservation’s importance must first be made with the general public – the constituents of those governmental officials who control the purse strings -- before public officials will make this work a governmental priority.

We have much to learn about how to share this responsibility, and it will require a transformation about how we do our work – not just within our institutions but among them as well.

We are preparing to make a new set of investments in the evolving national stewardship network in areas we have not explored as much as we would like. For example, we are planning on working directly with content creators in the private sector. We need to capture and preserve more of this material, particularly the ephemeral content that creators and distributors may not be interested in retaining themselves. We need to understand the synergies between activities in the private sector that are intended to sustain digital assets for future commercial repurposing and activities that support longer-term preservation and access for society. We also need to instill among private sector creators and distributors an awareness and appreciation for digital preservation. We will work with producers of music, film and digital images, among others, to agree on standards and metadata approaches that can be incorporated into their work to facilitate their preservation. This will be especially critical as electronic deposit of materials with the U.S. Copyright Office becomes a standard way of doing business.

The NDIIPP initiatives funded thus far have provided us with the necessary information to prepare for our next investments. We are working to take what we now know and develop networks among all the partners from our three areas of investment – content collecting and preserving, digital preservation research, and technical tools and services – to leverage this expertise. As we learn by doing we also catalyze new ideas and solutions.

If NDIIPP and other digital preservation programs worldwide can stimulate formation of the types of social networks necessary to sustain digital media across their lifecycle, then we will have achieved a framework to ensure that the intellectual heritage of the world remains preserved for generations to come.

For more information:

National Digital Information Infrastructure and Preservation Program: www.digitalpreservation.gov

Office of Strategic Initiatives Annual Review:
<http://www.digitalpreservation.gov/news/pdf/OSI2005review.pdf>

Section 108 Study Group: www.loc.gov/section108

American Memory: www.loc.gov/memory