Doing More with Less: The Future of Digital Preservation in a Constrained Environment

Laura E. Campbell, Library of Congress (USA)

Abstract

When one of the world's great visionaries, Thomas Jefferson, sold his personal library to the Library of Congress in 1815, he had what was then considered to be a "universal library," not in the sense that he had every book ever published, but that his collection was universal in scope – covering all subjects of knowledge and encompassing volumes from around the world, many of them in non-English languages.

Jefferson's love of knowledge and belief that all fields of endeavor and creativity were worthy of inclusion in the national library are the guiding principles upon which the Library of Congress was founded in 1800. And they guide us today as well, even though the universe of what is worth collecting has expanded far beyond anything even Jefferson could have imagined.

Of course, all of you know the reason why: The digital age has led a revolution comparable to the one started by Gutenberg more than 500 years ago. We have more of everything – more books and other printed publications; media invented in the last century, such as film and sound recording; and media that only exists in bits and bytes.

Through all of these information upheavals, the Library of Congress's mission has remained unchanged during its more than 200-year history, as we struggle to maintain that "universal" collection that Congress and the American people expect of their nation's library. The collection, storage and preservation of this inexhaustible flow of information is a problem facing all archival institutions large and small, as well as private industry, state and local governments and federal institutions.

The Library of Congress can no longer collect everything, and it can no longer assume the costs of collection and storage as it did when everything we had to deal with was in analog form. We have thus formed a network of more than 100 partners from the public and private sectors to collect and preserve this content that is at risk of loss. And, just as important, the network is catalyzing a nation of digital preservation activists who will help us spread the word – as well as the technical expertise – on the importance and know-how of digital preservation.

Introduction

The National Digital Information Infrastructure and Preservation Program (NDIIPP)[1] was originally conceived and funded in 2000 with $100 million of no-year money (available until spent) to develop a national strategy for collecting and preserving digital content that is at risk of loss if not saved now. This includes Web sites, data sets, electronic journals, maps, legal documents, legislative deliberations, television, photographs, correspondence, diaries – both published and unpublished materials. Collections today may contain several types of electronic material: audio, video, text and images. Many people now recognize how much of the human record, in all fields and subjects, is produced in digital form only. In 2000 it was very forward thinking to recognize the need for a national strategy and solutions to save at-risk digital content, just as we did in the print world.

During the last seven years we have invested $46 million in modeling and testing a distributed and collaborative approach to sharing the responsibility for collecting and preserving content in the digital age.

Today we are in a drastically different budget climate than in 2000. Last year the program was converted to an annual appropriation at a reduced level of funding, requiring a restructuring of our investments moving forward. We have invested in saving at-risk content and building a network of partners among public and private organizations, including new services that support the work of the network partners. We have funded research in preservation solutions and developed a storage and transfer infrastructure for partner content. We have created shared tools for digital content management. And, we have tackled tough intellectual property issues.

Because of scarce resources (in terms of money, talent and technical solutions), we focused on maximizing what any one organization could do by bringing diverse parties together and taking early action to save content. We realized we would have to learn by doing, it was not possible to spell out a detailed plan in advance and the process would be iterative. Specifically, we made a first round of investments in seeding the network, funding preservation research and technical infrastructure, followed by a period of evaluation when we made adjustments based on what we learned that helped shape a second round of investments. It was catalytic, bringing parties together to experience a new way of doing business. We worked collaboratively in teams on issues of what to collect, technical developments and intellectual property issues.

This is truly a collaborative undertaking with partners sharing their expertise. The partners bring a host of skills that are complementary. The whole is greater than the parts.

Preserving At-Risk Digital Content

In January this year, we made grants to 21 states to jump-start their efforts to preserve important state and local government information in digital form.[2] They are our newest partners. In keeping with our networking philosophy, these states were not individually funded. Rather, they are working in four consortial enterprises, tackling common problems – and their solutions – with the other states. One effect of this is that, on their own, these states have formed groups to determine best practices for the preservation of their digital records and content.

What are these states doing?

- The Arizona State Library, Archives and Public Records is leading a project with Florida, New York and Wisconsin to establish a low-cost, highly automated information network that reaches across multiple states. Results will include techniques for taking in large quantities of state data as well as developing a strong data-management infrastructure. Content will include digital publications, agency records and court records.
- The Minnesota Historical Society is working with legislatures in its own state as well as California, Kansas, Tennessee, Mississippi, Illinois and Vermont to explore enhanced access to legislative digital records. This will involve implementing a trustworthy information management system and testing the capacity of different states to adopt the system for their own use. Content will include bills, committee reports, floor proceedings and other legislative materials.
- The North Carolina Center for Geographic Information and Analysis has teamed with Utah and Kentucky to focus on replicating large volumes of geospatial data among several states to promote preservation and access. The project is working closely with federal, state and local governments to implement a geographically dispersed content-exchange network. Content will include state and local geospatial data.
- The Washington State Archives is collaborating with Colorado, Oregon, Alaska, Idaho, Montana, California and Louisiana in using its advanced digital archives framework to implement a centralized regional repository for state and local digital information. Outcomes will include establishment of a cost-effective interstate technological archiving system, as well as efforts to capture and make available larger amounts of at-risk digital information. Content will include vital records, land ownership and use documentation, court records and Web-based state and local government reports.

The advantages of this multiplier effect became apparent when we made our first digital preservation awards – to eight consortia comprising 36 institutions.[3] These eight groups are saving and preserving geospatial information useful to Congress in its legislative activities as well as to all Americans and the world. They are collecting political Web sites that historians will need when they write about how electronic technologies radically altered the political landscape. Another project is documenting the birth of the so-called "dot-com era," of the 1980s, when Internet start-up firms with little cash but lots of ideas were able to attract big money to their dreams, many of which went bust in a Darwinian decade whose history must be preserved. We are working with National Public Television to preserve its programming, most of which is now digital. Those are just a few of the projects we have sponsored in the area of content collection and preservation.

Another way that we are maximizing these initiatives is that, in most cases, our partners are required to provide in-kind matching resources to these projects. So we get double the payback for our investments.

Supporting Digital Preservation Research

A second area of investments for us is in digital preservation research. Again, we have required that these 10 projects provide matching resources, and we have teamed with the National Science Foundation.[4]

In one research project, the University of Arizona has partnered with Raytheon Missile Systems to investigate the semantics of data provenance, including the development of ways to automate the capture of provenance information in new product design and development. The ultimate goal of the project is to enable the development of autonomic and interoperable enterprise data management systems.

Our collaboration with the University of North Carolina at Chapel Hill developed a preservation framework for digital video by working with the complete series of NASA broadcast educational videos.

In another project, with Old Dominion University, researchers explored options to reduce digital preservation costs through the use of cheap and widely deployed infrastructure protocols such as NNTP or SMTP, whose operational and maintenance burden is shared among many partners.

We also made investments in new businesses that can provide services for a fee to the network partners.

Building the Technical Infrastructure

Our third area of investment is in services that support the technical infrastructure. One storage and preservation solution we are working with is Stanford University's LOCKSS,[5] for Lots of Copies Keep Stuff Safe. The beauty of LOCKSS is its simplicity and cost-effectiveness. By storing redundant copies of electronic journals at various sites throughout the country, all partners in the project can take advantage of another's copy of a particular item, should their own copy fail.

The LOCKSS Program runs on open-source software that provides libraries with an easy and inexpensive way to collect, store, preserve and provide access to their own local copy of authorized content. Another way that LOCKSS saves money is that its governance and administration are distributed to ensure that no single organization controls the archive or has the power to compromise the content's long-term safety or integrity.

Members of our digital preservation network have access to an inventory of more than two dozen tools. For example, one tool validates the integrity of digital files through mathematical techniques. Its purpose is to ensure the authenticity of digital objects in long-term archives. Another tool, which we call Hub and Spoke, provides a method for exchanging digital files and metadata among different types of digital management systems built on different platforms. It provides basic interoperability between repositories via a common profile. A scalable Web crawler capable of fetching, archiving and analyzing Internet-accessible content is also available. These tools are benefiting a large community of users.

Because the preservation of digital content often requires the ability to transfer it, we are working with a dozen institutions in a project that builds on our Archive Ingest and Handling Test,[6] which looked at what happens when a large archive of digital content is transferred among institutions, in this case Harvard, Stanford, Old Dominion and Johns Hopkins universities. The

content under test was the 9/11 Archive of George Mason University.

The Archive Test was designed in part to test the hypothesis of shared value. By designing a test in which the participants exported files to one another, we were able to simulate, albeit on a small scale, some of the shared effort we believe is required for varying types of institutions to operate in a loose federation.

In our current content transfer scenario we are doing cutting-edge research into the handling of a very large and very diverse body of content. This content includes Web sites, television programs and geospatial data. Few institutions have experimented with the transfer of such diverse content in such large volumes and at such high speeds.

Building on our success with the Archive Test, we initiated a collaboration with the San Diego Supercomputer Center to conduct tests of third-party storage.

We provided 6 terabytes of Web sites that the Library has harvested and preserved as part of our Web capture project. And we also supplied 580 gigabytes of digital image files of various resolutions from our extraordinary collection of images taken just before the Russian Revolution.

Our two main objectives:

- For San Diego to host the Library's content reliably and return it intact
- For the Library to be able to remotely access, process, analyze and manage that content.

We learned a great deal from this project, which was conducted from May 2006 through October 2007. The results will have deep implications for all archival institutions that transfer and remotely store their digital assets.

Among the lessons learned:

- Network transfer may be cheaper than transferring content via disk
- Use the simplest tools possible that can handle the frequency of transfer you need and amount of data that you need to transfer
- There is a crucial difference between reliability and availability of data. In addition to 100 percent reliability you need to plan for 100 percent availability.

The final report of our work with San Diego is available on our Web site at www.digitalpreservation.gov.

Developing Tools and Services

In another collaboration, we recently completed an agreement with a large software company to make available on our Web site the format specifications for software such as Word, PowerPoint, Excel and Office.[7] This project demonstrates that our digital preservation activities have synergy with preservation-supporting actions in the private sector, in this case the opening of previously undisclosed documentation. It is also important to note that Microsoft asked the Library to make this information available because of the trust it has in our commitment to long-term preservation.

Our caretaking of these format specifications over the very long term ensures that this technical information will be available to all who need it to carry out preservation "treatment" of content in these formats in the future. We also hope that this will encourage other private sector companies with proprietary software to do the same.

In addition to collecting and preserving formats, we are working with myriad organizations in the public and private sectors to develop standards for digital preservation.

The Library and Xerox[8] are working together on a project to develop better ways to store, preserve and access treasured digital images. The collection includes such images as a picture of the Wright brothers' first flight at Kitty Hawk and a panorama of San Francisco after the 1906 earthquake.

We are studying the potential of using the JPEG 2000 format in large repositories of digital cultural heritage materials such as those contained in the Library and other federal agencies. The eventual outcome will be the existence of leaner, faster systems that institutions around the country can use to store their riches and to make their collections widely accessible.

The images to be used from the Library's collection are already digitized (primarily in TIFF format), but JPEG 2000, a newer format for representing and compressing images, could make them easier to store, transfer and display.

JPEG 2000 holds promise in the areas of:

- visual presentation
- simplified file management
- decreased storage costs.

It also offers rich and flexible support for metadata.

Xerox scientists are developing the parameters for converting existing TIFF files to JPEG 2000 and will build and test the system, then turn over the specifications and best practices to the Library of Congress. The specific outcome will be development of JPEG 2000 profiles, which describe how to most effectively use JPEG 2000 to represent photographic content as well as content digitized from maps. The Library plans to make the results available on a public Web site.

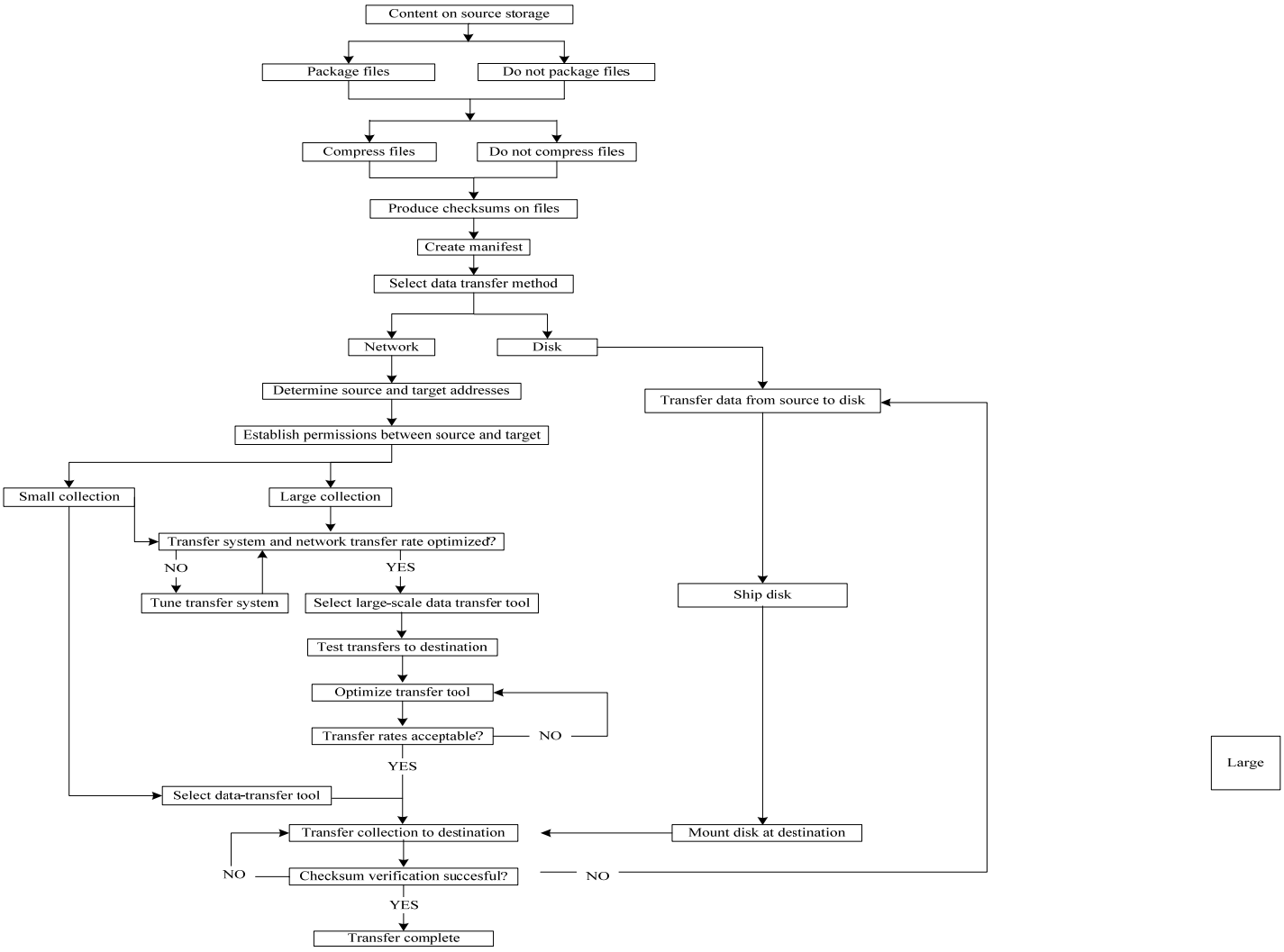
We have also taken the lead in setting standards used by the federal government. We have established multiagency working groups for still and moving images. We are thus taking the vast amounts of accumulated experience of these agencies and leveraging them for the benefit of all. We are working with 10 federal agencies in this project.

One of the projects the still-image group is working on is the development of tools to automate quality controls without human intervention.

Why? Because human judgment is very expensive, and human judgment is just that—everyone sees things differently and it also depends on environmental factors. The point is that not only can we not afford to have human resources devoted to comparing every digitized image to the original, we would not want that intervention even if we could afford it.

We thus need reliable tools that do the certain quality control work for us.

There are currently myriad standards related to imaging, but a lack of common, comprehensive guidelines. There is also a need for better workflow practices. Here is an example of one we used in the San Diego test:



Data Transfer Design Tree

This document describes decisions the Library of Congress has made in the course of its data-transfer tests. These decisions apply broadly to most data-transfer situations, and we hope this decision tree will help demystify the process.

The process consists of three overall parts:

- Preparing the data for transfer
- Transferring the data
- Confirming the successful transfer of the data.

We learned that:

- Although shipping disks is a familiar method, it has more possible points of failure, such as the disk being corrupted, dropped in transit or shipped to the wrong address.
- Network-based transfer is faster, more efficient and preferable, but it requires an initial investment in setup and optimization.

What It Costs:

- Bit storage costs are only a fraction of the total cost for "digital preservation"
- SDSC's sustainable single-copy 'bit storage' costs (2007 estimates):
    - Approximately $500/TB/yr for tape storage
    - Approximately $1500/TB/yr for disk storage
- Media costs are about 30% of the integrated 'bit storage' costs and total capital is about 50% of costs for both tape and disk
- Costs in dollars/per terabytes/per year increase, then flatten out and eventually slowly decline with the scale of installation
- Costs will decline with time, but the critical issue is which elements do not scale with media and technology advances
- Disk/tape integrated costs are converging.

Working with the Private Sector

Our ability to work with the private sector and with many different and often competing private sector entities will continue to play a key role in the distribution of the digital preservation burden among myriad organizations and institutions. Last August we announced eight partnerships with private sector creators and distributors of intellectual and creative content.[9] The partnerships grew out of a strategy meeting held by the Library in Los Angeles in April 2006 in which we gathered more than 50 private sector producers of digital content to assess their interest in, and plans for, the long-term preservation of their content. Participants in the meeting discussed a range of issues pertaining to digital preservation (e.g., standards) and explored potential relationships between the Library of Congress and those engaged in or associated with the creation of digital content in the United States today.

If you watched the Academy Awards in February, did you consider whether the nominated films, many of which were created digitally, will be accessible to future generations?

One of our partners in this program, which we call Preserving Creative America, is the Academy of Motion Picture Arts and Sciences, home of the Oscars. AMPAS has agreed to devote considerable resources to a host of motion picture-related educational, scientific and cultural endeavors, including the technical aspects of filmmaking and the preservation of motion pictures. The resulting Digital Motion Picture Archive Framework Project will build upon AMPAS's current research in digital preservation issues from the perspective of the major motion picture studios, extending the effort to include independent filmmakers and smaller film archives. Additional key components of the project will involve developing a case study system for investigating archival strategies for digital motion pictures and recommending specifications for image data formats across the production chain.

Even though digital is fast becoming the standard way in which we produce photographs, there is still no accepted standard set of rules for handling digital image files and maintaining information about them. The American Society of Media Photographers is acutely interested in this deficiency for obvious reasons and is working with us to expand an existing set of Universal Photographic Digital Imaging Guidelines, with recommendations for refined production workflows, archiving methods and best practices based on image use and capture methods. The Society will also work to promote the use of the guidelines through a Web site and awareness campaigns within the professional photographer community.

The adoption of digital recording has virtually eliminated the vital documentation once created on paper during the recording process. At the same time it has created tremendous unrealized potential to create and maintain all key information about a recording throughout its life cycle. The project with BMS/Chace focuses on creating a standardized approach for gathering and managing metadata for recorded music and developing software models to assist creators and owners in collecting the data. A standardized metadata environment will allow content creators, record labels, individuals and cultural heritage institutions to document, archive and manage "born digital" recordings effectively.

The works of "Doonesbury" creator Garry Trudeau and editorial cartoonist Pat Oliphant are national treasures. Together with the Universal Press Syndicate, we are modeling and testing the transfer of this digital content to the Library of Congress. The project is a case study for public-private partnerships for archiving digital content and will focus on aligning metadata practices, transfer procedures and continuing collection management in a manner consistent with the goals of digital preservation.

One of the most exciting projects in the Creative America initiative is the Preserving Virtual Worlds project, which is exploring methods for preserving digital games and interactive fiction. Major activities include developing basic standards for metadata and content representation and conducting a series of archiving case studies for early video games, electronic literature and Second Life.

The Library's Prints and Photographs Division is also working with the private sector in a pilot program with Flickr, the picture-sharing Web site. So far, more than 3,000 of the Library's most popular images -- from the Great Depression and from the 1900s through 1930s -- are reaching an entirely new audience for us.

And those users are reaching us. The Library has little documentation on many of the photos we have made available. Flickr users are telling us that they know where a certain photograph was shot, or that they recognize a relative in a photo or that they know the circumstances under which a photo was taken.

In other words, our users are helping us write the metadata at no cost to the Library.

A corollary benefit of the Flickr pilot is that it is helping us preserve the physical collections. The images we have made available are among our most used collections. For the majority of users, the digital images are sufficient, allowing the Library to increase the life of the physical prints and negatives. Low-resolution images are available through Flickr, which provides links to the high-resolution TIFF files on the Library's servers.

Recommending Changes to Copyright Law

Whenever you talk about preservation, you are almost always talking about the making of copies, and when you combine the making of copies with the rights of intellectual property owners, you wind up with a volatile mix of legal issues.

This is why the Library and U.S. Copyright Office, which is a part of the Library, formed the Section 108 Study Group.[10] Section 108 refers to the section of U.S. Copyright law that addresses the exemptions that libraries and archives have in order to fulfill their responsibilities to their users. Section 108 worked fine in the analog world, but the digital world presents an array of problems related to the fact that digital materials are so easily replicated. Once again, we have traded on our reputation as an honest broker to bring together 19 representatives of competing interests: libraries, content producers and content creators.

We will not be able to move forward with the digital preservation program until we resolve some of the intellectual property issues that hinder our work. The Study Group has just completed nearly three years of meetings, and its report makes recommendations for changes in the law that we believe will result in draft legislation for Congress, addressing exceptions for libraries and archives to collect, preserve and serve digital materials. We hope that libraries worldwide are able to benefit from this report.

That report, available on our Web site, recommends, for example, that the law be amended to permit a library or archives to authorize outside contractors to perform at least some activities permitted under section 108 on its behalf, provided certain conditions are met.

Under current U.S. law, there is a three-copy limit for libraries, modeled on best practices developed for microfilm preservation. But there appears to be no exact number of copies that would enable libraries and archives to preserve or replace analog works digitally, and it is impossible to anticipate how digital preservation technologies will develop. Even under current practice it is usually necessary to make numerous intermediate copies in order to generate a single digital "use" copy to replace a work in a library or archives collection. Over time, additional copies must be made to refresh and update that digital copy to ensure that it remains usable as technologies and formats evolve.

The complete list of recommendations by necessity travel through some very complex legal issues. Suffice it to say that unless copyright laws are amended for the digital age, the preservation of digital content will be severely at risk.

Conclusion

The Library and its partners currently have collected and preserved terabytes of at-risk digital content, and we have a plan in place to increase that amount nearly tenfold over the next five fiscal years. Our network of 130 partners in content, technology, research, government and business sectors reaches 25 states, and we plan to touch all 50 states by the end of 2013. We also have many partners throughout the world, and we are one of the charter members of the International Internet Preservation Consortium.

This National Alliance for Content Stewardship has developed, and will continue to develop, models of shared participation among preserving institutions that promise not to bankrupt the participants. Although the cost per unit of storage is still falling dramatically, that change alone is not enough to make digital preservation inexpensive. The steady decline in storage prices cuts both ways, making the creation of enormous new archives possible. Yet much of the cost in preserving digital material is in the organizational and institutional imperatives of preservation, not the technological ones of storage.

Institutions will only make the choice to share the burdens of preservation when they can do so without having to bear crippling costs, either financial or human. Given how much of the cost is in the initial design and implementation of a working architecture, anything that helps institutions adopt and effect digital preservation regimes will accelerate the spread of such systems. In addition, shared investment in handling digital data, in a manner analogous to union catalogs or shared collection building, will benefit from any efforts to simplify the sharing of data among institutions. The general principle is that reduced cost is an essential prerequisite to the spread of effective digital preservation.

Networking and collaboration in a supportive and catalytic environment in which everyone benefits are the keys to the success of digital preservation programs everywhere.

References

1National Digital Information Infrastructure and Preservation Program, www.digitalpreservation.gov

2 Digital Preservation Program Adds New Partners to Preserve State Government Digital Information, www.loc.gov/today/pr/2008/08-004.html

3 Library of Congress Announces Awards of $13.9 Million to Begin Building a Network of Partners for Digital Preservation, www.loc.gov/today/pr/2004/04-171.html

4 Library of Congress and National Science Foundation Announce Research Awards of $3 Million to Advance Digital Preservation, www.loc.gov/today/pr/2005/05-118.html

5 Library of Congress Announces Digital Preservation Award to Stanford University, www.loc.gov/today/pr/2006/06-129.html

6 Archive Ingest and Handling Test Final Report, www.digitalpreservation.gov/library/pdf/ndiipp_aiht_final_report.pdf

7 Sustainability of Digital Formats Web site, www.digitalpreservation.gov/formats/intro/specifications.shtml

8 Library of Congress Collaborates with Xerox to Test Format for Digitally Preserving, Accessing Treasured Images, www.loc.gov/today/pr/2007/07-213.html

9 Digital Preservation Program Makes Awards to Preserve American Creative Works, www.loc.gov/today/pr/2007/07-156.html

[10] Section 108 Study Group Report, www.section108.gov/