The International Journal of Digital Curation
Issue 1, Volume 3 | 2008

Evolving a Network of Networks: The Experience of Partnerships in the National Digital Information Infrastructure and Preservation Program

Martha Anderson,
Office of Strategic Initiatives,
The Library of Congress

July 2008

Abstract

The National Digital Information Infrastructure and Preservation Program (NDIIPP) was initiated in December 2000 when the U.S. Congress authorized the Library of Congress to work with a broad range of institutions to develop a national strategy for the preservation of at-risk digital content. Guided by a strategy of collaboration and iteration, the Library of Congress began the formation of a national network of partners dedicated to collecting and preserving important born-digital information. Over the last six years, the Library and its partners have been engaged in learning through action that has resulted in an evolving understanding of the most appropriate roles and functions for a national network of diverse stakeholders. The emerging network is complex and inclusive of a variety of stakeholders; content producers, content stewards and service providers from the public and private sectors. Lessons learned indicate that interoperability is a challenge in all aspects of collaborative work.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

When spider webs unite, they can tie up a lion. Ethiopian proverb

In the winter of 2000, a national digital preservation network began to form in the United States. The National Digital Information Infrastructure and Preservation Program (NDIIPP) was initiated by Congressional legislation that authorized the Library of Congress to work with other institutions to form a national network of partners dedicated to collecting and preserving important born-digital information. Guided by a strategy of collaboration and iteration, the Library of Congress and its partners have been engaged in learning through action (referred to as "learn by doing") that has resulted in an evolving understanding of the most appropriate roles and functions for a national network of diverse stakeholders.

Preserving our cultural heritage is not a mission that can be accomplished by a single institution. The amount of historical and creative content has reached astronomical proportions with the advent of the Internet. Technology has allowed any individual to become a publisher. Libraries and archives face a daunting task in their efforts to continue the tradition of preservation in the digital age. Although it was recognized early on that no one institution can do this alone, NDIIPP work has also taught us that the national network will be a complex interaction between networks rather than individual parties.

2000 to the Present: Developing the NDIIPP Network

When the U.S. Congress authorized NDIIPP, the Library started with the development of a plan and a strategy for moving forward. The plan, called "Preserving Our Digital Heritage," (Library of Congress, 2002) was approved by Congress early in 2003. During the development of the NDIIPP master plan, the Library met with hundreds of interested parties convened around the topics of preservation, technical architecture, research agendas and content collection and production. The result was the culmination of the initial research and planning phase and represented 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.

Congressional approval of the plan signaled the initiation of the first phase of network formation. The Library was eager to get to work on this exciting – yet daunting – program to save the creative and intellectual heritage of the nation in digital form. In the plan, the Library identified needs, how to address intellectual property issues and where to make investments of funding. Work began in three areas of endeavor – preservation partnerships, technical architecture and basic research.

Phase 1: Seeding the Network (2002-2005)

The first phase of network formation can be best characterized as launching small networks of partners with the common goal of preservation but with individual challenges of content and technical viewpoints. It began in September 2004 when NDIIPP funded content collecting and preservation projects comprising 36 institutions working with eight consortia. Each project consortia focused on specific content types and developed relationships and processes around the content. Each project set its own

technical agenda and devised its own methodology. In this phase there was an emphasis on "learn by doing."

This first set of preservation partnership investments totaled nearly US$14 million in funding to eight projects comprising 36 institutions. These institutions are selecting, collecting and preserving important digital materials such as:

- Geospatial data
- Social science datasets
- Political Web sites
- Historical and cultural materials from the American South
- Public television broadcasts
- Business records from the birth of the Dot-com Era

Other partners, added in later years, collecting important content so that it is available for future research are Portico, which is developing an archiving service for electronic journals; 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; and LOCKSS, a multi-site distributed archive of content.

Although the projects were developed around diverse content types, their activities 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 in which they are now engaged

In May 2005, the Library and the National Science Foundation awarded 10 university teams a total of US$3 million to undertake pioneering research to support the long-term management of digital information. These basic research awards were the outcome of a partnership between the two agencies to develop the first digital preservation research grants program[1].

A test completed in June 2005, called the Archive Ingest and Handling Test (AIHT), serves as an example of how NDIIPP is catalyzing joint problem-solving to achieve programmatic goals. AIHT tested the ingest of a digital 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 coincidently. We learned which aspects of digital preservation are institution-specific and which aspects are more general. In fact, the

[1]The Library of Congress: Digital Preservation: Partners http://www.digitalpreservation.gov/partners/presNSF.html

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 and cooperation among various institutions exists to the degree necessary to ensure interoperability (Wilson, 2003).

In a related project, the Los Alamos National Laboratory Research Library worked on building mechanisms to address challenges related to collecting, storing and accessing complex digital objects. 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 (Bekaert & Van de Sompel, 2005). In addition to the three investment areas, the Library 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 learned early on that we would not be able to move forward with the digital preservation program until we had 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, was convened in April 2005 under the sponsorship of the Library and the U.S. Copyright Office. Its objective was 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 made recommendations in March 2008 for changes that result in draft legislation for Congress, addressing exceptions for libraries and archives to collect, preserve and serve digital materials[2].

Lessons learned while seeding the network highlighted the individuality of institutional business processes and constraints. The characteristics of these initial collaborations were very much project-centric. The consortia and the Library were challenged by mechanisms for cooperative agreements and distributing funding across federal, state and private institutions. Bringing business relationships across such diverse organizations into agreement consumed time and resources. Early partnership building success was often marked by the completion and signing of a cooperative agreement.

Phase 2: Strengthening and Expanding the Network (2006-2008)

The current phase of the Digital Preservation Program is intent on strengthening and sustaining current partnerships while adding new types of partners and identifying tools and services for the network. The work of the first phase informed the second and current phase of network formation. In January 2005, the collecting and preserving partners identified some common tools and services needed to preserve digital content. Tools to work with metadata and tools to examine, characterize, and verify file formats

[2] Section 108 Study Group http://www.loc.gov/section108

were of highest priority. One of the most important services is storage for large volumes of files.

In May 2006, the Library began a 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 the Library's digital content and guarantee data integrity and access
- enable the Library to remotely access, manage, process and analyze that content

Two new communities are being developed during this phase – one with state libraries and archives, and another with the commercial content producers. The result of workshops conducted in 2005 with state librarians, archivists, and records managers informed a plan to fund multi-state demonstration projects whose results will assist all states in making decisions on preserving records and other state data that are increasingly available only in digital form.

Another set of investments is addressing the long-term preservation of creative content in digital form. Eight *Preserving Creative America* projects target preservation issues across a broad range of creative works, including digital photographs, cartoons, motion pictures, sound recordings and even video games. The work is being conducted by a combination of industry trade associations, private sector companies and nonprofits, as well as cultural heritage institutions. Several of the projects involve developing standardized approaches to content formats and metadata, which are expected to increase greatly the chances that the digital content of today will survive to become America's cultural patrimony tomorrow.

Although many of the creative content industries have begun to look seriously at what will be needed to sustain digital content over time, the *Preserving Creative America* projects will provide added impetus for collaborations within and across industries, as well as with libraries and archives. The awards also allow the Library to respect Congress's wishes that we enlist the private sector to help address the long-term preservation of digital content.

This phase can be characterized as one in which the partners identified common tasks and worked across projects. The partners began to form the larger network and during this phase identified functions of a preservation network that are applicable to a variety of content communities. In this phase the program began to see the emergence of defined partner roles within the network and the emergence of communities of practice. The partners as consortia identified their strengths and areas in which they could provide leadership and expertise to other partners.

Phase 3: Sustaining the Network (2008-2009)

The National Digital Information Infrastructure and Preservation Program is building a stewardship network of partners that operate in one or more functional roles:

(1) Committed Content Custodians;
(2) Information and Expertise, Development and Dissemination;
(3) Services; and

(4) Capacity Building.

Together, the interaction of partners playing various roles will strategically provide the necessary content, support and services to all the network's members. A layered model (Figure 1) illustrates the single or multiple roles any one organization would fulfill (Arms, 2006).

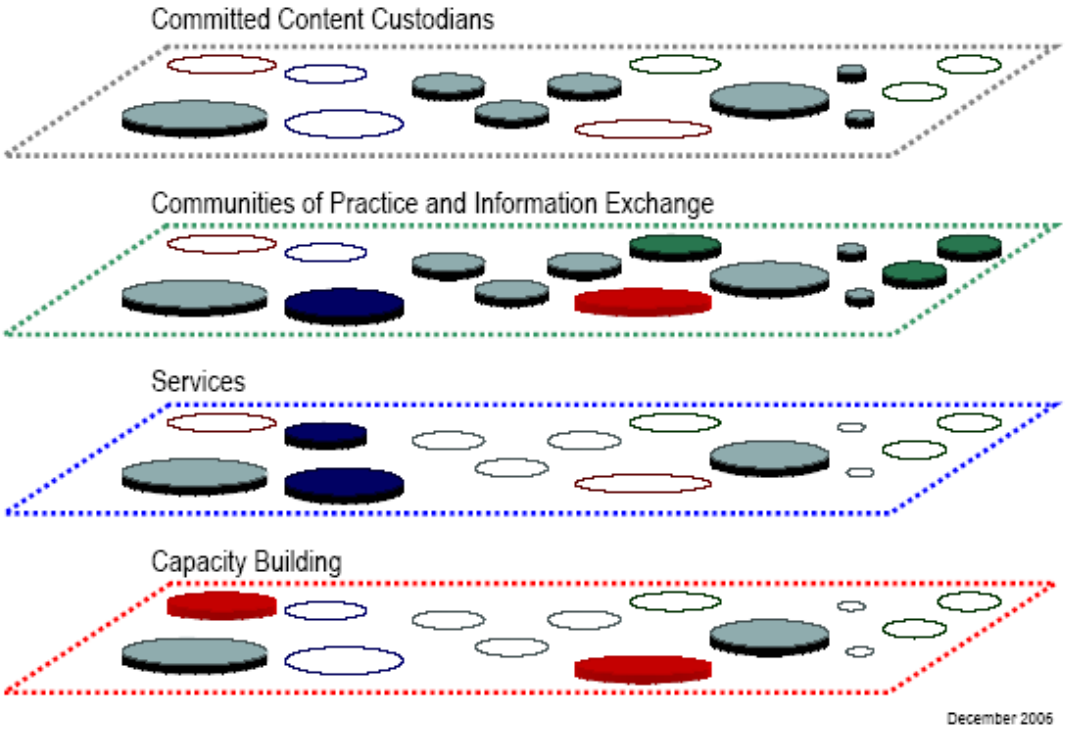

Figure 1. Layers in a Stewardship Network.

Layer 1: Committed Content Custodians.

The work of this group goes to the very heart of what NDIIPP is trying to achieve, for our committed partners have accepted responsibility for collecting specific types of digital content as part of what will collectively become the universal digital library. As of mid-2007, these partners have saved more than 66 terabytes of important at-risk content. The actors in this layer include government agencies such as national archives and libraries, universities, state libraries and archives, and special domain archives.

Layer 2: Information and Expertise, Development and Diffusion.

In this layer the focus is on the activities rather than on the actors, or partners themselves. It is where overlapping communities of practice will be constructed. It is the place to roll up your sleeves and make a contribution that assures that the content collected in Layer 1 is available to future generations.

Depending on the task, some activities may come and some may go. Organizations in this layer may see themselves more as "contributors" than as permanent "members." For example, standards-setting bodies will play a role, yet the work they do, while crucial to NDIIPP, may not be carried out specifically for the program.

Layer 3: Services.

This layer is for services that further the objective of long-term access to digital information. The players in this layer perform work that is useful to many entities and their work involves more than the mere sharing of expertise and information. Their services range from being totally centrally supported by the network to being partially subsidized, to being commercial and for-profit. Some examples of services in this layer are tool and format registries or Copyright registration and deposit tools. Open-source software developers such as DSpace and Fedora live here. One of our NDIIPP partners, LOCKSS, also inhabits this space.

Layer 4: Capacity Building.

The players in this layer will seek or provide funding and other support for activities in all four layers. The funding will be expected to produce results that benefit entities across the network. Government agencies at the federal, state and local levels will have a part to play. They will support training and education and the development of curricula in digital preservation.

Outcomes for this phase of NDIIPP work in addition to network formation include addressing access, developing and formalizing roles in the network, adopting interoperability standards, continuing to develop and refine services, issuing a plan for collecting content and providing a content directory of what has been saved so far. It is expected that by this time the recommendations for legislative changes suggested by our Copyright Study Group will be reviewed. Without changes to U.S. Copyright law, it will be difficult for libraries and archives to serve these materials to their users without violating intellectual property rights. A long-term funding strategy will help ensure that partners are able to continue their important work and their contributions to our universal digital library.

Phase 4: Formalizing the Network (2010-2015)

Although the rapidly changing technology and political landscapes make it difficult to project this far ahead, we do know we will formalize the network even further through a broader and deeper range of partners. Organizational roles and responsibilities will be refined and adopted. By this time, we hope that public awareness of digital preservation and why it is needed will be clear to policymakers, scholars and students as well as the general public. The vision is that of a galaxy of networks of content creators, producers and stewards interwoven with networks of service providers collaborating on standards and practices that sustain a large valuable collection of digital content.

Lessons Learned Within the Network of Partners

Collaborative digital preservation endeavors most often begin with metadata or format standards and workflow practices in order to promote interoperability. As essential as these efforts are, interoperability has become the signal word for agreement. One of the early lessons of the NDIIPP work is that there are interoperability challenges in every phase of the life cycle of digital objects. This paper highlights the challenges and some agreement in the planning and management, data curation and stewardship functions of the life cycle.

Planning and Management of Partners

Lesson: Organizational operations and business practices do not interoperate.

When establishing collaborative relationships between public and private, academic and government, commercial and academic, business operations are not always interoperable. In the NDIIPP experience, accounting systems and business practices of academic institutions are not compatible with federal government grant and procurement requirements. Monthly reporting and invoicing are very difficult for academic organizations but very common for public and commercial organizations. Sorting out roles and responsibilities around the acquisition and use of digital content is also very challenging because previous methods for inventory, acquisition and preservation are predicated on physical copies that can be more tightly controlled. Even within the same organizational domain, there are barriers to collaboration. The five universities engaged in the NDIIPP MetaArchive Project tackled the business relationship challenge by establishing a nonprofit (U.S. 501 (c) (3) nonprofit corporation called Educopia (McDonald & Walters, 2007).

Lesson: The NDIIPP network is an emergent network.

NDIIPP stated the goal of assembling a distributed network of partners as a strategy for digital preservation. The network was not constructed but rather is emerging from the work of the partnerships (Milward & Provan, 2006).

The consortia comprising the first NDIIPP funding initiative were selected for their demonstrated abilities in three areas: content selection, technical capacity, and potential to organize a network of partners. The bi-annual NDIIPP partners' meetings brought together participants from all NDIIPP-sponsored projects including the joint NSF/LC DIGARCH partners. The lesson learned is that although these partners shared a common interest and often articulated common problems, their work with diverse data and data communities was not conducive to thinking and working as a larger network. They needed some working sessions to discover and leverage the beneficial relationships.

One such session led by Clay Shirky asked the partners to identify project strengths that would benefit other projects. This exercise influenced the layered view of the stewardship network (Figure 1). It brought to the attention of all that the program needed to invest in projects that produced tools and services useful across the partnerships and that the partners needed to work together on common issues such as intellectual property, sustainability and collection policy.

At year three, there have been working sessions to define requirements for tools for file identification, verification and validation, metadata transformation and capture of content from the Web, shared storage solutions to address the growing demand for large volumes of data especially in the geospatial and Web domains, and shared collection development within the Web archiving domain. Established tools such as LOCKSS are being applied to varied data types to meet the needs of special preservation systems. There is a practical collaboration on large-volume transfer and storage mechanisms between the NDIIPP National Geospatial Digital Archive and the NSF DIGARCH research project at the University of Tennessee.

Data Curation and Stewardship

Lesson: Working within a diverse partner network increases the complexities for data interoperability.

From the viewpoint of the NDIIPP original eight preservation consortia, the OAIS concept of designated communities has been borne out for content types--social science datasets require different workflows and standards and currently serve different research communities from geospatial data. Interoperability challenges become greater as the designated user communities broaden their interest in content from various producer communities. An example is the wide adoption of geospatial data for commercial content services. From the NDIIPP partners, it is conceived that a useful research corpus could include political and government Web sites, polling data from the social science community and geospatial data created by state and local governments. Each of these data communities has vastly different metadata standards and practices. Access points are different. The content itself is comprised of a variety of formats that, in all three cases require different software for retrieval and display of the objects.

Lesson: Metadata in standardized formats very often represents an institutional context not easily transferable to a larger context.

The excellent work on metadata schemas over the last several years was useful to NDIIPP project teams but in the 2003 Archive Ingest and Handling Test (AIHT) it was revealed that each institution employed a different grammar for the same schemas. No two METS applications were alike even though there was some transferability across local archives. Clay Shirky, technical advisor for the project observed, "The goal should be to reduce, where possible, the number of grammars in use to describe digital data and to maximize overlap, or at least ease of translation, between commonly used fields. But it should not be to create a common superset of all possible metadata" (Shirky, 2005).

Lesson: At this time, the greatest common ground for preservation process, tools and standards lies at the bit level for digital content.

The Library is currently testing a process and protocols for transfer and verification of diverse data that can be applied at the bit level. All eight consortia are participating in an exercise to move content to an archive at the Library with a file manifest and a package manifest that provides information at a minimal level to retrieve and return the package to the source as well as to plan for more than just archival storage should that scenario develop (Sugimoto, 2006). Plugin architectures allow for diverse use of common validation and format tools such as JHOVE. The AIHT project demonstrated that interoperability for long-term preservation is data-centric and not system-centric. Common tools for data analysis, verification and validation were very useful but project participants cautioned against universal use of a single tool due to the possibility of inherent assumptions and logic that may not provide complete coverage and extraction of useful information.

Conclusions

Through NDIIPP, the original strategy of "learn by doing" has revealed the emergence of a complex network of partners that is best described as a network of networks. Each content community has identified, and is well on the way to solving, specific challenges for each content type – geospatial, digital television, Web content, digital images, digital sound recordings and datasets. At the same time, the various partners brought together through the program projects have been able to recognize and define functions that are best addressed through collaborative and common work. Each of the networks brings expertise and skill of value to the whole network. This network is organic rather than constructed and becomes stronger through shared expertise and common goals.

Acknowledgements

Over the years, the NDIIPP staff and partners have benefited from a valuable opportunity to gain a greater understanding of the timely challenges of digital preservation. This paper presents the work to-date of the NDIIPP program management staff and project teams. Thanks are due to their hard work and continued efforts.

References

Arms, C. (2006). *Sample layers in a stewardship network*. Internal draft. Washington, D.C.: The Library of Congress.

Bekaert, J., Van de Sompel, H. A. (2005, June). Standards-based solution for the accurate transfer of digital assets. *D-Lib Magazine vol. 11, (6).* Retrieved July 25, 2008, from http://www.dlib.org/dlib/june05/bekaert/06bekaert.html

Library of Congress. (2002). *Preserving our digital heritage: Plan for the National Digital Information Infrastructure and Preservation Program*. October 2002. Retrieved July 25, 2008, from http://www.digitalpreservation.gov/library/pdf/ndiipp_plan.pdf

McDonald, R. H., & Walters, T.O. (2007). Sustainability models for digital preservation federations. Paper presented at *DigCCurr 2007, Chapel Hill*.

Milward, H.B, & Provan, K.G. (2006). *A manager's guide to choosing and using collaborative networks*. Washington, D.C.: IBM Center for Business of Government.

Shirky, C. (2005, December). AIHT: Conceptual issues from practical tests. *D-Lib Magazine vol. 11, (12)*. Retrieved July 25, 2008, from http://www.dlib.org/dlib/december05/shirky/12shirky.html

Sugimoto, S. (2006). A collaboration model between archival systems to enhance the reliability of preservation by an enclose and deposit method. Presentation in *DLF Fall 2006 Forum Program*. Washington, D.C.: DLF. Retrieved July 25, 2008, from http://www.diglib.org/forums/fall2006/presentations/sugimoto-2006-11.pdf

Wilson, B. (Ed.). (2005, December). Five views of the archive, ingest and handling test. *D-Lib Magazine vol. 11,(12)*. Retrieved July 25, 2008, from http://www.dlib.org/dlib/december05/12editorial.html