Sustainability Models for Digital Preservation Federations

Robert H. McDonald
Chronopolis Project Manager
San Diego Supercomputer Center
University of California, San Diego
Email/AIM: mcdonald@sdsc.edu

Tyler O. Walters
Associate Director for Technology & Resource Services
Library and Information Center
Georgia Institute of Technology
Email: tyler@gatech.edu

Abstract

The Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIPP – www.digitalpreservation.gov) has brought together a variety of partners over the last few years (2004-2007) to take an in-depth look at various digital preservation technologies and the organizational strategies to implement digital preservation programs (LeFurgy, 2005). This national program has been a success and the Library of Congress plans to extend it through 2010 to develop these partnerships and solidify the sustainability of the preservation initiatives that have emerged.

Like many such national programs, NDIIPP is both a benefactor and catalyst of applied and theoretical work on digital preservation being conducted in the United States. This paper will present the sustainability models that have emerged at the international, national, state/regional, and local levels and contrast those with models that have emerged out of the work of the NDIIPP partnership known as the MetaArchive Cooperative (www.metaarchive.org). This overview will address governance issues for federated digital preservation programs that span across unilateral institutional governance.

Introduction

Based on the major findings published in the Task Force on Archiving of Digital Information (Waters and Garrett, 1996) it was clear to the authors as early as the mid-90s that:

> *Long-term preservation of digital information on a scale adequate for the demands of future research and scholarship will require a deep infrastructure capable of supporting a distributed system of digital archives.*

This paper explores current and proposed models of governance and institutional support for the sustainability of digital preservation initiatives, evolving over the course of the last ten years. As most of these programs have evolved in the last five years, it is critically important to combine information on them into a single document that reviews governance and sustainability and discusses options for future growth. This type of medium-range planning (Barton and Walker, 2003) will be necessary for digital preservation programs at all levels as the growth of born digital information begins to outweigh that of print collections (Beagrie and Greenstein, 1998). Curators of these digital resources (Hedstrom and Montgomery, 1998) will become the collaborative stewards of information, existing across many types of unilateral entities and will be responsible for building the basic collaborative infrastructures necessary to support a distributed system of digital preservation archives. The management and organizational skills curators will need to accomplish this monumental task will also be examined.

Organizational and Governance Models for Digital Preservation

In his 2003 Council on Library and Information Resources Report on National Digital Preservation Initiatives, Neil Beagrie utilized surveys and personal correspondence to construct an overview of national-level digital preservation initiatives. His purpose was to provide an event horizon for the newly formed Library of Congress NDIIPP program. In this document, we will take a similar approach. However, the authors will draw from selected examples to further elucidate the semantic model for digital preservation collaboration described in the diagram below. Figure 1 shows a prototypical collection disbursement across the digital preservation collaborative spectrum. This includes preservation components at the local institutional level, the state or regional levels, as well as the national and international levels. In the following sections, these layers are broken down with current examples to show the benefits and collaboration involved. It is envisioned that private and non-governmental entities will overlap with collaborations at the state, regional, national, or international levels depending on their funding and the missions of their charter institutions.

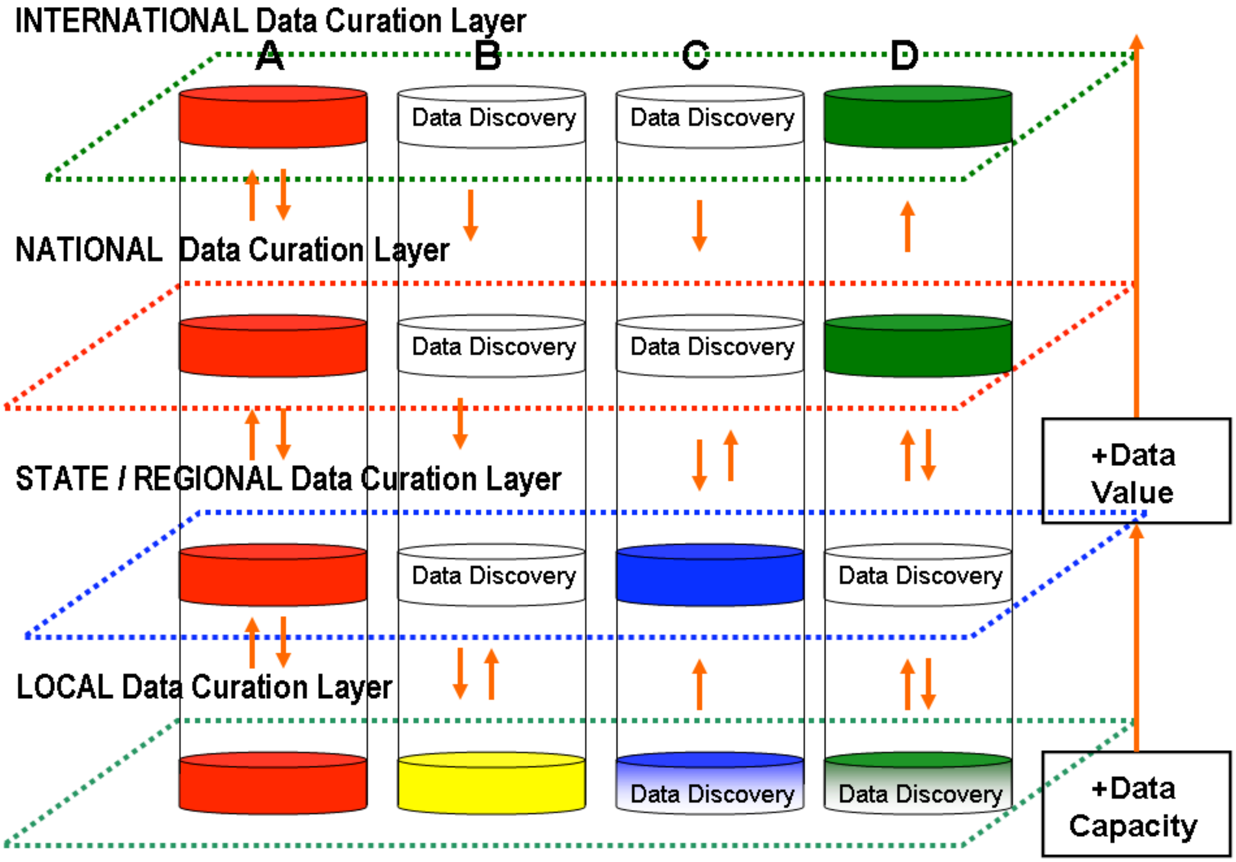

Figure 1 – Figure 1 shows a model comparison of four collections (**A, B, C, D**) as ingested within the collaborative data curation layers of sustainable governance. **Model A** shows a collection that is of moderate size but that has immense value to the public. **Model B** shows a collection that is of value only to the local organization. **Model C** shows a collection that is too large in size to be stored locally and is being curated at a state/regional/private consortium to take advantage of infrastructure scalability. **Model D** is most likely a large scientific data set that is being stored locally for short-term use and being stored nationally and internationally for long-term reuse and preservation. Note that data sets are discoverable from any repository layer regardless of where data is stored.

In order to understand how these layers of collaboration will function across homogeneous and oftentimes institutional-based repositories, concrete examples are given. The following examples highlight an individual collaboration that is either ongoing or in initial stages and then a collection of data or digital objects that is related to the project. Each collection is described within the parameters of generic collections A, B, C, and D in Figure 1. A fifth model type, used as a framework by non-governmental member-based organizations operating under nonprofit status, will also be discussed. This fifth model is similar to the national or regional models shown above depending on the mandate of the organization.

International

The International Internet Preservation Consortium (IIPC, 2007) provides a perfect example of international collaboration. This group is comprised of the following partners:

- Bibliotheque Nationale de France (National Library of France)
- British Library
- The European Archive Foundation
- Internet Archive (U.S.)
- Kansalliskirjasto (The National Library, Finland)
- Koninklijke Bibliotheek (National Library of the Netherlands)
- Kungl. Biblioteket, National Library of Sweden
- Landsbokasafn Islands – Haskolabokasafn (National and University Library of Iceland)
- Library and Archives Canada
- Library of Congress (U.S.)
- The Library of Virginia (U.S.)
- Národní Knihovna České Republiky (National Library of the Czech Republic)
- National Archives and Records Administration (U.S.)
- National Library of Australia
- Netarchive.dk (The Royal Library, Denmark)
- Nasjonalbiblioteket (The National Library of Norway)
- Schweizerischen Nationalbibliothek (NB)(Swiss National Library)

The IIPC is comprised of state and national libraries and private and non-governmental agencies working to preserve Internet content that spans the globe. During the course of this initiative, the content involved will represent content Model A and possibly content Model D in Figure 1 in that the web content will have a local home and be collected at the state/regional/non-governmental agency level as well as at the national level by the appropriate national library. This material will inevitably grow to such a state and size that only like-minded national libraries or independent entities such as Internet Archive or Chronopolis ™ (Moore et. al, 2005) will be able to provide adequate storage, thus making some of this content fit within Model D and requiring deep infrastructure to support large-scale web archiving.

National

Work done by the U.S. Library of Congress within the MetaArchive Cooperative illustrates how national projects can fit within the multiple layers (local/state/regional) of the model represented in Figure 1. The authors would like to point out that currently there are national programs well underway in Australia, Canada, France, Germany, Italy, Netherlands, and the United Kingdom at a variety of levels and governance. However, most follow the national library or national archives model for collaboration, and some, but not all have digital depository status as part of their national legal mandate.

The NDIIPP Partnership sponsored by the Library of Congress started work in 2004 creating eight original partnerships (LeFurgy, 2005). These partnerships were designed to fit the model of collaboration shown in Figure 1 with the Library of Congress being the national entity and the eight partnerships serving as state/regional aggregators of content. This program required that the digital content already exist at the local institutional level for funding to occur. The MetaArchive Cooperative (Arms, McDonald, Nicol, Walters, 2005) was a prototypical NDIIPP partnership, albeit one of the largest partnerships funded under NDIIPP as it brought together southeastern cultural heritage materials from the primary local institutions represented by the research libraries of Auburn University, Emory University, Florida State University, Georgia Institute of Technology, University of Louisville, and the Virginia Polytechnic Institute and State University.

State/Regional

In this layer, we will look at two models, one which falls under state university governance and one which falls under a state agency governance model. These examples are the University of California (UC) California Digital Library's (CDL) Digital Preservation Repository (DPR) and the State Library of Virginia's Archiving the Web Program. Both of these programs are working to preserve content for their parent institutions, in this case the University of California and the State of Virginia.

CDL is examining programs that benefit all ten of the UC system campuses as well as their respective research libraries, thus fitting Models A, B, and C in Figure 1. Content such as that hosted in the Calisphere collection, which is ingested into the DPR via local selection at a UC system research library, fits the framework of Model A. This collection is brought into an aggregated collection like Calisphere, which, depending on content, would reach a national and/or international preservation level in terms of representing the history of the State of California. Representative of Model B, content that is strictly local in importance, such as archival business or legal records, would have a short life-time retention cycle. Content such as large-scale web crawls of the entire UC system possibly can be stored in local infrastructures, but would require tremendous resources and is representative of Model C.

The State Library of Virginia is archiving web sites based on a hierarchal system tied to various Executive Administrations of the State of Virginia. While this is analogous to work being done on a national level by the U.S. National Archives and Records Administration (NARA) Electronic Records Archives (ERA) (NARA, 2007), it is becoming an increasing burden for state archives and libraries to host archival and access copies of this born digital material. Parts of this test case were accomplished in conjunction with the private entity Internet Archive and their Archive-It program. More needs to be done to distribute the infrastructure necessary to preserve this material. How can entities such as a state university system library support organization, like CDL, and their respective state libraries and archives, work to provide state and regional infrastructures that are supported by national and international replication?

Local

To provide examples of local models, the authors will examine a prototypical IR and its place within the state/regional and national context. The Georgia Tech institutional repository, SMARTech, has been in operation since 2004 and contains about 11,000 items. In the past year, it has been searched over 1 million times and nearly 500,000 items were downloaded. Powered by the DSpace open-source software, it contains a heterogeneous mix of annual reports, conference papers and audio/video of sessions, lecture series materials and recorded symposia, pre-prints and post-prints, proceedings, research and technical reports, white papers and work papers series, theses and dissertations, and other student works. Initiated and supported centrally by the Georgia Tech Library and Information Center, this repository is far from being an isolated entity.

SMARTech is enmeshed in many other technology infrastructures, ranging from being harvested by a statewide repository (GKR) and utilities like Google Scholar and OAIster to being preserved through the MetaArchive Preservation Network, built by the MetaArchive Cooperative, an LC/NDIIPP partnership. Metadata for SMARTech content will appear in a newly planned statewide institutional repository, the Galileo Knowledge Repository (GKR) (Jannik, Graham, Spasser, 2005). The GKR effort, which is being led by the University of Georgia, Georgia State University, the Medical College of Georgia, and Georgia Tech (under the aegis of the University System of Georgia), will further integrate SMARTech content at the state university system level and represent the type of arrangement shown in Model D of Figure 1. A private, regional approach to digital preservation has been implemented since 2004 for SMARTech. The content is harvested via a modified form of the LOCKSS software and is distributed and replicated in the six-university MetaArchive Preservation Network, the first established LOCKSS private network. The organizational and sustainability model of the MetaArchive Cooperative is described as a “non-governmental, private network” in more detail below.

Non-Governmental and Private

The models and examples described above indicate federated digital preservation activities at the international, national, state/regional, and local levels. In addition to these, there are several federations that have formed following a different model, one supported by a nonprofit status within the United States. These organizations are focused on digital preservation but have based their governance models around the type of nonprofit status entity that has proven successful with the open source software development community (ASF, 2007). These are often referred to as private or non-governmental member-based organizations. They frequently are comprised of like-minded organizations that want the freedom of working in a collaborative environment outside of the bounds of a typical university or library setting. Three prime examples of this type of organization within the digital preservation community are the LOCKSS Alliance (LOCKSS, 2007), the DSpace Foundation (Walker, 2006), and the LC/NDIIPP partnership known as the MetaArchive Cooperative and its nonprofit management entity, the Educopia Institute (Educopia Institute, 2007).

In order to provide sustainability for the MetaArchive Collaborative outside of the NDIIPP infrastructure, the members have formed a nonprofit private (U.S. 501(c)(3) nonprofit corporation) entity called Educopia (Walters, 2006).

The mission of the MetaArchive/Educopia Cooperative’s Preservation Network is to:

“…create an enduring and stable, geographically dispersed archive of digital materials pertaining to the American South that can, if necessary, be drawn upon to restore collections at the originating institutions.” (Educopia Institute, 2007)

Once the MetaArchive Cooperative was initiated, its steering committee began investigating how to sustain the Cooperative’s organizational and technological models. The partnership with LC/NDIIPP gave the MetaArchive access to a wide variety of financial and human resources and placed its work within the context of a national digital preservation agenda. Access to the other NDIIPP and DIGARCH research project partners has supported the MetaArchive’s work and helped spur on its innovations.

To achieve a measure of organizational structure and sustain the MetaArchive's activities, a cooperative charter for the MetaArchive has been developed and is supported by all institutions involved. The Charter details the types of partnerships supported by the MetaArchive Cooperative which include:

> Sustaining Partners – develop and test the MetaArchive Preservation Network technology, operate a preservation node, and engage in activities that sustain the Cooperative as an organization
>
> Preserving Partners – operate a preservation node, ingest collections from partner institutions, and make the node available for testing. (Development and Preservation partners currently are comprised of the original MetaArchive partners).
>
> Contributing Partners – as cultural memory institutions, possess digital content to preserve via the MetaArchive Preservation Network. They contribute fees for this service and do not operate a node.

The Cooperative Charter further addresses the MetaArchive Cooperative's organization, governance, and communication structures. These are largely accomplished through committees, which include the steering, content, preservation, and technical committees. Individual representatives from partner institutions serve fixed terms on the committees, ensuring broad participation in MetaArchive activities. The Charter also details the services provided to partners, including digital preservation (network development and maintenance, content ingestion, and retrieval), format migration, digital collection disaster recovery, digital preservation network consulting, and LOCKSS services. Lastly, the Charter also includes technical specifications for the MetaArchive Preservation Network that development and preservation partners must follow and a copy of the memorandum of understanding between the six universities. After reaching consensus on the Cooperative's basic organizational and operational structure, the issues of organizational sustainability and technological growth and innovation became the next focus of the steering committee.

Three aspects of sustainability are chiefly being considered: the continuing need for financial resources, the integration of MetaArchive work with other digital initiatives to inform the Network's continued development, and the need for an economically efficient management structure to support the activities. The partners realize these needs represent much more than the technical work on which they initially agreed. Subsequently, the steering committee explored the idea of establishing a nonprofit management entity to host and guide the MetaArchive Cooperative. The result was the Educopia Institute, a nonprofit organization that provides oversight for the Cooperative and for other future digital projects. The intention of Educopia is to provide a low-cost, low-overhead conduit for completing digital library and scholarly communications projects that will advance the digital preservation infrastructure needed to support modern research, teaching, and learning. Educopia's five-member board of directors, which have MetaArchive representation through two officers from Emory and Georgia Tech, will review new projects, collaborations, awareness-building activities, and funding possibilities that will advance digital preservation and broader stewardship goals.

Management and Organizational Skills for Digital Curators

In light of the federated governance models described, what skills do digital curators need to initiate and sustain digital preservation federations?

Beyond the realm of technical skills required to actually "curate" digital collections, there are many organizational and relationship-building skills that are an absolute. In many instances, depending on how much a library can afford to invest in local digital curation, this organizational role may be the only position a curator holds and everything else may be outsourced to federations at the state, regional, or national level. Increasingly, digital preservation work is carried out in consortially-based arrangements or locally in repurposed staffing and workflow (Merrill, Morrow and Roosa, 1991)(Kennedy, 2005). The central task becomes that of building an organizational structure that facilitates the technical work to be done: the fundraising required; and the research, development (NSF and LC, 2004), and standards-building that supports future digital curation activities (Walters, 2006).

Simply put, today's digital curators must be capable of relationship-building. They must be effective project managers and understand the technical requirements needed to reach out and engage and solicit potential partners. They must be able to describe the challenge at hand, visualize the future solution, and communicate it effectively to others. Once initial partners come together, this digital curator will need to facilitate discussions regarding the roles and responsibilities of each institution. This is always a delicate dance, making sure that each partner is integrating their strengths, while addressing their particular challenges. Trusting partners to help design your solutions can be tough. Diplomacy, persistent communication, and transparency are required to work through difficult issues and see that the relationship is beneficial.

There are many other highly desired skills, such as the ability to lead or interact firsthand in fund-raising initiatives and garner other resources through persuasive communication. Project management skills in a highly diffuse and complex collaborative setting involving emerging technologies are paramount. Also, keeping in mind the crucial importance of knowledge sharing throughout the project, both internally and with external groups who are focused on similar issues, will help greatly as federation members strive for new and successful technical approaches. The 21st century digital curator is a renaissance person, both technical and intensive, but who thinks expansively and is people-oriented. This type of curator fosters collaboration by regarding human resources and seeing the value in joint efforts. Only with this complete set of skills will they be able to construct the preservation federations necessary to meet current and future challenges to the longevity of digital information.

In Figure 2, we have taken the model for a digital curation manager and mapped it into a diagram, showing the typical influx of knowledge, skills, and training as applied to organizations involved in managing digital repository infrastructures. In addition to technical documentation and policy planning, of prime importance is an ability to build internal institutional knowledge of the ongoing long-term collaborations that are supporting the digital curation lifecycle. Training both internally and from external hires must come from multidisciplinary areas of archival studies, library science, information science, computer science, and computational science.

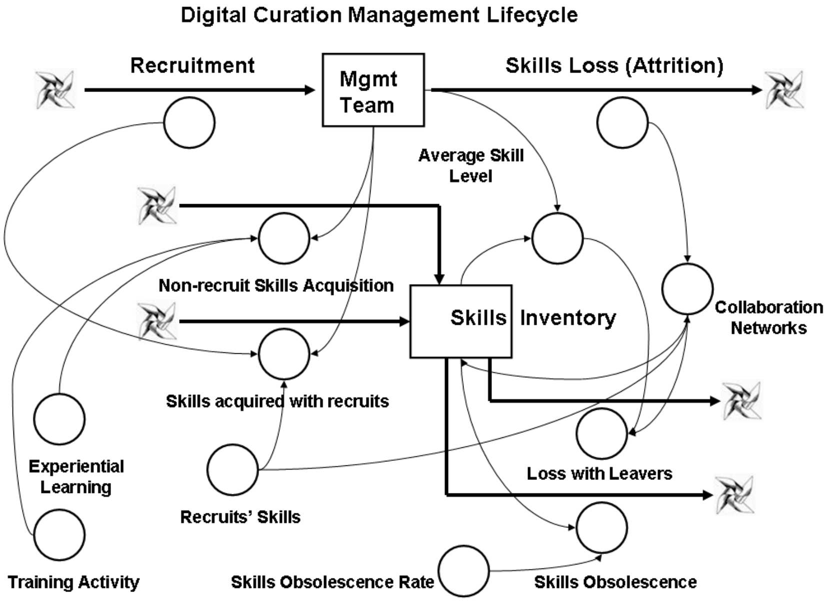

Figure 2 – Adapted from Winch (1998), Figure 2 illustrates how critically important it is that collaboration and networking skills be reinvested into local digital preservation management.

Conclusions

Sustainability of our many digital preservation initiatives must be tied to our known cultural heritage organizations for long-term preservation. That being said, it is clear that no one institution will be able to support every piece of the digital preservation agenda (Rusbridge, 2006). Having institutions which specialize in one area and then federate their services with others is one way to address this issue. Accurate models for ongoing costs of digital preservation are still at an early stage (Chapman, 2003)(Moore, 2007). Currently, it is unknown whether the institutions we will rely on to provide overall digital curation and preservation infrastructure are sustainable. Having diverse strategies that involve many different collaborations and partnerships (Lavoie and Dempsey, 2004) will ensure a diversity of practices offering preservation that at a minimum matches expectations from print curation models. Offering multiple copies of digital information curated in different ways, creating collaborative efforts with nonprofit oversight, like the Educopia Institute, or cultural heritage oversight and hiring and supporting personnel with the necessary digital curation skills, will help to provide the best chance for long-term access to digital information in the event of methodological or technical failures.

Acknowledgements

Mr. McDonald's involvement with this work was partially funded by the National Archives and Records Administration (NARA) and the National Science Foundation (NSF). Views and conclusions contained in this report are the authors' and should not be interpreted as representing the official opinion or policies, either expressed or implied, of the Government, or any person or agency connected with them.

Both authors wish to acknowledge NDIIPP funding from the Library of Congress, which has furthered the governance and sustainability work that has been accomplished by the MetaArchive Cooperative as well as by the MetaArchive partner institutions (Auburn University Libraries, Emory University Libraries, Florida State University Libraries, Georgia Institute of Technology Library and Information Center, University of Louisville Libraries, and the Virginia Polytechnic Institute and State University Libraries).

References

Apache Software Foundation (ASF) (2007). FAQ on the Apache Software Foundation.
http://www.apache.org/foundation/faq.html

Arms C., R.H. McDonald, L. Nicol and T. Walters (2005). Building a Collaborative Digital Preservation Network: NDIIP and the METAARCHIVE Experience. *Proceedings of the Educause 2005 Annual Conference*.
http://www.educause.edu/LibraryDetailPage/666?ID=EDU05199

Barton, M. and J. Walker (2003). Building a Business Plan for DSpace, MIT Libraries' Digital Institutional Repository. *Journal of Digital Information*, 4(2).
http://jodi.tamu.edu/Articles/v04/i02/Barton/

Beagrie, N., and D. Greenstein (1998). A Strategic Policy Framework for Creating and Preserving Digital Collections. Version 5.0 (Final Draft). ELib Supporting Study, p. 3. London: Library Information and Technology Centre, South Bank University.
http://www.ahds.ac.uk/about/publications/index.htm

Chapman, S (2003). Counting the Costs of Digital Preservation: Is Repository Storage Affordable?.*Journal of Digital Information*, 4(2), 1-15.
http://jodi.tamu.edu/Articles/v04/i02/Chapman/chapman-final.pdf

Educopia Institute (2007). Website of the Educopia Institute.
http://www.educopia.org

Educopia Institute (2007). MetaArchive Cooperative Charter.
http://www.metaarchive.org/pdfs/MetaArchiveCharter0906.pdf

Hedstrom, M. and S. Montgomery (1998). Digital Preservation Needs and Requirements in RLG Member Institutions. Mountain View CA: RLG.
http://www.rlg.org/preserv/digpres.html

International Internet Preservation Consortium (2007). Website of the IIPC.
http://www.netpreserve.org

Jannik, C. T. Graham, and M. Spasser (2005). Institutional Repositories. Proceedings of the Georgia COMO Conference.
http://hdl.handle.net/1853/11081

Kennedy, M. (2005). Reformatting Preservation Departments: The Effect of Digitization on Workload and Staff. College and Research Libraries, 66(6), 543-551.

Lavoie, B. and L. Dempsey (2004). Thirteen Ways of Looking at...Digital Preservation. *D-Lib Magazine*, 10(7/8).
http://www.dlib.org/dlib/july04/lavoie/07lavoie.html

LeFurgy, W. (2005). Building Preservation Partnerships: The Library of Congress National Digital Information Infrastructure and Preservation Program. Library trends, 54(1), 163-172.

LOCKSS Alliance (2007). LOCKSS Alliance Website.
http://www.lockss.org/lockss/LOCKSS_Alliance

Moore, R.L., J. D'Aoust, R.H. McDonald, and D. Minor (2007). Disk and Tape Storage Costs Models. Proceedings of the 2007 IS&T Archiving Conference.

Moore, R.W., F. Berman, B. Schottlaender, A. Rajasekar, D. Middleton, and J. JaJa. (2005). Chronopolis: Federated Digital Preservation Across Time and Space. Proc. of the IEEE-CS International Symp. on Global Data Interoperability, p. 171.

Merrill-Oldham, J., C. Morrow, and M. Roosa (1991). *Preservation Program Models: A Study Project and Report*. Washington: Association of Research Libraries.

National Archives and Records Administration. (2007). Electronic Records Archives Website.
http://www.archives.gov/era/

NSF and Library of Congress Joint Report (2004). It's About Time: Research Challenges in Digital Archiving and Long-Term Preservation: Final Report.
http://www.digitalpreservation.gov/about/NSF.pdf

Rusbridge, C. (2006). Excuse Me: Some Digital Preservation Fallacies. *Ariadne* 46 Jan. 2006.
http://www.ariadne.ac.uk/issue46/rusbridge/

State Library of Virginia (2006). Governor Mark R. Warner Administration Web Archive.
http://www.lva.lib.va.us/whatwehave/webarchive/warner_web_archive.htm

Thomas, C.F., R.H. McDonald, A.D. Smith, and T.O. Walters, "The New Frontier of Institutional Repositories: A Common Destination with Different Paths," *New Review of Information Networking* 11:1 (May 2005): 65-82.

Walker, J.H. (2006). Founding a DSpace 501(c)(3) NonProfit.
http://wiki.dspace.org/static_files/1/15/Non-profit_white_paper_-_Final.pdf

Walters, T. (2006). Digital Sustainability: Weaving a Tapestry of Interdependency to Advance Digital Programs. *Proceedings of the Sustaining Digital Libraries Conference*, Emory University, Atlanta, GA, October 6, 2006.
http://smartech.gatech.edu/handle/1853/12178

Walters, T. (2006). Strategies and Frameworks for Institutional Repositories and the New Support Infrastructure for Scholarly Communications. *D-Lib Magazine* 12(10).
http://www.dlib.org/dlib/october06/walters/10walters.html

Waters, D., and J. Garrett. (1996). Preserving Digital Information: Report of the Task Force on Archiving Digital Information. Washington, DC: Council on Preservation and Access. New York: Research Libraries Group, Inc.
http://www.clir.org/pubs/abstract/pub63.html

Winch, G. (1998). Dynamic Visioning for Dynamic Environments. *Journal of the Operational Research Society* 49 354-61.