Chapter 1: Preserving Our Collections, Preserving Our Missions

Martin Halbert (University of North Texas)

Katherine Skinner (Educopia Institute)

As the collections of cultural memory organizations become increasingly digital, preservation practices for these collections must likewise turn to digital techniques and technologies. Over the last decade, we have witnessed major losses of digital collections, both due to large-scale disasters (e.g., hurricanes Katrina and Rita, the 2003 power grid failure of the northeastern U.S. and southeastern Canada) and more isolated, local-level events (media failures, human errors, hacker activities, and smaller-scale floods and fires). From these losses, we are learning how vulnerable our digital collections are and how urgently we need sustainable digital preservation practices for our cultural stewardship community.

Paradoxically, there is simultaneously far greater potential *risk* and far greater potential *security* for digital collections as compared to physical and print collections. Risk, because such collections are as ephemeral as the electrons with which they are written, and can be catastrophically lost because of both technical and human curatorial failures much more easily and quickly than our physical and print-based holdings. Security, because digital collections, unlike physical artifacts, can be indefinitely reproduced and preserved with perfect integrity and fidelity. For all intents and purposes, anything less than perfect continuity of digital collections implies complete corruption and loss of data. Thus we as cultural stewards must create fail-safe methods for protecting and preserving those collections that we deem to be of sufficient cultural and historical importance.

The apparatuses, policies, and procedures for preserving digital information are still emerging and the digital preservation field is still in the early stages of its formation. Cultural memory organizations are experimenting with a variety of approaches to both the technical and organizational frameworks that will enable us to succeed in offering the perfect continuity of digital data that we seek. However, most cultural memory organizations are today

underprepared for the technical challenges incurred as they acquire, create, and preserve digital collections.

As a result, troubling trends are already developing within our community that may be counterproductive to our overall aims. For example, many cultural memory organizations are today seeking third parties to take on the responsibility for acquiring and managing their digital collections through contractual transfer and outsourcing of operational arrangements. The same institutions would never consider outsourcing management and custodianship of their print and artifact collections; the very notion is antithetical to cultural memory organizations, which exist specifically for the purpose of preserving and maintaining access to these collections. Yet institutions are today willingly giving up their curatorial responsibilities for their digital collections to third parties, precisely at the time that these digital collections are becoming their most important assets.

The central assertion of the MetaArchive Cooperative, a recently established and growing inter-institutional alliance, is that cultural memory organizations can and should take responsibility for managing their digital collections, and that such institutions can realize many advantages in collaborative long term preservation and access strategies. This assertion is based both on the shared convictions of our members and on the successful results that MetaArchive has achieved in recent years through coordinated activities as a cooperative association.

Authored by members of the MetaArchive Cooperative, *A Guide to Distributed Digital Preservation* is intended to be the first of a series of volumes describing successful collaborative strategies and articulating specific new models that may help cultural memory organizations to work together for their mutual benefit.

This volume is devoted to the broad topic of distributed digital preservation, a still-emerging field of practice within the cultural memory arena. Digital replication and distribution hold out the promise of indefinite preservation of materials without degradation, but establishing effective processes (both technical and organizational) to enable this form of digital preservation is daunting. Institutions need practical examples of how this task can be accomplished in manageable, low-cost ways.

We have come to believe that the use of the LOCKSS (Lots of Copies Keep Stuff Safe) software developed by the Stanford University Libraries-based LOCKSS team for collaborative digital

preservation purposes is one effective, practical, and affordable strategy that many collaborative digital preservation initiatives may wish to consider. Portions of this volume thus delve into the specific topic of deploying the LOCKSS software to create Private LOCKSS Networks (PLNs). These are membership-based geographically distributed networks that are dedicated to the long-term survival of digital archive. The MetaArchive Cooperative has successfully operated a shared digital preservation infrastructure based on this model for more than six years, and has advised other groups in the implementation of similar networks. With eight known international examples in operation today, PLNs are arguably the first well-established approach to distributed digital preservation within the cultural memory arena. This is not to say that PLNs are the only approach to distributed digital preservation. Indeed, we hope that this book will someday stand as one among many guides and that practitioners creating other promising frameworks will produce documentation about the technical and organizational approaches they use in order to foster additional communities of development in this important field.

The remainder of this introductory essay will serve to outline the early consensus on the emerging field of distributed digital preservation (DDP) and the rationale for networks based on the use of the LOCKSS software that has been modeled by the MetaArchive Cooperative, the Alabama Digital Preservation Network, and other PLNs over the last six years. Our hope in this guide is to begin to advance the current conversation on these topics among cultural memory organizations, topics that we believe are central to the continued vitality and success of such institutions.

AT-RISK CONTENT AND THE EMERGING DDP FIELD

Most cultural memory organizations do not yet have a digital preservation program, although most are aware of their need for one. According to the 2005 Northeast Document Conservation Center (NEDCC) Survey by Liz Bishoff and Tom Clareson, 66% of all cultural memory institutions report that no one is responsible for digital preservation activities, and 30% of all archives have been backed up one time or not at all.[1] These statistics should be staggering and alarming for leaders of cultural memory organizations that expend major funding on the creation and acquisition of digital collections. As the service programs in

cultural memory organizations become increasingly focused and dependent on long-term access to these digital collections, the gap that exists in preservation efforts for these collections becomes all the more critical to address. Yet, the previously mentioned gap in our collective understanding of how to respond to this challenge is a reality; many institutions simply do not know what to do.

What is at stake if we do not build this proposed new field of digital preservation? The National Digital Information Infrastructure Preservation Program (NDIIPP) has highlighted the scope of the collections now at risk:

> Technology has so altered our world that most of what we now create begins life in a digital format. The artifacts that tell the stories of our lives no longer reside in a trunk in the attic, but on personal computers or Web sites, in e-mails or on digital photo and film cards…. When we consider the ways in which the American story has been conveyed to the nation, we think of items such as the Declaration of Independence, Depression-era photographs, television transmission of the lunar landing and audio of Martin Luther King's "I Have a Dream" speech. Each of these are physically preserved and maintained according to the properties of the physical media on which they were created. Yet, how will we preserve [the following] essential pieces of our heritage?
>
> - Web sites as they existed in the days following Sept. 11, 2001, or Hurricane Katrina?
> - Web sites developed during the national elections?
> - Executive correspondence generated via e-mail?
> - Web sites dedicated to political, social and economic analyses?
> - Data generated via geographical information systems, rather than physical maps?

- Digitally recorded music or video recordings?
- Web sites that feature personal information such as videos or photographs?
- Social networking sites?

Should these be at a greater risk of loss, simply because they are not tangible?[2]

A great deal of content is in fact routinely lost by cultural memory organizations as they struggle with the enormous spectrum of issues required to preserve digital collections, including format migration, interoperability of systems, metadata to make the collections intelligible, and a host of other challenges. If this range of challenges was not enough, best practices for the most basic requirement of all are still poorly understood, namely how to ensure the long-term continuity of the bytes of data that fundamentally comprise digital collections.

Backups versus Digital Preservation

There are some that would dispute the above statements. Backing up the content on servers to tape and other static media is a long-standardized component of system administration. Why do we differentiate data backups from digital preservation programs? As the Joint Information Systems Committee (JISC) wrote in 2006:

> Disaster recovery strategies and backup systems are not sufficient to ensure survival and access to authentic digital resources over time. A backup is short-term data recovery solution following loss or corruption and is fundamentally different to an electronic preservation archive.[3]

Backups have always been tactical measures. Tape backups are typically stored in a single location (often nearby or collocated with the servers backed up) and are performed only periodically. As a strategy, backups are designed to address short-term data loss via minimal investment of money and staff time resources. While they are certainly better than nothing, backups are not a comprehensive solution to the problem of preserving information over time.

Digital preservation is strategic. A digital preservation program entails forming a geographically dispersed set of secure caches of critical information. A true digital preservation program will

require multi-institutional collaboration and at least some ongoing investment to realistically address the issues involved in preserving information over time. It also requires the creation and maintenance of preservation policies and procedures that guide the long-term curation of digital collections.

WHY DISTRIBUTED DIGITAL PRESERVATION?

In the context of this critical need for a new consensus on how to preserve digital collections, a growing number of cultural memory organizations (including those of the MetaArchive Cooperative) have now come to believe that the most effective digital preservation efforts in practice succeed through some strategy for distributing copies of content in secure, distributed locations over time. This conceptual strategy is a straightforward carry-over of the practices that in the chirographic (handwritten) world of antiquity enabled scholars to preserve content through millennia of scribal culture. But in the digital age this strategy requires not only the collaboration of like-minded individuals, but also an investment in a distributed array of servers capable of storing digital collections in a pre-coordinated methodology.

A single cultural memory organization is unlikely to have the capability to operate several geographically dispersed and securely maintained servers. Collaboration between institutions is essential, and this collaboration requires both organizational and technical investments. Not only a pre-coordinated technological solution, but also strong, long-term inter-institutional agreements must be put in place, or there will be insufficient commitment to act in concert over time. The following quote from a joint National Science Foundation (NSF)/Joint Information Systems Committee (JISC) study captures the opportunity presented by this situation succinctly:

> The increased number and diversity of those concerned with digital preservation—coupled with the current general scarcity of resources for preservation infrastructure—suggests that new collaborative relationships that cross institutional and sector boundaries could provide important and promising ways to deal with the data preservation challenge. These collaborations could potentially help spread the burden of preservation, create

> economies of scale needed to support it, and mitigate the risks of data loss.[4]

The experience of the MetaArchive Cooperative echoes this sentiment, namely that any effective implementation of distributed digital preservation requires both a robust technical infrastructure and strong inter-organizational arrangements. By "robust technical infrastructures" we especially mean strategies combining geographic distribution to multiple locations and security of individual caches, a combination of approaches that maximizes survivability of content in both individual and collective terms. Maximizing security measures implemented on individual caches reduces the likelihood that any single cache will be compromised. Distribution reduces the likelihood that the loss of any single cache will lead to a loss of the preserved content. This combination of strategies enabled documents to survive over millennia in the scribal world. We do not yet know if they will have similar results in the digital world, but they offer the most promising strategy to date.

A CALL TO ACTION

A central purpose for the MetaArchive Cooperative as we wrote this guide was to build documentation to help strengthen the distributed digital preservation and Private LOCKSS Network (PLN) communities and to encourage institutions to create and engage in collaborative preservation strategies with each other. Cultural memory organizations understand preservation issues in ways that other entities do not, and as a community, we must value both the training and the mindsets that librarians and curators bring to the virtual table as we pioneer solutions for preserving our digital collections. Philosophically and practically, our mission as museums, libraries, and archives is twofold: to provide access to and to preserve those objects deemed by curatorial experts to be most important for current and future generations. We need to provide preservation services more urgently than ever before due to the vulnerability of our digital assets. If cultural memory organizations do not take an active role in the preservation of our own collections, and rather cede this responsibility to external agents to do it for us, we run the risk of undermining our own stability as institutions.

To put that more plainly, to outsource one of our two key missions in the digital medium is to begin acting as brokers rather than

curators—a dangerous step in any time, but particularly so in one so fraught with economic tensions. If we want to continue serving as cultural stewards of the digital age, we must be active players, not passive clients of third-party preservation services. We learned a difficult lesson in the university library community in the 1990s when we chose to pay for temporary access to major journals that were digitized by corporate entities rather than creating and owning that digital content for ourselves; we cannot afford to repeat that mistake in the realm of preservation.

One of the greatest risks we run in not preserving our own digital assets for ourselves is that we simultaneously cease to preserve our own viability as institutions. One of the costs to us as institutions if we ignore, postpone, or outsource our duty as cultural stewards of a digital age is that other institutions will fill the gap that we leave, likely to the detriment of our institutional community and with it, to our cultural memory.

So how can we affordably approach digital preservation for ourselves? This guide provides a sketch of the current DDP landscape and the promising models that are emerging within it, and outlines for interested institutions a pathway forward.

COLLABORATIVE APPROACHES TO PRESERVATION

Organizationally, the DDP value proposition is both simple and radically different from that of most businesses. It advocates seeking to reduce both short- and long-term costs by investing in commonly owned solutions. DDP networks require deep and long-term commitments from their members in order to serve the preservation needs for which they are designed. We need new technical and organizational models for undertaking this collaborative work as we work to sustain our activities.

In the following chapters, readers will learn about different approaches to DDP. One of these approaches has been pioneered by groups of institutions using a common technical framework, that of the Private LOCKSS Network (PLN). However, the resources required to run these PLNs varies widely depending upon the specific decisions made by each PLN at both a technical and organizational level. Some PLNs are run within existing organizational structures and administered by the Stanford LOCKSS team. Other PLNs are independent entities. Some PLNs use the standard tools provided by the LOCKSS team; others

couple these tools with additional preservation tools and services to meet more specific preservation needs. PLNs, then, can be created with different levels of technical and administrative complexity. Each approach has its pluses and minuses, but all have been shown to accomplish the central aim of distributed digital preservation. It is up to the institutions that host and participate in a PLN to decide which approach will work best for them.

We hope that this guide will help to disseminate information about some of these emerging models and in doing so, assist other cultural memory groups in their efforts to create distributed digital preservation networks.

ENDNOTES

1. See http://www.ils.unc.edu/digccurr2007/slides/bishoff_slides_8-3.pdf (last accessed 01-29-2010).
2. See http://www.digitalpreservation.gov/importance/ (last accessed 07-21-2008).
3. JISC. “Digital Preservation: Continued access to authentic digital assets." (November, 2006)
4. Berman and Schottlaender, “The Need for Formalized Trust in Digital Repository Collaborative Infrastructure, NSF/JISC Repositories Workshop (April 16, 2007) http://www.sis.pitt.edu/~repwkshop/papers/berman_schottlaender.html (last accessed 01-20-2010).