The MetaArchive Cooperative: A Collaborative Approach to Distributed Digital Preservation

Katherine Skinner and Martin Halbert

Abstract

As identified by the U.S. Congress, there is a national need for collaborative approaches to digital preservation services for cultural, historical, and political repositories. Responding to this need, the Library of Congress established the National Digital Information Infrastructure and Preservation Program (NDIIPP) in 2003 to produce a "national network of partners collaborating on digital preservation" that would pioneer a variety of digital preservation services for cultural memory organizations. Among these collaborative ventures, the MetaArchive Initiative developed an organizational model and operated a technical infrastructure (building on the LOCKSS software developed at Stanford University) for preserving the digital assets of cultural memory organizations in a geographically distributed framework. Subsequently, MetaArchive transitioned from a project to a program with the founding of the MetaArchive Cooperative in 2007. This article focuses upon the relationship between MetaArchive and NDIIPP, highlighting MetaArchive's commitment to enable institutions to host their own preservation solutions rather than outsourcing this core mission. It details the strategies that the MetaArchive Cooperative has employed to support, sustain, and grow its cross-institutional collaboration; explores an array of logistical and organizational issues that have arisen; and discusses the strengths of particular organizational structures for fostering and sustaining collaborative work between peer institutions.

Introduction

In the print world, librarians, archivists, and curators have long built and used information management architectures to safeguard our cultural

legacy for future generations. In the early twenty-first century, we are only beginning to institutionalize such information management architectures to effectively sustain *digital* information over *decades*, let alone centuries or millennia (Berkman, 2008). As a nation, we are in danger of losing an enormous portion of the cultural, political, and historical data that will allow current and future citizens and scholars to understand these decades of change from a print-based to digitally oriented society.

Studies of current practices in government archives provide a snapshot of the gravity of the problem and point us toward potential solutions. The Center for Technology in Government's *State Government Digital Information Preservation Survey and Report* found that the current capacity for digital preservation of state governmental information is very low, approaches are inconsistent, and that there is no standard way of prioritizing at-risk materials for preservation. The report recommended that this community "build digital preservation partnerships within and among state library, archives, records management, IT, and other interested and responsible parties," and noted that "partnerships have emerged as the most viable strategy . . . for preserving state government digital information" (Pardo, Burke, & Kwon, 2006).

Similar studies have been conducted with libraries, archives, museums, and historical societies. The Northeast Document Conservation Center's (NEDCC) 2005 online survey discovered that more than 88 percent of the museums, libraries, archives, and historical societies who participated in their survey were "collecting, acquiring, or creating digital assets," but that only 29 percent of these respondents had documented "policies addressing the preservation and management of these digital assets." Also disturbing, the majority of respondent institutions devoted 5 percent or less of their budget to *any* type of preservation activity, and 9 percent devoted none at all. A subsequent colloquium NEDCC hosted for digital preservation experts recommended that "small and medium-sized institutions will need the assistance of experts to assess the preservation status and needs of their expanding digital collections" (Clareson, 2006).

The Task Force on the Future of Preservation in the Association of Research Libraries (ARL) developed a set of recommendations in April 2007. Among them, one was titled: "Support the Library Community's Ability to Provide Stewardship for Their Collections." The report commented that "'Community' is the key word in this recommendation," then continued, "Although all preservation requires concerted local investment, it is evident that significant work can be accomplished only through coordinated activity, e.g., development of shared collections, support for centers of expertise and shared preservation services that offer economies of scale, and opportunities for preservation staff to develop new skills" (Teper, 2007).

The common themes emerging in these and other recent studies of cultural, political, and historical archives' digital preservation readiness and practice are clear:

- Most cultural memory organizations have not yet implemented sufficient preservation operations for their digital collections, and
- Most experts recommend that *only* through cooperative efforts can we effectively address the enormous challenges of digital preservation.

Cooperative work in *digital* preservation for cultural, historical, and political archives provides the opportunity for chronically underfunded but richly knowledgeable archival communities to jointly create and maintain an infrastructure that is low cost, open source, replicable, and designed with information management strategies at its core.

The MetaArchive Approach to Digital Preservation

Anticipating and responding to the need for cooperative practices in the emergent field of digital preservation, the initial NDIIPP-supported MetaArchive project sought to collaboratively establish a secure digital "archive of archives," or MetaArchive. As one of the original eight initiatives contracted by the Library of Congress, MetaArchive (http://MetaArchive.org) began in 2004 as a venture led by Emory University with Georgia Tech, University of Louisville, Virginia Tech, Auburn University, Florida State University, and the Library of Congress. Following the success of our initial endeavor, we transitioned in 2007 to an independent, unincorporated, international membership association, the MetaArchive Cooperative, for the purpose of supporting, promoting, and extending our collaborative approach to distributed digital preservation practices. We are continuing to work with the Library of Congress on a second project, and also enjoy the support of the National Historical Publications and Records Commission (NHPRC), along with membership fees paid into the Cooperative by all member institutions.

During the initial project with the Library of Congress (2004–7), the MetaArchive partner institutions envisioned designing and implementing a technical and organizational infrastructure for the preservation of at-risk digital cultural content. The concept of digital preservation had, by this time, become an issue of critical concern to the principal investigators of this project. Each felt strongly that preservation is a core mission of libraries and other cultural memory organizations, and that as such, it is not a task well-suited for outsourcing. We actively sought to create a solution that would enable cultural memory organizations to take responsibility for their own digital preservation needs—including the myriad levels of information management required to ensure the safety, longevity, and accessibility of collections—rather than relying on vendors and service providers to do so for them externally.

To provide an initial focus and scope for the project, the project participants proposed to prioritize digital content in a subject area of national value: the cultures and histories of the U.S. South. By focusing on a subject domain of shared importance to our participant institutions, we intended to both provide our collection with clear parameters and to test the efficacy of forming archives around specific subject areas or other genre types.[1]

The project group members set out in 2004 to jointly develop the following:

- A prioritized conspectus of at-risk digital content in this subject domain held at the partner sites
- A distributed preservation network infrastructure based on the LOCKSS software (http://www.lockss.org)
- A preserved body of critical content from our partner sites
- A formal cooperative agreement to facilitate our ongoing collaboration

Below, we provide a brief overview of the initial NDIIPP MetaArchive project's successes and a few of the lessons we learned along the way in each of these areas. For more details, please see the project documentation and final report at http://metaarchive.org.

Creating a Conspectus of At-risk Content

Before tackling the actual preservation of at-risk digital Southern culture resources, we sought to clearly define the scope of the collection we intended to amass, including its subject domain, what formats it would contain, cataloging requirements, copyright considerations, and harvesting prioritization. Scoping the collection required us to think, not only about this first subject-based archive that we were establishing, but also how we wished to develop additional archives in the future. We created a scope document to record our initial project decisions—including our subject domain definition, what standard metadata we would capture, what formats we welcomed, how we intended to prioritize collections for ingest, and our early recommendations to partners concerning copyright status and preservation.[2] We also created a schema to enable us to record collection-level preservation metadata for all collections prior to ingest.[3]

We produced a conspectus tool to enable our partners to use this thirty-eight-field schema to create collection-level entries for each collection they intended to submit to the MetaArchive for ingest. The resulting Web form allows participants to record collection-level preservation metadata. The collection-level metadata records information regarding the collections' contents, structure, form, and ingest route—details that are imperative both for securely maintaining and monitoring the collection

within the archive *and* for using the collection to restore in-house masters if technical disaster should strike a participating institution.

Building on this work as we moved to a Cooperative structure in 2007, we began documenting these practices for an increased membership and for a variety of archive types. As our Cooperative begins to provide "full preservation" (meaning, format migrations for standard file types) as well as bit-level preservation (which we have always provided), this documentation also specifies the file types that we will support with each level of preservation.

Among the most important findings of our early project work are the following points:

- We need collection-level metadata as a tool and system component; we do not need to require institutions to provide item-level metadata. Item-level metadata provides detailed descriptive, technical, and administrative information about each item in a collection. It is expensive to produce, and as such, many digital collections are not cataloged at the item level. The result is that metadata becomes a barrier to participation in many of the emerging preservation solutions that require PREMIS and other metadata forms for all submitted objects. The MetaArchive Cooperative determined that collection-level metadata would enable (1) the processing and monitoring of collections, and (2) the information needed to restore a collection.[4] As a result, MetaArchive *requires* collection-level metadata for each ingested collection to be entered in a common schema using a Web tool that we provide as part of our system. We also *recommend* that institutions that have created item-level metadata submit that metadata as part of the collection to be ingested. In this way, we are able to preserve all existing aspects of a collection— including its item-level metadata, where that exists—without producing an artificial barrier to entry.
- In order to be effective, preservation requires internal institutional documentation. The Cooperative highly recommends that its participants produce in-house documentation to record their decisions and practices. These should include: (1) procedures for suggesting collections for preservation; (2) curatorial decisions (what collections to preserve, rights issues, risk rankings); (3) timelines for adding new content/re-evaluating collections for inclusion; (4) responsible parties and processes for monitoring their preserved materials (i.e., who checks to ensure that their preserved collections are ingested correctly and preserved within the system, and how/when do they do so?); and (5) institutional policies for reconstructing collections if technical disaster does strike their in-house copies. Without such documentation, no institution can take full advantage of the preservation solution(s) that they employ.

PRODUCING A PRIVATE LOCKSS NETWORK

From the earliest stages of project planning, the partners agreed that it is questionable whether any *centralized* scheme for preservation can succeed over the long haul. All of us instead believed that *decentralized* mechanisms for mobilizing group efforts between cooperating institutions hold great promise for effective long-term models for preservation. This conviction led the group to utilize the open source Lots of Copies Keep Stuff Safe (LOCKSS) software, developed by Stanford University, as the foundation for MetaArchive's distributed technical system.

At the highest level, a LOCKSS network is comprised of *content provider sites* whose content is being crawled and ingested by *preservation nodes* (LOCKSS caches) that are administered by institutions that contract with each other to preserve this content in a collaborative, distributed manner. The LOCKSS software provides a systematic way for these LOCKSS caches to constantly check in with one another for signs of file degradation or "bit rot," as well as for a node's disappearance due to technical or physical catastrophe. If a problem with a file is detected, the LOCKSS caches conduct a "poll" to determine which copy has been compromised. Once they reach quorum, the software can repair that damaged copy. Similarly, if an entire LOCKSS cache disappears from the network due to technical failure, the node can be safely re-created via the other caches in the network. Each of these LOCKSS caches is under separate administrative control, which guards both against natural/technical disaster and the threat of economic failure of any one node. Any node in the system can drop out—whether intentionally or due to technical or organizational failures—without jeopardizing either the network or its contents. Myriad testing scenarios, both by the LOCKSS team and the MetaArchive project group, have consistently demonstrated the stability of this preservation system's technical infrastructure.

As of 2004, when the MetaArchive project began, LOCKSS had been used exclusively as a public network tool, and its primary application had been in the e-journal market. In that instance, LOCKSS had been adopted by institutions that held a shared sense of investment in the materials that they preserved, which were e-journals that each subscribed to and that all wanted to ensure would be available in the long term. Using LOCKSS, these institutions preserved e-journals in a geographically distributed network with the understanding that if the access copies their scholars relied on were to disappear due to technical or organizational failure, their LOCKSS network would provide a copy for public access.

MetaArchive's use of LOCKSS differed substantially from this public network model. We sought to preserve materials that were of value to each institution, but not necessarily to the group of institutions as a whole. We also sought to expand the number of formats, sizes, and file types that LOCKSS preserved by ingesting our institutions' diverse arrays of archival

materials. Finally, we formed our network as a *Private* LOCKSS Network, or a closed system within which materials are stored with no intention of providing public access to them in either the short or long term. The preserved materials are made available only to the institution that submitted them for ingest, not to other parties. In this way, MetaArchive has intentionally separated the function of preservation from that of access. We encourage institutions to provide access to their collections through other means, but we do not seek to fulfill that function for them in our "dark" archival system.

We learned several important lessons as we implemented our technical framework, including the following:

- In order to sustain a cooperatively run technical infrastructure, you must distribute both the technical components (LOCKSS caches) *and knowledge* of how to manage the system. One of the core strengths of our approach to preservation is the distributed nature of our system. Rather than creating an organizational infrastructure in which all major systems knowledge is held by a centralized project staff, the Cooperative members have sought to spread the knowledge out among each institution's technical staff members as well, ensuring that (1) sustainability is increased through this distributed knowledge—we are not dependent on one or two central staff members to run the system, but rather have shared expertise to draw on across all member institutions; and (2) the Cooperative does not need to incur the costs associated with employing and hosting central staff—which allows the Cooperative to keep its costs low and provides a major savings for our member institutions.
- Our technical infrastructure, though lightweight, requires systems attention at each Preservation site. Because we geographically distribute the LOCKSS caches, the institutions that host these caches bear responsibility for setting up and maintaining their servers as part of the overall network. Also, because we hold a strong belief in distributing knowledge about how to manage the system, the system administrators and technical staff members who help to set up and maintain these servers may also become part of the "knowledge bearers" who learn how to manage the overall system.

Preserving Critical Content

From the time our technical infrastructure became operational we have been actively ingesting and preserving a body of content that includes collections from each of our participating institutions in the Southern Digital Culture subject domain.

Among the collections ingested and preserved in our network during the first project phase with the Library of Congress are digital masters of preservation scans of historic photographs (e.g., *Alabama Cooperative*

Extension Service ACES Photographs, 1920s to 1960s; *Photographs of the Historic American Buildings Survey of Georgia*; and *Jean Thomas, The Traipsin' Woman*; as well as a variety of university archives holdings at each institution); text files (e.g., electronic theses and dissertation files, university archives documents); sound and video files (e.g., Virginia Tech Special Collections and University Archives Sound Recordings and the oral history interviews of the Bernheim Foundation, a historic environmental trust in Kentucky); "born-digital" materials (e.g., the multimedia Internet journal *Southern Spaces'* html, image, sound, and video files and Virginia Tech's *We Remember* site memorializing the events of April 16).

In order to ingest a collection, the contributing institution undertakes a four-part process: (1) identifying the collection and ensuring it is accessible via the Internet for ingest (usually through a secure pathway); (2) completing a collection-level entry in the Conspectus Database for the collection; (3) writing a "plugin," or set of code that sets the parameters for ingest; and (4) alerting the LOCKSS caches that there is a new collection ready to ingest.

In the extension phase of our project work (2007–9), we are continuing to ingest and preserve a wide range of content. We are actively populating two additional archives, one for electronic theses and dissertations in partnership with the Networked Digital Library for Theses and Dissertations (NDLTD) and another for an international collection on the history of the slave trade. We establish new archives for specific subject domains and genre types at the request of our member institutions, which means that as MetaArchive continues to grow, so will the diversity of collections stored in its archives.

Among those things we have learned about content management and preservation are the following:

- Just because an institution has content does not mean that its content is ready for ingest. We anticipated that all collections would be either Web-accessible (necessary, as LOCKSS ingests collections using the Internet) or easily mountable for ingest. A major project finding is that many of the collections most at risk were those that were stored off-line on CDs or DVDs. These collections are often poorly monitored and poorly labeled. Each institution struggled with the unaccounted-for expense of locating, checking, and providing metadata for these collections in order to prepare them for ingest. It quickly became apparent that these "data wrangling" activities were paramount to the success of our project and, indeed, to digital preservation activities at large.
- The organization of an institution's collections can help or hinder its preservation readiness. In order to successfully ingest and preserve a collection, that collection must be "fixed," or stable and unchanging. If a collection is growing (or shrinking), the ingest conducted in a particular

month and year cannot account for the growth and/or shrinkages that take place thereafter—thus compromising the preservation process for that collection. For example, if an institution stores its electronic theses and dissertations in an access-oriented system in folders marked by year, it is quite likely that the 2008 folder may not yet include restricted dissertations (those held back from access due to author request or other factor). When the restriction period ends, the institution may add particular ETDs into the 2008 folder. If the MetaArchive nodes are not notified of this change, there is no way to ensure that the change will be accurately reflected in the preservation archive.[5] Such issues must be addressed, either by internal institutional policies or by organizational structures, in order to ensure that the content an institution intends to preserve is indeed the content that it is preserving.

CREATING A COOPERATIVE AGREEMENT

At the beginning of our 2004–7 NDIIPP project, the project participants intended to create a lightweight "Cooperative Agreement" document that would govern our ongoing collaboration beyond the contract-funded period, and that would also provide a model and template for future networks.

We quickly realized that a simple Cooperative Agreement would not provide the level of formality needed by a preservation-oriented Cooperative. Long-term preservation demands a higher level of commitment and dependability from member institutions. Under the guidance of legal counsel, the Steering Committee determined that a Charter and legal Membership Agreement best fitted our Cooperative's arrangement.

The resulting MetaArchive Cooperative Charter accomplishes two interrelated goals: (1) it defines the Cooperative's mission and operating principles, membership responsibilities, governance structure, and services and operations, and (2) it formalizes the relationships between member institutions with a legal contract. The Charter was first released in March 2007 and is updated annually at the MetaArchive Steering Committee meeting (*Resources*, n.d.).

As we created the Charter and Membership Agreement, the question arose: with whom are we making these agreements? As we will elaborate below, the idea of carrying forth with one peer institution from the Cooperative serving as the "lead" belied the very purpose of creating a balanced Cooperative structure. Instead, we sought to create a sustainable infrastructure with its own mission. In order to fulfill this goal, we determined that we needed a legal entity beyond the member institutions that could act as the Cooperative's fiscal and administrative agent. This would effectively decouple the MetaArchive Cooperative from the member institutions, removing its dependence on its initial participants and in doing so, promoting its sustainability. The group thus decided to form a nonprofit

entity to oversee the operations of the MetaArchive Cooperative. This organization, the MetaArchive Services Group, was incorporated in Atlanta in 2006, and was granted 501(c)(3) status in 2007.

The success of our project-turned-program can be credited to several important factors. First, and perhaps most important, we have been a highly engaged partnership comprised of principal investigators and project staff who take an active role in the project's work via regular meetings in Atlanta, weekly conference calls, and a listserv. We have also experienced the strong support of the LOCKSS team. As several collaborating institutions have described, we have a "roll up your sleeves" mentality across all formal and informal participants—somewhat of a rarity in collaborative projects.

We also had an outstanding range of expertise across our institutional partnerships. This included technical experience (system administration and programming), library experience (collections scoping and metadata), and organizational experience (documentation and legal factors). Each institution brought a different skill set to the project that, when combined, created a stellar project team.

From an early stage of project development, we began actively sharing the results of our research and production with other digital library efforts (including other NDIIPP-supported groups) and attempting to integrate some of their findings into our model where appropriate. This, too, enriched the work that we completed during and beyond the project period.

But one of the most important and foundational elements in our work has been our ongoing relationship with the Library of Congress and other NDIIPP-supported groups. In the next section, we will discuss the significance of being a part of this particular cluster of projects.

NDIIPP, MetaArchive, and the Emerging Digital Preservation Landscape

It is important to understand the context of the MetaArchive Cooperative as relates to the NDIIPP and the emerging landscape of digital preservation in the early twenty-first century. This is the dawn of the digital preservation field, and we now stand on intellectual ground that will serve as our foundation for many, many years to come. For the benefit of current and future debates in this emerging field, it may be useful to briefly discuss what differentiates the MetaArchive Cooperative from other NDIIPP projects and digital preservation initiatives.

The NDIIPP is the first systematic digital preservation program chartered by the government of the United States:

> In December 2000, the U.S. Congress passed legislation establishing the National Digital Information Infrastructure and Preservation Program (NDIIPP) in the Library of Congress (PL 106-554). The legislation charges the Library to lead a national planning effort for the long-term

preservation of digital content and to capture current digital content that is at risk of disappearing. The legislation calls for the Library to work collaboratively with representatives of other federal, research, non-profit, philanthropic, library, and business organizations . . . The over-arching programmatic focus of NDIIPP is on ensuring preservation of historically significant digital content through the establishment of a national network of committed partners, collaborating in a digital preservation architecture with defined roles and responsibilities. The ultimate goal of this national preservation program is to build a nationwide network of partners to share responsibility for digital content and to seek national solutions (Library of Congress, 2003).

From the inception of the program, the strong emphasis made clear in the program announcement and in subsequent presentations by the repetition of the phrasing, was on creating a *national network of partners collaborating on digital preservation.* NDIIPP leaders understood from the beginning that no single strategy or entity would have the capacity or could successfully assume the responsibility for digital preservation of the nation's cultural materials. This focus on cultivating an aligned network of cultural content stewards can be understood as a *decentered* approach that emphasizes decentralized cooperation over centralized hierarchical arrangements. The *degree* of decentralization that would have the best results in digital preservation efforts was unknown at the time the NDIIPP was inaugurated. Within the program, many experimental arrangements were funded to foster intermediate organizations and relationships, ranging from centralized corporations to broadly distributed structures. However, there was a clear conceptual move in the creation of the NDIIPP toward decentralized structures.

The leaders of the MetaArchive Cooperative also strongly believed in the conceptual strengths of decentralized organizational networks, and had a specific interest in exploring the idea of alliances of cultural memory institutions (libraries, archives, historical research associations, and museums) to address the pressing needs and scale of problems in digital preservation. This preference is not shared by all organizations in the emergent digital preservation field. Some consider it a controversial stance and place more emphasis on vesting digital preservation responsibilities in centralized institutions *external* to cultural memory institutions. In this alternate scenario, cultural memory organizations may arrange to contractually outsource their digital preservation activities to external parties.

The conceptual difference in these strategies manifests itself in the approaches their advocates have proposed to address the digital preservation challenge. The difference of emphasis constitutes a creative tension and unresolved issue at this early stage of the digital preservation field and has great implications for what form digital preservation practice will take in the future and for how we understand the functions of cultural memory organizations in the digital age.

The most visible distinction between the decentralized and centralized approaches is that the decentered approach would task cultural memory organizations with digital preservations activities (albeit taking advantage of inter-institutional alliances to leverage efforts), while the centralized approach recommends that cultural memory organizations outsource digital preservation activities to specialized external agencies (typically corporations). The NDIIPP has been careful to fund and advance both approaches in the interests of exploring all fronts simultaneously. The ramifications of pursuing one or the other of these two approaches are not immediately obvious to decision makers at cultural memory organizations, who may believe they need only to make a tactical decision based on immediate costs or efforts required to pursue either course of action.

But these two approaches surely have significantly different implications for the future of libraries, archives, and museums. Historically, cultural memory organizations have by definition considered the function of preserving cultural materials one of their core missions and reasons for existing at the individual institutional level. If an entity outsources its core mission, it thereby transfers responsibility for that mission to another entity and calls into question the need for its own existence. Centralized approaches that outsource preservation functions for digital cultural materials are problematic precisely because of this issue. *Cultural memory organizations that see digital preservation as a secondary or merely technical responsibility are missing the point that our cultural memory is rapidly becoming digital, and that the core of their future activities may well focus on these digital knowledge resources.* Outsourcing this core mission of preservation may eventually amount to a systematic restructuring of the sphere of cultural memory institutions, centralizing this function in a relatively small handful of specialized corporations.

Would this be a bad thing? The argument has been made that consolidating cultural memory functions in this way can result in greater economies of scale and therefore greater efficiency. It would ultimately shift the balance of responsibility for preservation activities toward corporations and away from institutions like universities and historical research associations in a manner reminiscent of the transition in the scholarly communication cycle that occurred in the late twentieth century, when the digital publication and/or distribution of scholarly journals was increasingly transferred to corporations like Reed Elsevier. The problem with this kind of restructuring is that while it arguably may improve efficiencies of scale, it changes the equation of control of cultural memory in ways that are *not* ultimately advantageous for cultural memory organizations. The question "What kind of institution should control our cultural memory?" is one of social choice, and not simply one of tactical expediency or immediate costs. Oligarchic solutions to governance are often put forward as more efficient, but they often prove to be undesirable in the long run because

of how they change the basic social decision-making process. It may seem unwarranted to extrapolate these larger conclusions from what may seem like simple technical decisions, but the lack of that extrapolation was precisely what led to what is now considered a widespread crisis in scholarly communication in the form of monopolistic journal subscription price escalation in the hands of corporations.

Bringing this discussion back to the NDIIPP and the emerging landscape of digital preservation in the early twenty-first century, our point here is that there is a long-term strategic importance in the MetaArchive Cooperative and similar approaches that emphasize that cultural memory institutions must assume responsibility for preserving their intellectual assets (whether these occur in analog or digital forms). Such decentralized approaches have the best chance of accomplishing their aims in an affordable and sustainable manner through collaborative efforts.

By the midpoint of the first round of NDIIPP projects in 2006, the leaders of the NDIIPP had realized that their first investment priority needed to be "building a distributed storage platform to help preserving institutions attain redundant and geographically disbursed storage of digital materials at low cost" (Smith, 2006), which had also been the first priority of the MetaArchive Cooperative during its startup phase. By 2006, MetaArchive had been operating such a distributed storage platform for several years. At that point, we expanded our mission by inviting additional cultural memory organizations to join our network, and by advising other consortia in Alabama, Arizona, and elsewhere that sought to follow our Private LOCKSS Network model for establishing such distributed preservation approaches. The previously barren landscape of digital preservation was slowly starting to germinate. The hope of the MetaArchive Cooperative was to act as a kind of Johnny Appleseed who, contrary to the popular image of this frontiersman, did not scatter seeds randomly but helped farmers to systematically grow nurseries that could supply others with apple trees and associated products as a way of benefitting frontier society in terms of sustainability. By fostering a particular kind of self-sufficient approach to digital preservation, the MetaArchive Cooperative continues to work to catalyze change in the emerging digital preservation landscape.

The MetaArchive Cooperative has benefitted significantly from its formative strategic partnership with NDIIPP, becoming the first organization dedicated to providing distributed digital preservation services for archival and other cultural memory materials. NDIIPP has also benefitted significantly from the work of the MetaArchive Cooperative in that the group has become a catalyst for advancing the NDIIPP vision of a robust ecology of layers in the emerging content stewardship network, including custodians, communities of practice, digital preservation services, and capacity building (Anderson, 2007). In the next section we will briefly discuss at a high level how the MetaArchive Cooperative plans to expand and sustain

its operations, and in the process, to advance all of the elements of a robust ecology of digital preservation as envisioned by NDIIPP.

Growing and Sustaining the MetaArchive Cooperative and Distributed Digital Preservation Networks

In this article we have put forward the MetaArchive Cooperative as an exemplar of a new approach to digital preservation, a decentralized and cooperative approach that continues to vest preservation functions in cultural memory alliances rather than corporations external to the cultural memory sphere. How will such alliances sustain themselves over time? The leaders of the MetaArchive Cooperative have devoted a great deal of thought to this question for obvious reasons, and we here offer our initial prospective thoughts.

There are many practices and functions that once seemed problematic for libraries and other cultural memory organizations to undertake. For example, not too long ago, computer systems and services associated with them were arcane and impossibly expensive for libraries to consider incorporating into their program offerings. Times changed and as computer systems became more broadly distributed throughout all aspects of society, we recognized that computer-based services of many different sorts were not only possible in libraries but that they forwarded essential parts of the library's mission, including promoting access to collections. Even the most reactionary understanding of a library today would include the range of computer systems it uses to provide essential public services and staff functions. This change was certainly influenced by the much larger shifts in availability and adoption of computer systems throughout society, as mentioned. But the integration of computer systems into library practice required experimentation and leadership on the part of many individuals and groups of librarians. Experiments conducted during the last quarter of the twentieth century were critical to the adaptation of libraries to the changed landscape of the new millennium.

There is a similar need for experimentation and change today in the emerging field of digital preservation. The MetaArchive Cooperative and similar endeavors will only succeed to the degree that they can successfully cultivate such experimentation and leadership in cultural memory organizations during the coming years. On a tactical level, our cooperative must continue to grow for some period of time through the addition of new members to continue to be viable in terms of our core mission of distributed digital preservation. On the strategic level, if our vision of the future of digital preservation as a key feature of the twenty-first-century library, archive, and museum is to flourish we must continue to convince additional new groups to adopt decentralized and cooperative strategies in which cultural memory alliances retain responsibility for preservation functions.

The MetaArchive Cooperative is pursuing this goal of growth. On the tactical level, we are now in the process of admitting new members and continuing to expand the Cooperative with new preservation node sites and new subject/genre-based archives. These activities have proven to be relatively straightforward, requiring little or no marketing of the Cooperative as word of mouth has led to many new institutions wishing to join. The expansion of the MetaArchive Cooperative as a service provider of distributed digital preservation functions addresses one layer of the stewardship network envisioned by NDIIPP, and an important one as there are still very few options for institutions seeking to establish a digital preservation program. The advantage of the MetaArchive Cooperative is that it enables institutions to develop these services for themselves, rather than simply buying services from a vendor. In other words, rather than selling fish, we teach groups how to fish for themselves. Expanding the membership of the cooperative further distributes not only the content preserved in the network but also the minimal costs of coordinating the activities of network members. Further, it expands the number of individual institutions that are actively undertaking efforts as custodians and stewards of digital content (as opposed to simply ignoring the problem or implementing a black-box solution over which they have limited control at best), another layer of the NDIIPP ecology.

On the strategic level, the MetaArchive Cooperative has begun offering workshops and consulting services to other consortia interested in setting up distributed digital preservation networks similar to ours. This has resulted in a series of new "affiliates," new Private LOCKSS Networks devoted to preserving some body of cultural memory content. These networks are typically separate entities from the MetaArchive Cooperative, but implement some version of our organizational model and technical infrastructure. This advances our broader aim of encouraging cultural memory organizations to take on distributed digital preservation responsibilities through alliances and, concomitantly, the long-term sustainability of such preservation networks, which is an even higher priority for us than the growth of our own cooperative. If we can foster these kinds of communities of practice, we feel we will have succeeded in advancing the cause of content stewardship via digital preservation networks more broadly.

We are also actively seeking to better understand the possibilities and affordances of the MetaArchive Cooperative as a forum for our members to use as they focus on additional elements of the digital preservation challenge beyond bit-level preservation of collections, such as format migration, automated metadata enhancement, and pooled resource access portals. We also want to improve and mature the core offerings and capacity of the MetaArchive Cooperative as a digital preservation service, an aim for which we have recently been awarded funds to pursue by the National Historical Publications and Records Commission (NHPRC).

Groups of institutions within the Cooperative will be actively seeking out grants and contracts from additional federal and private funding agencies interested in advancing the state of the art in all of these areas in coming months as well.

On both the tactical and strategic levels, we are building the capacity for digital preservation within a large number of cultural memory organizations. Our hope is to serve as an effective element of the much larger landscape of digital preservation actors in a synergistic way, as advocates for cultural memory organizations becoming aware of and taking responsibility for this emerging function that holds critical importance for the future of such institutions.

In the next section we will discuss the details of how we are cultivating the MetaArchive Cooperative and pursuing these larger aims of mobilizing efforts between cultural memory organizations.

Sustaining MetaArchive: The Cooperative Approach and the MetaArchive Services Group

In order to foster and sustain the MetaArchive Cooperative's collaborative work and approach to distributed digital preservation, we have, of necessity, paid close attention to our organizational structure. We have undertaken a three-pronged approach to the intertwined endeavors of (1) growing the services we offer and (2) encouraging libraries and other cultural memory organizations to play an active role in the preservation of their own digital collections rather than outsourcing that work to external parties. First, we have established an organizational structure that we believe holds great promise for sustaining not just this program, but also for stimulating similar programs in the digital library field. Second, we have explicitly diversified our mission: as discussed above, we are not only interested in growing our own network, but are equally committed to helping other groups use our technical and organizational model to form and grow their *own* networks. And finally, we are actively working to form alliances with other collaboratives, consortia, and cooperatives as part of our belief that we can accomplish far more in partnership with each other than any single institution or any single group can in isolation.

From Collaborative to Cooperative

There are many different configurations that collaborative networks can adopt and adapt to fit their needs. Depending on the collaborative network's goals, mission, and stage of growth, the appropriate organizational structure or business model takes on different characteristics and may vary from very informal to very formal.

The original MetaArchive project (2004–7) started with a fairly informal structure that is familiar to many cultural memory organizations. The six participating institutions entered into this work together through creat-

ing a proposal and work plan to submit to a funding entity. Once awarded the sponsored funding, the six institutions began to work through the tasks laid out in that original proposal. As is common in sponsored funding projects, there was a lead institution, Emory University, with whom subcontracts were signed by each of the partner institutions.

Anticipating that the very substance of the project—long-term preservation—would necessitate a longer-term arrangement than a three-year project could provide, the participants had proposed to create documentation that would govern their ongoing collaboration as one of the four main deliverables of the project. In 2006–7, when the collaborating institutions sought to complete this documentation, we considered the various tasks that our organization would need to manage. Using a government-oriented guide developed by H. Brinton Milward and Keith G. Provan, we broke these tasks down into five main categories: accountability, legitimacy, conflict, commitment, and design (2006).

In the conversations that followed, we focused on the following broad points:

- On accountability: we needed a structure that would clearly designate someone *or* some group of people who were authorized by the network to assign tasks, hold individual members accountable for completing those tasks, and ultimately respond to those members who did not take adequate measures to complete tasks.
- On legitimacy: we sought to establish with our members and with non-members/potential members in target groups (cultural memory organizations) a sense that our enterprise is legitimate and can meet their preservation needs.
- On conflict: we wanted to design clear channels for conflict management. We also wanted to designate someone who would have the authority to make hard decisions and steer the program forward if differences of opinion were to erupt.
- On commitment: we sought to foster a sense of ownership in the network across all members. We also wanted our membership structure and operational guidelines to provide members and prospective members with a clear sense of their roles in the network and of the ways that network resources would be disseminated across member institutions.
- On design: we believed that each of the above points could be managed more easily if we chose a network design that would foster the sense of cooperation and that would simultaneously establish clear lines of direction.

With the help of the Milward and Provan report, as well as legal advice from a project member, Dwayne K. Buttler, and an Atlanta-based law firm, Kilpatrick and Stockton, we determined that we had at least three main options for establishing a stable structure for MetaArchive's growth and

sustainability: distributed, centralized with a lead organization, or centralized with a formed management entity. There were benefits and drawbacks to each option. With the flexibility of the distributed model came less certainty and more fragility; with the more fixed structures came a higher overhead. We weighed each of the options in turn.

As a "distributed" network, we believed we would encourage a high level of investment from all of our collaborating institutions. However, the multi-ownership model would not provide a clear channel for leadership. Who would take responsibility for such functions as making sure that the network was running properly, administering funds, documenting our technical and organizational practices, evaluating (and where appropriate, satisfying) emergent standards, hosting events and workshops, and recruiting new members? And who would serve as the clear administrative voice for tasks such as conflict resolution and direction setting for the group? Such concerns led us to look beyond the "distributed" formation.

As a centralized network formed with a lead organization, we would ensure that administrative responsibility was clearly designated. However, the idea of one of the peer institutions that comprised the MetaArchive's membership becoming the official "lead," not just of a short-term project, but of the long-term sustainability of our entire network, quickly proved problematic. Perhaps most importantly, this would necessitate entrusting an institution to lead the collaborative network without letting its institutional goals interfere with the network goals—an improbable scenario. Likewise, if we named a "lead" institution, how would we ensure the stability of the network if and when leadership changes occurred at the lead institution due to staff turnover and the like? Also, this arrangement would require all members to pay their membership fees to a peer—something that none of our institutions felt particularly comfortable doing. And finally, we questioned the effectiveness of the "lead" institution's authority, given that the institution would also play a member role in the network. For all of these reasons, we began considering the idea of engaging a management entity.

The major drawback of establishing a centralized structure with a management entity was the much-increased overhead. Someone would need to research nonprofit formation practices, prepare documentation, and provide administrative leadership in order to create and run the entity. However, the benefits of this management form for our collaborative network significantly outweighed the costs. Specifically, by founding an external organization to manage the network, we would achieve clear leadership—key in managing accountability, conflict, and commitment, as well as in establishing our legitimacy with other institutions and with other collaborative networks. We would ensure that the focus of the management entity stayed on the network, not on other institutional goals. We also would be more able to negotiate and form relationships with other groups, including existing consortia.

Once we determined that all of our partners were interested in setting up a nonprofit business to manage our group, we further distinguished the nature of our enterprise by selecting a name: the MetaArchive Cooperative. We chose the "cooperative" formation quite deliberately as it fit the nature of the business we wanted to establish. In cooperatives, members "buy in" at varying levels depending on their degree of commitment, membership dues directly support the activities of the cooperative, and central staffing is usually kept as minimal as deemed possible. The members of cooperatives have a clear sense of shared ownership—something we felt was imperative to the success of our distributed preservation system.

In 2006, we founded the MetaArchive Services Group, a nonprofit 501(c)(3) host organization for the MetaArchive Cooperative. We named a board of highly successful pioneers in digital library work, including David Seaman (then the director of the Digital Library Federation), Greg Crane (founder of the Perseus Digital Library), Rachael Bower (founder of the Internet Scout Project), and Martin Halbert and Tyler Walters (both instrumental in the founding of MetaArchive). Under the leadership of the board and MetaArchive Services Group Director Katherine Skinner, the MetaArchive Services Group provides administrative services for the MetaArchive Cooperative, including billing, managing, and distributing membership fees; assisting with the administrative functions in organizing and hosting meetings for the MetaArchive Steering Committee and workshops for members and the broader public; holding members accountable for work they have promised to complete; and cultivating relationships with other consortia. The MetaArchive governance, which consists of a steering committee, provides structured leadership for the development of the Cooperative and its network.

We designed our Cooperative Charter and Membership Agreement to document our practices and our roles and responsibilities for both members and the organization.[6] As a cooperative, we seek to effectively pool our resources in order to minimize the cost of digital preservation. To this end, we have created a pricing model that allows the Cooperative to perform its core duties: monitor the network, run annual steering committee meetings and other communications, expand our membership, cover travel that is related to establishing relationships with other consortia or welcoming in new members, and insure institutions such that we can replace member equipment when appropriate and necessary due to technical failures. The fees are intentionally set as low as possible: we are, after all, a library-based initiative that is resolved to serve libraries and other cultural institutions, not to make a profit.

Conclusion

The investment we have made and will continue to make in the Cooperative has an important strategic purpose. As we have discussed, a central

objective of the MetaArchive Cooperative is to create tools and organizational structures that will enable cultural memory organizations to preserve their own digital collections (a core mission of all of our libraries) rather than outsourcing that task to external parties. We fear that if such options are not both available and attractive, many institutions will turn to commercial vendors, entrusting them as the new custodians of our digital heritage. This is problematic in two senses. First, commercial vendors and cultural memory organizations have very different functions. Vendors seek to use information resources to make money; cultural memory organizations by and large work to make information resources accessible, both for today's citizenry and that of the future. In outsourcing, then, we risk moving our cultural resources from the hands of trained information managers and custodians to those with a more capitalistic approach to information management. This may be appropriate for particular types of collections, but certainly should not be the only pathway available to cultural memory organizations. Second, if cultural memory organizations choose to outsource most or all of their digital preservation work, they effectively lose control over their materials. A vendor, after all, may or may not take care of the information management tasks in the ways that librarians, archivists, curators, and other professionals have been trained to do. And finally, by outsourcing instead of working on in-house solutions, cultural memory organizations will lose the ability to perform one of their long-held core missions. This is precarious for the stability of cultural memory organizations into the future—we risk becoming an expendable "middleman" rather than the primary custodian of our cultural objects.

For these and other reasons, we hope to continue assisting cultural memory organizations in preservation—not only of their digital materials, but also of their own livelihood. We are currently in dialog with more than five dozen institutions that have approached the Cooperative with an interest in joining our network. We have assisted three Private LOCKSS Networks in creating networks of their own, and are assisting several more groups as they do so in the coming year. We are heartened by the steady growth of both our Cooperative and of active Private LOCKSS Networks.

We are grateful to Congress and the Library of Congress for funding much of the technical and organizational work that we have conducted to pioneer a distributed digital preservation solution. We are also grateful to the National Historical Publications and Records Commission for helping to fund our transition from a project to a sustainable program. We look forward to helping cultural memory organizations of all types to use and adapt our organizational model and technical framework in order to successfully address one of their core missions during the transition from physical to digital media: that of preservation.

NOTES

1. After solidifying the organizational model and infrastructure, we expanded coverage to additional subject and genre areas.
2. See project documentation at MetaArchive Cooperative (n.d., NDIIPP Project Documentation 2007–9).
3. This schema is based in such commonly used existing schemas as DC, MODS, CLD, and RSLP, and maps to PREMIS (the release of which post-dated our initial work).
4. For migration purposes, the Cooperative will need to know precise information about the format, size, and file type of every object in our system—information that is necessary in order to ensure that we perform the correct migrations on each file. To this end, we will soon extract technical metadata at the item level for every item in the system using a leading, open-source metadata extraction tool.
5. The LOCKSS software does return regularly to the ingest site for its active collections (those that continue to be accessible by Internet) and will thus record changes that occur within the collection and store a record of those within the preservation collection; however, this is not a dependable way to capture changes within collections. For more information, please see the forthcoming publication by Skinner and Halbert.
6. This work was greatly assisted by Evelyn J. Schneider Endowed Chair for Scholarly Communications at University of Louisville, Dwayne K. Buttler. As a lawyer who works within a university library, Dwayne was able to help us to strike a balance between creating an agreement that would be legally binding and communicating our mission, goals, and responsibilities effectively.

REFERENCES

Anderson, M. (2007). Evolving a network of networks: The experience of partnerships in the National Digital Information Infrastructure and Preservation Program (NDIIPP). *Third International Digital Curation Conference*. Washington, DC. Retrieved June 26, 2008, from http://www.dcc.ac.uk/events/dcc-2007/programme/presentations/Day_Two/Martha_Anderson_1400-1430.ppt

Berkman, P. A. (2008) Once in a hundred generations. In K. Skinner and M. Halbert (Eds.) *Strategies for sustaining digital libraries.* Atlanta: Emory University Digital Publications.

Clareson, T. (2006). NEDCC survey and colloquium explore digitization and digital preservation policies and practices. *RLG DigiNews, 10*(1). Retrieved April 5, 2007, from http://www.rlg.org/en/page.php?Page_ID=20894#article1

Library of Congress. (2003, August). National Digital Information Infrastructure and Preservation Program, Program Announcement to Support Building a Network of Partners, Washington, DC. http://www.digitalpreservation.gov/index.php?nav=4

MetaArchive Cooperative. (n.d.) NDIIPP Project Documentation 2007–9. Retrieved January 14, 2009, from http://metaarchive.org/NDIIPPdocumentation.html

MetaArchive Cooperative. (n.d.). Resources. Retrieved January 14, 2009, from http://metaarchive.org/resources.html

Milward, H.B., & Provan, K. G. (2006). A manager's guide to choosing and using collaborative networks. *Networks and Partnerships Series* IBM Center for the Business of Government. Retrieved June 30, 2008, from http://www.businessofgovernment.org/pdfs/ProvanReport.pdf

Pardo, T. A., Burke, G. B., & Kwon, H. (2006). Preserving state government digital information: A baseline report. Retrieved May 1, 2007, from http://www.ctg.albany.edu/publications/reports/digital_preservation_baseline/

Skinner, K., & Halbert, M. (Eds.). (forthcoming). *Guide to distributed digital preservation.* Atlanta: Emory University Digital Publications.

Smith, A. (2006). Distributed preservation in a national context: NDIIPP at mid-point. *D-Lib Magazine, 12*(6). Retrieved June 30, 2008, from http://www.dlib.org/dlib/june06/smith/06smith.html

Teper, T. (2007). The future of preservation in ARL libraries. *ARL Bimonthly Report, 251,* 9–11. Retrieved May 4, 2007, from http://www.arl.org/bm~doc/arlbr251preserv.pdf

Katherine Skinner is the digital projects librarian for the Emory University Libraries, providing leadership and strategic direction for the library's digital initiatives that are supported through sponsored funding. Skinner also serves as executive director of the MetaArchive Services Group, an independent nonprofit 501(c)(3) educational organization dedicated to improving scholarly communication in socially responsible ways (http://metaarchive.org). She is a co-principal investigator on the SouthComb Cyberinfrastructure for Scholars Project (http://southcomb.org), a founder and editorial board member of *Southern Spaces* (http://southernspaces.org) and program manager for the MetaArchive Cooperative, a distributed digital preservation service organization supported by the Library of Congress and the National Historical Publications and Records Commission (http://metaarchive.org). She holds a BA from the University of North Carolina at Chapel Hill and a PhD in American studies from Emory University.

Martin Halbert is director of digital programs and systems at the Emory University Libraries and directs all digital library services and systems functions for the Emory General Libraries. He is responsible for researching and leading library information technology initiatives, including all digital scholarly communication projects of the MetaScholar Initiative (http:// MetaScholar.org). Halbert provides a leadership role within the library for computer systems operations, development, planning, and integration. He is the principal investigator for research projects with budgets totaling $4.8 million. He is the founding president of the MetaArchive Services Group, an independent nonprofit 501(c)(3) educational organization dedicated to improving scholarly communication in socially responsible ways (http://metaarchive.org). With support from the Library of Congress in 2003, he established the MetaArchive Cooperative, a growing consortium of cultural memory organizations that provides distributed digital preservation services (http:// MetaArchive.org).