4th International Digital Curation Conference
December 2008

Comparison of Strategies and Policies for Building Distributed Digital Preservation Infrastructure: Initial Findings from the MetaArchive Cooperative

Martin Halbert,
Emory University

December 2008

Abstract

This paper discusses the importance of a particular approach to building and sustaining digital content preservation infrastructures for cultural memory organizations (CMOs), namely *distributed* approaches that are *cooperatively* maintained by CMOs (rather than centralized approaches managed by agencies external to CMOs), and why this approach may fill a gap in capabilities for those CMOs actively digitizing historical and cultural content (rather than scientific data). Initial findings are presented from an early organizational effort (the MetaArchive Cooperative) that seeks to fill this gap for CMOs. The paper situates these claims in the larger context of selected exemplars of DP efforts in both the United States and the United Kingdom that are seeking to develop effective DP models in an attempt to recognize those organizational aspects (such as the governmental frameworks, cultural backgrounds, and other differences in emphasis) that are UK and US-specific.

The 4th International Digital Curation Conference taking place in Edinburgh, Scotland over 1-3 December 2008 will address the theme *Radical Sharing: Transforming Science?*.

Introduction

Digital preservation (DP) is an emerging field within the still broader field of data management. While I will examine the larger context of different DP initiatives, my narrower primary concern is with preservation of digitized cultural materials in cultural memory organizations (CMOs). By "cultural memory organizations" I mean small to medium-sized libraries, archives, museums, and historical associations, and not enormous national agencies like the US Library of Congress or the British Library. This paper will make some claims about the importance of a particular approach to building and sustaining digital content preservation infrastructures for CMOs, namely *distributed* approaches that are *cooperatively* maintained by CMOs (rather than centralized approaches managed by agencies external to CMOs), and why this approach may fill a gap in capabilities for those CMOs actively digitizing historical and cultural content (rather than scientific data). I will then describe initial findings of an organization (the MetaArchive Cooperative) that seeks to fill this gap for CMOs.

Because the field of digital preservation is still at a relatively early stage of development this paper will also attempt to clarify the context of this discussion and what is *not* being claimed. I will first briefly highlight ambiguities in the emerging digital preservation field that may obscure my claims about the DP needs of CMOs. I will then review the context of selected exemplars of DP efforts in both the United States and the United Kingdom that are seeking to develop effective DP models in an attempt to recognize those organizational aspects (such as the governmental frameworks, cultural backgrounds, and other differences in emphasis) that are UK and US-specific. There are obviously DP projects underway in many other countries; my purpose in focusing here on the US and UK is to understand the current context of the DP landscape in two countries engaged in major DP leadership efforts, countries that simultaneously have many similarities and subtle differences in orientation to the challenges of this emerging field. Without this contextual framing of other parts of the broader field (such as scientific data preservation efforts, and their acceptance of DDP), my claims about DDP and the smaller context of CMOs may make little sense.

In the process of examining the current DP landscape I will also highlight a gap that I believe exists in DP options for CMOs. Finally, I will try to make a case for the strategy that I term *distributed* digital preservation (DDP) in the context of this gap, and then make some limited claims based on the initial experience of an organization that is attempting to fill this gap. The MetaArchive Cooperative is a nonprofit organization of CMOs established in 2003 to foster the advancement of DDP approaches and to use them to preserve valuable research archives of primarily cultural (rather than scientific) materials that are digital in form. The early conclusions of the MetaArchive Cooperative will be offered as insights that may be useful when considering DDP approaches for other CMO associations, but I will also acknowledge the aspects of the effort that may be more aligned with US institutions.

The CMO DP Problem

CMOs hold virtually innumerable archives of idiosyncratic material that are rapidly being digitized in local initiatives. This digital content has important long term value for both research and cultural identity purposes. But CMO professionals frequently lack effective, scalable DP infrastructures. This lack of access to effective means for long term preservation of digital content is aggravated by a lack of consensus on DP issues and professional roles and responsibilities.

An Emerging Field

There is a lack of consensus on the terms, practices, and scope of activity comprising the still emerging field variously termed "digital curation", "data curation", "digital preservation", or "digital librarianship." Should this new field emphasize *curation* with a strong emphasis on repository systems that will provide *access* to the content in question (echoing the emphasis of museum curators on displays of material curated), or *preservation* with a strong emphasis on repository systems that will ensure the *long-term survival* of the content in question (echoing the emphasis of archival conservators on conservation of documents)? Clearly both preservation and access are important, but many programs are wondering where they should devote their initial emphasis on a spectrum of prioritization that has access on one extreme and preservation on the other. More broadly should we consider this a single field of unified practice, or multiple fields with many different areas of specialized expertise and practice?

Clifford Lynch has critiqued the phrase "digital curation" (Day, 2007), commenting on the way that it ambiguously depends on borrowed concepts of "data curation" from the sciences, and the fact that there is no widespread understanding of what skills a "digital curator" should possess. Lynch's comments are emblematic of the uncertainty driving the quest to understand and define digital curation by collaborative efforts such as the Digital Curation Curriculum (DigCCURR) project (http://ils.unc.edu/digccurr/) in the United States and the Digital Curation Centre (DCC) (http://www.dcc.ac.uk/) in the UK. Various discussions appear to situate digital curation as a conceptual descendant of everything from curatorial work in archives to data mining in corporate IT departments. If anything, there is even less consensus on what constitutes the scope and practice of "digital preservation", which variously seems to include all of the topics described previously. These two phrases suffer from what I will call *neologistic ambiguity*: they attempt to articulate critically important new professional responsibilities still emerging (and therefore potentially broad) but which may have many very specific priorities and expectations from practitioners from disparate areas (which are inherently narrow). For example, some people automatically assume that a digital preservation project *must* address *both* access and preservation issues, whereas others assume it *only* addresses preservation.

Marking this ambiguity over the field(s) and terminology, I will focus in this paper on digital preservation rather than digital curation for several reasons. First, I think that preservation concerns, while they obviously relate to access in an extraordinary number of ways, can be separated from access questions in at least some contexts. Second, preservation is prior to access (if the content hasn't survived, it can't be presented). Third, there is much more emphasis in the field today on access and curation than preservation (very much echoing the historical emphasis on access over preservation in research centers and libraries). Finally, although both are ambiguous phrases, digital preservation is simply more familiar to me than digital curation, as DP has been the phrase used most often in the National Digital Information Infrastructure and Preservation Program (NDIIPP) context of the United States in which I have worked. The original legislation that created the NDIIPP did not attempt to define digital preservation, but simply set forth an exploratory agenda:

> "Americans look to libraries to facilitate research in complete, authentic, original, undistorted sources. But we do not yet know how to preserve digital content, or even which content to preserve. Building a digital preservation infrastructure that will work alongside the one

> already in place for print and audiovisual materials poses great technical challenges. But to an even greater degree, it requires forging the legal, economic, and social agreements that will ensure that important digital data are deposited in their original form into a trusted repository for safe custody… In December 2000, Congress passed PL106-554 establishing the National Digital Information Infrastructure and Preservation Program (NDIIPP). It charges the Librarian of Congress to lead a nationwide planning effort for the long-term preservation of digital content. The conference report urges the Library to work jointly with key government agencies…"
> (Library of Congress, 2002)

Given the ambiguity of these early days, it seems useful to preface any claims with a comparative examination of representative digital preservation efforts now underway in different disciplines and nations, as well as observations concerning the different strategies and policy assumptions underlying these efforts.

Comparison of Selected Digital Preservation Efforts

The following is not intended to be a comprehensive list of all digital preservation efforts, but rather is a selection that demonstrates the broad underlying models and patterns of the field. In each case I have tried to identify comparable efforts in the US and the UK for contrast in the strategic approaches of the two countries. My purpose here is twofold; 1) to frame the landscape of digital preservation efforts and the gap that I think exists in digital preservation options for small cultural memory organizations, and 2) to identify the shared belief across the field in the importance of distributed digital preservation strategies, albeit to very different levels of commitment.

National Scientific Research Agency Efforts

There is a strong emphasis in many digital curation/preservation efforts on ensuring access to scientific data created by means of public funding for long term research purposes. The sentiment behind such efforts is captured by the *Declaration on Access to Research Data from Public Funding* of the Organisation for Economic Co-operation and Development (OECD, 2004), which states:

> "Recognising that an optimum international exchange of data, information and knowledge contributes decisively to the advancement of scientific research and innovation; Recognising that open access to, and unrestricted use of, data promotes scientific progress and facilitates the training of researchers; Recognising that open access will maximise the value derived from public investments in data collection efforts... [the governments of 34 OECD nations] DECLARE THEIR COMMITMENT TO: Work towards the establishment of access regimes for digital research data from public funding...

Given that both the US and UK were signatories to this declaration, it is not surprising that curation of scientific information and data sets has been emphasized in major reports published by large government agencies in the two countries that support such research. Examples of agencies with this focus in digital preservation include the Joint Information Systems Committee (JISC) for the Support of Research in the United Kingdom (Digital Archiving Consultancy, 2003) and the National Science Foundation (NSF) in the United States (NSF, 2005). The general point is that since such research was funded by public funds, the information produced with these funds should be

made available to the public permanently as a basic social good. Several models for research agency data services will be examined briefly to understand this part of the DP landscape.

PubMed Central Efforts

While JISC and the NSF are still working toward policies to support this goal, the most successful example to date has been the US NIH Public Access Policy (http://publicaccess.nih.gov/), which was directed by the following statement contained in the text of Public Law 110-161:

> "The Director of the National Institutes of Health shall require that all investigators funded by the NIH submit or have submitted for them to the National Library of Medicine's PubMed Central an electronic version of their final, peer-reviewed manuscripts upon acceptance for publication, to be made publicly available no later than 12 months after the official date of publication…" (Consolidated Appropriations Act of 2008, p. 344)

This public access policy formalized a program first championed by Harold Varmus, director of the National Institutes of Health (NIH), and implemented by the National Library of Medicine (NLM) in the form of the PubMed Central (PMC) repository (http://www.pubmedcentral.nih.gov/) in the year 2000. The PMC has become an enormous and rapidly growing body of publicly available medical scientific literature. The UK PubMed Central (UKPMC) was modeled on the PMC and its aims, and implemented at the British Library (BL) with similar requirements of deposit by funding agencies such as the Welcome Trust. While the two PMC programs acknowledge the importance of DP, their creation was primarily driven by the desire to provide public access to content created with public sector funds.

Notably, these programs were established operationally at leading national libraries (the US NLM and UK BL) with sponsorship by major funding agencies that control the purse-strings of research efforts (the US NIH and the UK Welcome Trust, respectively) and by threatening to withhold future grants in the case of non-compliance have some likelihood of being able to enforce deposit of publications (although it seems problematic that authors are themselves left with the task of securing IP rights from publishers for PMC deposit). These two programs are similar enough that I will jointly label them the PMC Model, and observe that it has so far been moderately successful, and can work in the case of large and centralized research funding agencies as a way to provide public access to scientific information.

While the PMC programs clearly emphasize access, they have begun to devote more attention to the long-term survivability of their accumulated content. The PMC FAQ states: "The long term goal of PMCI is to create a network of digital archives that can share some or all of their respective locally deposited content with others in the network." This goal is presumably informed by the historical observation and generally accepted belief that print content which survived over centuries did so by being replicated in multiple secure repositories that were geographically distributed, a belief that also informs the LOCKSS project described later in this paper. The PMC Model may increase the long term survivability of agency-funded research publications, but it does nothing for the small archive seeking a means of preserving its locally digitized content.

Social Science Dataset Archives

Two notable examples of repositories that have preserved access to social science research datasets over a period of decades are the US Inter-university Consortium for

Political and Social Research (ICPSR) and its associated digital preservation alliance termed Data-PASS, and the UK Data Archive. These repositories have preserved access to thousands of datasets for more than four decades, motivated by precisely the aims articulated in the above cited OECD declaration.

The ICPSR is a membership-based organization with a successful record of preserving access to social science statistical datasets. (CRL, 2006a) Although it is technically an operating unit of the Institute for Social Research at the University of Michigan, the ICPSR has hundreds of member institutions around the world, and a premier council of researchers and data professionals who provide oversight for ICPSR activities. Recently the ICPSR has led a collaborative digital preservation effort termed Data-PASS (part of the NDIIPP, which will be discussed later in this paper) that also involves several other organizations such as the Odum Institute and Roper Center which preserve datasets. There are several points to highlight concerning ICPSR and Data-PASS. First, although ICPSR is not a governmental agency, its predominance as an effective long term preservation and access center for social science datasets over decades has made it a de facto trusted source of authentic datasets for researchers. Second, despite this trusted status and successful record of preserving data over long periods, the ICPSR sought to partner with other repositories to implement a collaborative DDP strategy when the opportunity presented itself in the context of NDIIPP. In this strategy a preservation network of geographically dispersed sites is created to securely replicate copies of data as a way of ensuring long term survival of information. This echoes the goal of the PMC to create a similar network of allied archives replicating its content for long term preservation purposes.

The UK Data Archive (UKDA) has operated at the University of Essex since 1967, and functions in many ways as the UK equivalent of the US ICPSR, preserving access to social science datasets for researchers over the long term. In 2004 it began collaborating with The National Archives (TNA) of the UK on mutual digital preservation activities as part of the JISC Digital Preservation and Asset Management Programme. While this collaboration did not result in a full blown effort to create a DDP infrastructure, the UKDA storage strategy was designed to keep "up to six copies of the same data file on at least four separate preservation servers" (UKDA/TNA, 2005), although this infrastructure is all located on the grounds of the University of Essex. (UKDA, 2008)

A general point here is that in both of these cases even canonical repositories of particular kinds of scientific data are increasingly seeking to create shared digital preservation infrastructures with multiple nodes and partners. Another similarity between these two efforts is that both the US ICPSR and UKDA actively chose to pursue major new collaborative digital preservation efforts as part of national solicitations, the NDIIPP in the US, the JISC Digital Preservation and Asset Management Programme in the UK (although these solicitations were significantly different in nature, as discussed below in the section on cross-disciplinary national efforts). And finally (and obviously), I will note again that although successful, this model does little to address the needs of CMOs.

Big-Science Agency Efforts

"Big Science" funding agencies like the US National Science Foundation (NSF) and UK Higher Education Funding Council for England (HEFCE) obviously have both the financial means and motivation to be successful in this kind of endeavor. Additional "Big Science" efforts with characteristics superficially resembling the PMC Model are now being considered in the US, UK, and other countries. The UK appears

to be significantly farther along in planning such efforts with a recently released study commissioned by the HEFCE articulating the case for a UK Research Data Service (UKRDS). (Serco Consulting, 2008) The UKRDS would form an umbrella "organization structure and governance approach" for prescriptively developing and sustaining technology, standards, training, marketing, and funding for the management, preservation and curation of research data. The US NSF has not yet proposed a similarly coherent service or coordinated service framework, although the 2005 report *Long-Lived Digital Data Collections* (NSF, 2005) and 2007 NSF *Cyberinfrastructure Vision for 21st Century Discovery* report (NSF, 2007) both highlight the importance of systematic management of scientific data as a social good, as well as talking in terms of cyberinfrastructure "tools" and "environment".

The UKRDS report examines a comprehensive set of access and preservation functions and three organizational models for accomplishing these functions: the status quo, a centralized model, and a decentralized "hybrid/umbrella" model. The report considers various advantages and disadvantages associated with each of these models, and ultimately favors the umbrella approach as the most flexible, least invasive, and easiest to potentially achieve. Quite apart from its recommendations regarding next steps in planning a prospective UKRDS and the characteristics that such a service would have, there are two notable points of context to highlight about the UKRDS report and how (unlike the case of the PMC Model) US and UK assumptions seem to differ concerning how to proceed in improving access and preservation to scientific data. The UK strategy for planning and implementing scientific data service initiatives appears to be more prescriptive and centrally coordinated than the typical US approach which most often simply announces a competitive grant opportunity and relies on coordination to take place at the level of institutions jointly applying for funds. An example of this less centrally coordinated style in the US is the NSF-funded National Science Digital Library (NSDL), which set forth only general categories and guidelines for applications in its first NSF competitive award funding cycle in 2000 (Zia, 2001) and has continued under a very loose pattern of coordination in the years since. The recent NSF Datanet solicitation likewise does not lay out a prescriptive set of aims but simply seeks to create "a set of exemplar national and global data research infrastructure organizations (dubbed DataNet Partners) that provide unique opportunities to communities of researchers to advance science and/or engineering research and learning." (NSF, 2008) The resulting services selected for funding will presumably not be required to address the entire data access and preservation challenge, unlike the proposed umbrella program of the UKRDS. Although UK funding agencies also typically make competitive awards, a greater degree of prescriptive direction is evident in solicitations from JISC and other UK agencies. This may provide a greater capability to target a grand challenge such as the data access/preservation question.

In the US there has been much less faith in large prescriptively coordinated governmental solutions to emerging problems. This may seem curious given the fact that the successful centrally coordinated PMC model originated in the US, but is consistent with the general skepticism by recent US administrations toward any sort of governmental solutions to large-scale social problems. This may change under the next administration of the US government, but in general it still seems likely that UK science agency efforts will continue to be characterized by relatively more centrally coordinated and top-down approaches to data access and preservation. And once again to state the obvious, none of these big-science efforts help DP efforts in small libraries, in either the UK or the US.

Cross-Disciplinary National Efforts

There have been several prominent national efforts in both the US and UK that crossed disciplines and institution types. I have previously mentioned the National Digital Information Infrastructure and Preservation Program (NDIIPP) undertaken by the US Library of Congress as a prominent endeavor, and notably one that selected projects with a highly entrepreneurial character as start-up digital preservation projects. The NSF DataNET effort is still in its inception, and the NSF NSDL was not a digital preservation effort, but it is likely that DataNET projects selected will (like NSDL and NDIIPP projects) be highly entrepreneurial in character. There have been a number of major cross-disciplinary efforts in the UK focused on digital preservation, examples being the Digital Curation Centre (DCC) and the JISC Digital Preservation and Asset Management Programme, as well as pan-european efforts with significant UK involvement, such as the Preservation and Long-term Access through NETworked Services (PLANETS) consortium, the Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR) project, the Network of Expertise in Long-Term Storage of Digital Resources (NESTOR) project, the DigitalPreservation-Europe (DPE) effort, and others.

There is not sufficient space in this paper to discuss all of these programs, so I will simply highlight some comparative points that I take away. All of the UK efforts were collaborative efforts between different institutions, but (unlike the US NDIIPP) were not vested by the government in the national library, the BL. Both US and UK efforts were primarily funded by governmental agencies, in the US by congressional appropriation and in the UK primarily by JISC funding. Both US and UK efforts primarily focused on institutions of higher education (universities), although they have also attempted to involve information technology corporations at times.

It is hard to generalize, but there does seem to be a more coordinated and prescriptive element to the UK efforts. The NDIIPP only put forward broad solicitation guidelines, and did not attempt to strongly direct the projects that it funded, in fact taking a relatively "hands-off" approach. While there have been attempts to synergize NDIIPP efforts with the NSF by means of NSF solicitations that were funded by NDIIPP, unlike the JISC efforts there was no attempt to recommend or enforce either best practices or established standards. The UK efforts are better aligned than the US efforts, with more focus on the creation of shared training programs, standards documentation, and other coordinating functions. Both the UK and US have funded significant tools development, again with more emphasis on coordination in the UK through frameworks such as the JISC Integrated Information Environment (http://www.jisc.ac.uk/whatwedo/themes/informationenvironment).

There does seem to be a greater emphasis on preservation of content (rather than technology alignment or training) in the US NDIIPP, and more inclusive attention to cultural content (rather than just scientific research data). This may arise from the greater focus of the Library of Congress on historical information, as contrasted with the emphasis placed on scientific data by JISC. This difference in emphasis also plays out in affiliated non-governmental efforts, and government initiatives like NDIIP have thereby begun to indirectly benefit CMO DP needs.

Non-Governmental Efforts

There are a small number of non-governmental efforts that I would highlight as relevant to this discussion of digital preservation. They each epitomize a particular model of digital preservation and engagement with the community, in these cases primarily concerning electronic journal (e-journal) content.

The LOCKSS Alliance (http://www.lockss.org) is an international nonprofit association of libraries that preserves e-journal content to which the member libraries subscribe. (Reich, 2001) The LOCKSS network is comprised of servers running the LOCKSS open-source software, and allows libraries to cache copies of e-journal content in multiple secure geographically dispersed locations as a way of safeguarding this content. The LOCKSS Alliance is not operated by any governmental agency, although it has received NDIIPP funding from the Library of Congress. The LOCKSS Alliance has approximately two hundred member institutions in countries around the world. LOCKSS is subject domain neutral, being focused on e-journals rather than scientific data.

Portico (http://www.portico.org) is a nonprofit vendor of digital preservation services based in the US, initially focusing on e-journal content. Portico is also not operated by any governmental agency, although it has also received NDIIPP funding. Portico is much more centralized philosophically than LOCKSS and does not implement a geographically distributed digital preservation network with many multiple nodes, but does cache content in at least two locations.

LOCKSS and Portico are successful approaches for libraries seeking to preserve their subscribed e-journal content, but still provide no facilities for preserving locally digitized content. However, the Portico service is now considering providing such services, and LOCKSS has generated several spinoffs that directly address CMOs, as described in the next sections.

Cultural Memory Organizations and Distributed Digital Preservation Strategies

Long term preservation of information has historically succeeded primarily by caching copies of content in secure archives that are geographically distributed. While these are still early days for the emerging field of digital preservation, it seems likely that preservation of digital information will likewise succeed over long periods of time through similar strategies of securely distributing copies of content. Clearly, most of the large governmental scientific DP efforts described above have a fundamental faith in the importance of DDP as a strategy for long term survival of information.

This claim likewise underlies a rapidly growing series of projects based on the LOCKSS model for distributed digital preservation (Reich, 2002); projects which reuse the LOCKSS open source software in new network implementations. These projects include the original LOCKSS network for e-journal content preservation (http://www.lockss.org), the MetaArchive Cooperative which is the main focus of this paper, the Alabama Digital Preservation Network (http://www.adpn.org), the distributed "stacks" of the Arizona PeDALS project (http://rpm.lib.az.us/pedals/), the CLOCKSS shared archive maintained by publishers and research libraries (http://www.clockss.org), and several other similar efforts now in development.

PLNs versus other DDP Systems

The networks which followed the original LOCKSS network have now come to be generically termed "Private LOCKSS Networks", or PLNs for short. While PLNs are certainly not the only technological solution for implementing secure and distributed digital preservation networks, they are becoming an identifiable trend with a certain amount of momentum.

Let me say this again more directly so I am not misunderstood: ***PLNs are not the only way to implement a DDP strategy.*** Depending on organizational requirements,

there are any number of effective DDP solutions, ranging from complex approaches such as IRODS (https://www.irods.org) and the Sherpa-DP AHDS (Knight, 2007) to much simpler file replication technologies (which may admittedly entail more manual intervention). However, the PLN model offers a relatively simple, low cost mechanism for disparate CMOs to quickly establish a network for cooperative digital preservation purposes. There are nevertheless pros and cons of such cooperatives.

Cooperatives for Digital Preservation

The importance of collaboration and trust in creating DDP infrastructures has been emphasized in the literature recently. (Day, 2008) I claim that the creation of DDP infrastructures for CMOs can most effectively accomplished through cooperative approaches between existing institutions to build up their capacity for ensuring data viability over long periods. This claim is based on several factors:

1. Digital technologies provide a mechanism for replicating data indefinitely. Unlike the content of their physical archives, CMO digital archives can be securely replicated for preservation purposes in multiple locations. All that is required is an appropriate networked infrastructure, typically servers in institutionally maintained server rooms.
2. While most contemporary CMOs engaged in digitization activities (even fairly small ones) will typically maintain such a server room operation, they do *not* typically maintain multiple rooms, at least not multiple rooms that are significantly separated geographically (a pre-requisite for disaster survivability). It is also not economically realistic for any one cultural memory institution to establish a geographically dispersed infrastructure for survivability. While Big-science agencies have the both the resources and the mandate (as described above) to create such robust infrastructures, CMOs are unlikely to have this capability within the foreseeable future.
3. The organizational model of a *cooperative* (an enterprise in which the infrastructure of the endeavor is owned and operated by its users) is logical for CMOs seeking to collaboratively enable distributed preservation in this manner. Libraries and archives are accustomed to maintaining facilities with long-term commitments to external groups in order to realize mutual benefit in preserving knowledge artifacts for posterity. A cooperative structure allows such institutions to maintain their autonomy while still working together.
4. While it might be possible to contract for such a distributed infrastructure from vendors, CMOs often lack the funding to pay for the expense typical of such vendors solutions.
5. Cooperatives are likely to be the most *affordable* way for cultural memory insitutions to acquire access to a robust, distributed digital preservation infrastructure. The economics of low cost replication network solutions like PLNs are such that they take advantage of the sunk costs of already-established server room operations. The incremental costs of installing and maintaining a LOCKSS-based appliance server in terms of hardware and labor are so low that they are far outweighed by the benefits of obtaining access to a distributed set of facilities. Cooperatives that leverage sunk costs in this way are always likely to beat out profit-centered solutions, which are usually focused on more highly capitalized corporate or large government clients as mentioned above.

While cooperatives may arguably be the most affordable organizational model for CMO DDP, traditionally oriented institutions like libraries and archives may very well be hesitant to take up the task of establishing high technology cooperatives.

New Roles for Cultural Memory Organizations

If they are to realize the benefits of scalable shared DDP infrastructure, libraries, archives, and other CMOs must now rise to the challenge of building and sustaining distributed preservation infrastructures for digital content, as well as accumulating the associated expertise needed to support such infrastructures. This claim may seem either radical or obvious, depending on where one sits and how one understands the missions of cultural memory institutions.

Some leaders of libraries and archives balk at conceptions of their institutions that internalize responsibilities for maintaining networked infrastructures. They may see libraries and archives as utilizing some of the capabilities of networked infrastructures, but they do not see such infrastructures as part of the mission or scope of operations of their organizations. Rather, they understand networked systems as something maintained by the phone company, or Google, or some entity "out there". They see their institutional missions as properly limited to the practices involved in maintaining print or other physical forms of cultural memory.

Conversely, some leaders of libraries and archives understand their mission as properly encompassing all of the changing practices involved in providing for the information needs of the research communities they serve. This may entail the maintenance of many different types of information infrastructures, some print, and some digital.

As access to information across all segments of society becomes increasingly digital rather than analog, organizations that see their missions constrained by the analog forms of information will become increasingly limited and marginalized. It is incumbent on leaders of libraries, archives, and other cultural memory organizations to understand their missions and roles as encompassing the larger scope of the digital universe of knowledge and not only the print universe. They will fail to effectively serve their clienteles otherwise.

The concrete implication of this assertion as relates to preservation is that, yes, creating and maintaining digital preservation networks are part of what libraries and archives do in the 21st Century. It is not someone else's job, digital preservation is part of the responsibility of maintaining research collections for posterity. This does not minimize the challenge of such activities; these are not trivial tasks, they are challenging, and unlikely to be easily accomplished by any one library or archive. This means that our cultural memory institutions must work together to accomplish their ends. The most effective way to do this is through distributed infrastructures, as maintained above and supported by the commitment to DDP solutions in much larger efforts. But there are admitted challenges to sustaining CMO DDP networks.

Sustaining Distributed Digital Preservation Infrastructures

To sustain distributed digital preservation infrastructures, groups must successfully undertake a variety of tasks. The most difficult elements of creating and maintaining distributed digital preservation networks are not technical, but organizational. Because such networks have been uncommon or non-existent in the past, there are many foundational requirements, such as analyzing business and cost models, undertaking long-term strategic planning, and simply figuring out how to run them effectively.

Unincorporated associations are one way of accomplishing this task. Projects such as the Alabama Digital Preservation Network demonstrate that groups of libraries can implement PLNs without the formation of new organizations. These kinds of organizational structures are very lightweight, can form quickly, and have reasonable

likelihood of succeeding over time.

However, it is my contention that there is also need for new kinds of collaborative organizational frameworks, specifically nonprofit cooperatives dedicated to the distributed digital preservation of cultural research information. I make this claim after five years of practical efforts to advance the practice and understanding of collaborative inter-institutional digital preservation. The following are some of the reasons I would highlight as findings informing this claim:

1. CMOs are competitors, not for monetary profit, but for institutional prestige. Alliances and other frameworks for unincorporated associations often suffer from the fact that some single larger institution must function as the leadership of group efforts. This basic fact serves to undermine many collaborative efforts. If University X is the functional recipient of funds or membership fees by other universities, the other universities see the effort as "University X's project." They understand funding that goes to University X as contributing to the greater glory of that institution and not their own, even if such funding is directly dispersed for joint expenditures.
2. Separate organizations that do not comprise a competing CMO or parent research institution avoid this problem. This is why universities and other research centers are often more willing to direct jointly pool funding to a commercial vendor than a mutual collaboration. But outsourcing key research functions to commercial entities can lead to a disastrous loss of control; witness the so-called crisis in scholarly serials resulting from the outsourcing of these publications to vendors like Elsevier.
3. Cooperatives can avoid both pitfalls. Because all assets of the endeavors are owned and retained by the individual CMOs, there is neither a perceived or real loss of either control or prestige by participants. However, an incorporated nonprofit organization can accomplish functions that an unincorporated association cannot. It can serve as a legal entity to make contractual commitments to, it can collect funds and disperse them, etc. These capabilities are essential to mobilizing group efforts effectively.

If digital preservation cooperatives are acknowledged as an effective mechanism, what are best practices for the creation of such entities?

The MetaArchive Cooperative

The MetaArchive Cooperative (http://www.metaarchive.org) provides a model for an incorporated nonprofit organization of research libraries created as part of the U.S. National Digital Information Infrastructure and Preservation Program that has established an effective model for shared distributed digital preservation infrastructures development. The MetaArchive Cooperative was created by a group of research libraries as a means of mobilizing efforts for distributed digital preservation, after consideration of various organizational options highlighted the following points:

- If libraries were to make commitments to mutualistically preserve digital content, the libraries involved needed to make such contractual obligations to some legal entity. In practical terms, the libraries could not make one-to-one agreements with each other as this quickly became an N x N scaling problem as the preservation network grew. There needed to be some central entity to which all of the contractual commitments could be directed.
- This central entity could not be one of the constituent members of the network, both because the network commitments needed to be able to survive the withdrawal of any one member and because no one research center was willing to

assume the overhead of being the organizational host for such agreements.

- For reasons described in the previous section, an independent agency was needed that would not itself be perceived as a competing research institution. A cooperative that was a) operated by member institutions and b) which owned nothing itself could effectively serve in this capacity. By organizing the cooperative as an operation of a specially-tasked U.S. 501(c)3 nonprofit organization, the distributed digital preservation operation was made much more palatable to prospective participants.
- Because MetaArchive made the decision early on to embrace the PLN technical strategy, there was no need for the organization to own or develop proprietary software solutions. The cooperative could focus on the work of digital preservation and not developing new technical solutions.

Although it is a nonprofit cooperative and now international in scope (Hull University in the UK is a member), the MetaArchive Cooperative may have some admittedly entrepreneurial characteristics more typical of US organizations. For example, it has been unconcerned with seeking governmental mandates or sanctions at the state or federal level beyond the sponsorship it received from the US NDIIPP. MetaArchive has been more concerned with the simple aims of providing workable low-cost DDP solutions for CMOs. Our shared infrastructure is still relatively young, but is based on a technology proven for even larger networks (LOCKSS), a technology that we do not claim is perfect or the only DDP solution, but simply an *effective one*. There will likely be problems that we have not yet foreseen in scaling the network up to hundreds of global nodes, but such problems are to be expected.

Conclusions

While there are many large scale governmentally sponsored digital preservation initiatives now underway in the US and UK, most of these efforts do not address the needs of small cultural memory organizations for preservation of local digitized content. The MetaArchive Cooperative has now been in operation for several years, and is rapidly gaining experience in how to marshal efforts among cultural memory organizations for distributed digital preservation. With the steady addition of new constituent members it offers an example of an effective strategy for organizing the efforts of disparate libraries and archives around digital preservation functions. It has become a center of excellence in private LOCKSS network implementations, now offering training and joining options for institutions seeking to apply such solutions. The founding members of the MetaArchive Cooperative look forward to a future of collaborating distributed digital preservation organizations for our shared global cultural memory, a future that can be sustained as a long-term social priority.

Acknowledgements

The MetaArchive Cooperative would like to acknowledge the generous support of both the US Library of Congress National Digital Information Infrastructure and Preservation Program and the US National Historic Publications and Records Commission.

References

[report] Beagrie, N. (2003). *National Digital Preservation Initiatives: An Overview of Developments in Australia, France, the Netherlands, and the United Kingdom and of Related International Activity*. (CLIR Publication No. 116). Council on Library and Information Resources, Washington, D.C. Retrieved October 6, 2008, from http://www.clir.org/pubs/reports/pub116/pub116.pdf

[proceedings]Berman, F., et al (2007). The Need for Formalized Trust in Digital Repository Collaborative Infrastructure. *NSF/JISC Repositories Workshop (held April 17-19, 2007 in Phoenix, Arizona, USA)*. Retrieved July 25, 2008 from http://www.sis.pitt.edu/~repwkshop/papers/berman_schottlaender.html

[proceedings] Brophy, P. & Fisher, S. (2001) Evaluating the Distributed National Electronic Resource. *Proceedings of the First ACM/IEEE Joint Conference on Digital Libraries.* pp. 144-145.

[report] CRL (2006a). ICPSR Audit Report For the period ending 24 October 2006. Center for Research Libraries. Retrieved October 6, 2008, from http://www.crl.edu/PDF/ICPSR_final.pdf

[report] CRL (2006b). Portico Final Report Draft. Center for Research Libraries. Retrieved October 6, 2008, from http://www.crl.edu/PDF/Portico_Final_Report_10-06.pdf

[report] CRL (2007). LOCKSS Audit Report. Center for Research Libraries. Retrieved October 6, 2008, from http://www.crl.edu/PDF/LOCKSS_Audit_Report_11-07.pdf

[statute] Consolidated Appropriations Act of 2008 (2007). H.R. 2764, 110th Cong. Retrieved October 6, 2008, from http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=110_cong_bills&docid=f:h2764enr.txt.pdf

[journal article] Day, M. (2007). Report from the DigCCurr 2007 International Symposium on Digital Curation, Chapel Hill, NC, April 18-20, 2007. *The International Journal of Digital Curation*, Issue 1, Vol. 2, p. 102-111. Retrieved October 6, 2008, from http://www.ijdc.net/ijdc/article/view/28/31

[journal article] Day, M. (2008). Toward Distributed Infrastructures for Digital Preservation: The Roles of Collaboration and Trust The Roles of Collaboration and Trust. *The International Journal of Digital Curation*, Issue 1, Vol. 3, pp. 15-28. Retrieved October 6, 2008, from http://www.ijdc.net/ijdc/article/view/60/61

[report] Digital Archiving Consultancy. (2003) *e-Science Curation Report Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision prepared for: The JISC Committee for the Support of Research (JCSR).* Retrieved October 6, 2008, from http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

[report] Knight, G. & Anderson, S. (2007). *SHERPA DP: Final report of the SHERPA DP project.* Retrieved October 6, 2008, from http://www.jisc.ac.uk/media/documents/programmes/preservation/sherpa%20dp%20final%20report.doc

[report] Library of Congress. (2002). *Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and preservation Program.* Retrieved October 6, 2008, from

http://www.digitalpreservation.gov/library/resources/pubs/docs/ndiipp_plan.pdf

[report]National Science Foundation (NSF). (2003). *It's About Time: Research Challenges in Digital Archiving and Long-Term Preservation*. Retrieved October 6, 2008, from http://www.digitalpreservation.gov/library/resources/pubs/docs/about_time2003.pdf

[report] National Science Foundation (NSF). (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. Retrieved October 6, 2008, from http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

[report] National Science Foundation. (2007). *Cyberinfrastructure Vision for 21st Century Discovery*. Retrieved October 6, 2008, from http://www.nsf.gov/pubs/2007/nsf0728/index.jsp

[report] National Science Foundation. (2008). *Sustainable Digital Data Preservation and Access Network Partners (DataNet)*. Retrieved October 6, 2008, from http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.pdf

[report] OECD (2004). Declaration on Access to Research Data from Public Funding. (Organisation for Economic Co-operation and Development, Paris). Retrieved 20/12/07 from http://www.codataweb.org/UNESCOmtg/dryden-declaration.pdf

[report] Serco Consulting. (2008). *UKRDS Interim Report:The UK Research Data Service Feasibility Study*. (Version: v0.1a.030708) Serco Consulting, London. Retrieved October 6, 2008, from http://www.ukrds.ac.uk/UKRDS%20SC%2010%20July%2008%20Item%205%20(2).doc

[Internet journal] Reich, V. & Rosenthal, D. (2001). LOCKSS: A Permanent Web Publishing and Access System. *D-Lib Magazine 7,(6)*. Retrieved October 6, 2008, from http://www.dlib.org/dlib/june01/reich/06reich.html

[Internet journal]Reich, V. (2002) Lots of Copies Keep Stuff Safe As A Cooperative Archiving Solution for E-Journals. *Issues in Science and Technology Librarianship*. Fall 2002. Retrieved July 25, 2008, from http://www.istl.org/02-fall/article1.html

[Internet journal]Rosenthal, D., et al. (2005). Requirements for Digital Preservation Systems: A Bottom-Up Approach. *D-Lib Magazine 11,(11)*. Retrieved July 25, 2008, from http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html

[report] UKDA/TNA (2005). *Assessment of UKDA and TNA Compliance with OAIS and METS Standards.* Retrieved October 6, 2008, from http://www.data-archive.ac.uk/news/publications/oaismets.pdf

[report] UKDA (2008). *UK Data Archive Preservation Policy*. Retrieved October 6, 2008, from http://www.data-archive.ac.uk/news/publications/UKDAPreservationPolicy0308.pdf

[journal article] Zia, L. (2001). The NSF National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL) Program: New Projects and a Progress Report. *D-Lib Magazine*, Volume 7, Number 11. Retrieved October 6, 2008, from http://www.dlib.org/dlib/november01/zia/11zia.html