February 2009

MetaArchive: A Cooperative Approach to Distributed Digital Preservation

Katherine Skinner

Emory University, katherine.skinner@emory.edu

Martin Halbert

Emory University, martin.halbert@emory.edu

Follow this and additional works at: <https://docs.lib.purdue.edu/atg>

Part of the [Library and Information Science Commons](#)

Recommended Citation

Skinner, Katherine and Halbert, Martin (2009) "MetaArchive: A Cooperative Approach to Distributed Digital Preservation," *Against the Grain*: Vol. 21: Iss. 1, Article 6.

DOI: <https://doi.org/10.7771/2380-176X.2489>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Endnotes

1. Jones, Maggie. "Archiving E-Journals Consultancy (Final Report)," October 2003, accessed January 16, 2009; available from http://www.jisc.ac.uk/uploaded_documents/ejournalsfinal.pdf; Internet.
2. Kenney, Anne R., Richard Entlich, Peter B. Hirtle, Nancy Y. McGovern, and Ellie L. Buckley. "E-Journal Archiving Metes and Bounds: A Survey of the Landscape," September, 2006, accessed January 16, 2009; available from <http://www.clir.org/PUBS/abstract/pub138abst.html>; Internet.
3. Sparks, Sue, Hugh Look, Adriene Muir and Mark Bide. "Scoping study for a registry of electronic journals that indicates where they are archived," **Rightscom and Loughborough University**, 2008; accessed January 2009, 33; available from <http://www.jisc.ac.uk/media/documents/programmes/preservation/ejournalregstudy.pdf>; Internet.
4. *ibid*, 13.
5. SUNCAT includes serials' information from over 60 UK libraries, including the **British Library**, the **National Libraries of Scotland and Wales**, some of the largest Higher Education institutions in the UK and a number of specialist libraries. <http://www.sunecat.ac.uk/>.
6. <http://edina.ac.uk>. In addition to SUNCAT, EDINA also provides organisational support for the **UK LOCKSS Alliance**, <http://edina.ac.uk/lockss/> following project work at the **Digital Curation Centre**. The **University of Edinburgh** is a **CLOCKSS Archive Node**.
7. <http://www.issn.org/>; Internet.
8. The **Directory of Open Access Journals (DOAJ)** requires ISSN assignment.
9. Accessed on January 19, 2009; available from <http://www.issn.org/2-22637-What-is-an-ISSN-L.php>; Internet.
10. Terry Morrow, Neil Beagrie, Maggie Jones and Julia Chruszcz, Tee Em Consulting, "A Comparative Study of e-Journal Archiving Solutions," May 2008. http://www.jisc-collections.ac.uk/news_and_events/news_articles/ejournals_pr.aspx.
11. These six, **LOCKSS (Lots Of Copies Keep Stuff Safe)**, **CLOCKSS (Controlled LOCKSS)**, **Portico**, **e-Depot** (from the Dutch national library, Koninklijke Bibliotheek, KB), **OCLC Electronic Collections Online (ECO)**, and **British Library e-Journal Archiving Programme**, were selected on relevance to UK, with some relation to the 12 e-journal archiving initiatives included within the *CLIR Report* (op cit) as having met seven indicators of viability.
12. Kenney et al, op. cit., 3.
13. *ibid*, 32.
14. *ibid*, 16.
15. Comments should be sent to the author at <p.burnhill@ed.ac.uk>. A full version of this article is to appear in *Serials* in time for the UK Serials Group Meeting in April 2009.

MetaArchive: A Cooperative Approach to Distributed Digital Preservation

by Katherine Skinner (Digital Projects Librarian, Emory Univ., 540 Asbury Cir., Atlanta, GA 30322; Phone: 404-783-2534; Fax: 404-727-0827) <katherine.skinner@emory.edu> <http://metaarchive.org>

and Martin Halbert (Director for Digital Innovations, Emory University, 540 Asbury Cir., Atlanta, GA 30322; Phone: 404-727-2204; Fax: 404-727-0827) <martin.halbert@emory.edu> <http://metaarchive.org>

What role will the Library take in digital preservation? On first glance, the question seems relatively easy to answer. As the library continues to transition from its centuries-long focus on print assets to a combination of print and digital resources, it will take an active role in the preservation of our digital cultural resources that is similar to that which it has long undertaken in the print realm.

Or will it?

Of late, many of us in the library field have become preoccupied with the concept of digital preservation — and rightly so. We wonder aloud about the forms that digital preservation will take, the amount it will cost, the rigor demanded in its implementation, and the feasibility of different organizational approaches to digital preservation.

But what does it mean to participate? How do we want to be involved? And what role(s) should we, as librarians and archivists, aspire to take in the realm of digital preservation?

Questions such as these led to the founding of the **MetaArchive Cooperative**, a collaborative network of institutions that have banded together to communally approach the challenges of preserving digital assets. The original six members founded this **Cooperative** due to their strong belief that libraries both could be and should be *actively engaged* in the creation and maintenance of their own digital preservation solution. They knew that alone, none of these institutions were likely to create and maintain — much less sustain — a robust digital preservation solution. However, they believed that if they approached the issue as a group and built a *shared* infrastructure, they could accomplish together what no one institution had the resources to achieve in isolation.

The MetaArchive Cooperative: A Shared Digital Preservation Infrastructure

The **MetaArchive Cooperative** (<http://MetaArchive.org>) formed to enable cultural memory organizations to effectively and mutually preserve their archival digital assets *for themselves*. **MetaArchive** began in 2004 as one of the original eight initiatives contracted by the **Library of Congress** under the **National Digital Information Infrastructure and Preservation Program (NDIIPP)**.¹ The venture was led by **Emory University** in collaboration with **Georgia Tech**, **University of Louisville**, **Virginia Tech**, **Auburn University**, **Florida State University**, and the **Library of Congress**.

The network established by this group was the first major effort to build and operate a private implementation of the open source **LOCKSS (for Lots of Copies Keep Stuff Safe)** software for digital preservation (<http://www.lockss.org>), an approach that has since been termed a **Private LOCKSS Network**, or **PLN**. The **MetaArchive PLN** is a distributed preservation infrastructure that meets the **OAIS Reference Model** standards for repositories.²

Technically speaking, the foundation of the network is the open source **LOCKSS** software developed at **Stanford University**, which enables a group of **LOCKSS** caches, or node servers, to work together across geographical space to replicate and preserve content.³ **MetaArchive** is the only **PLN** in operation thus far that does not depend on the **LOCKSS** team to administer the network; we run a separate cache manager (coded in collaboration with the **LOCKSS** team) to monitor our network. The **MetaArchive Cooperative** has created and layered additional modules on top of the **LOCKSS** framework to provide our members with administrative tools, including a conspectus database and the cache manager. The conspectus database enables members to capture collection-level metadata for preservation decisions and actions, and the cache manager serves as a monitoring tool for network-wide tracking and troubleshooting activities. We are in the process of packaging these open source software components for use by other **PLNs**, and plan to release this software through **SourceForge** next year.

The organizational framework that we have constructed has been as integral to our success as the technological platform upon which we have built our preservation services. After running the network for three years, we transitioned from a sponsored-funding-supported project to an independent, membership association in 2007, a transition that has been greatly assisted through the support of the **National Historical Publications and Records Commission**. As part of this work, we founded a 501c3, the **MetaArchive Services Group**, to administer the **Cooperative**. All of the components of the network we run are owned and maintained by our member institution. This decentralized apparatus enables the **Cooperative** and its services to be *independent*

continued on page 38

of each member — our members learn how to run and operate their own preservation node for the network, building their internal knowledge of the preservation process. They also are given opportunities to contribute to the software development efforts undertaken by the **Cooperative**.

The mission of the **MetaArchive Cooperative** is to support, promote, and extend our collaborative approach to distributed digital preservation practices. We have made our organizational model available to others as an example of how to create shared digital infrastructure. To this end, we not only run our own network, but also provide training and consulting assistance to other groups that wish to found similar preservation networks. We host workshops and make all of our documentation freely available to other collaborative projects and programs.

Unlike the public **LOCKSS** network, where participant libraries preserve journal content in which they all have a vested interest, the collections in the **MetaArchive** network are the unique holdings of each participant library and archive. In other words, **MetaArchive's** members cannot rely on the incentive that drives participation in the public **LOCKSS** network — a shared body of content to which all subscribe and upon which all rely. Instead, the **MetaArchive** network requires a strong commitment between constituent institutions — each participates in order to preserve their own data in exchange for preserving other institutions' data.

So what *are* the drivers in this **PLN** scenario? Topping the list are a strong sense of community engagement and a strong belief in the library's cultural stewardship role. Our members share the conviction that libraries have a vested interest in preserving their own digital assets. Each has determined that they do not want to cede all of their digital preservation activities to external groups, and do want to participate in creating their own preservation solution. Building alone is a costly proposition, so these institutions have coupled their resources in order to achieve their preservation goals in a community-based effort.⁴

To enable this, **MetaArchive** formed as a **cooperative**, not a vendor. **MetaArchive's** members do not pay for services, but rather make an investment to create and sustain their own preservation infrastructure. The **Cooperative** is more than a technical solution for preservation. It also functions as a learning environment in which members gain experience in developing and enacting a full preservation plan for their assets. Each member both contributes to and benefits from the expertise and the technical infrastructure developed by the overall community. In keeping with these principles, membership fees are kept at the absolute minimum required for the operation of the **Cooperative**, and range from \$300 to \$5K per year, together with a fee of \$2 every 3 years per 1 GB of content contributed. These

continued on page 40

Katherine Skinner & Martin Halbert

against the grain people profile

Skinner is Digital Projects Librarian
Halbert is Director for Digital Innovations
Emory University, 540 Asbury Circle, Atlanta, GA 30322
Skinner Phone: (404) 783-2534 • <katherine.skinner@emory.edu>
Halbert Phone: (404) 727-2204 • <martin.halbert@emory.edu>
Fax for both: (404) 727-0827 • <http://metaarchive.org>

PROFESSIONAL CAREER AND ACTIVITIES: Dr. Katherine Skinner is the Digital Projects Librarian at **Emory University** and provides leadership for the university's digital projects that are supported through grants or other sponsored funding sources. In this role, she has coordinated efforts involving interdisciplinary interest groups from more than three dozen universities worldwide, including faculty members (in the sciences, social sciences, and humanities), information technologists, librarians, curators, and campus administrators. She is a Co-Principal Investigator on the **SouthComb Project (Andrew W. Mellon Foundation)**, the **MetaArchive Project (Library of Congress)**, and the **MetaArchive: A Sustainable Digital Preservation Service Project (NHPRC)**. She's a founder and an editorial board member of the peer-reviewed Internet journal, *Southern Spaces* and a consultant on the **TransAtlantic Slave Trade Database Projects** at **Emory (NEH)**. She recently co-edited a monograph entitled *Strategies for Sustaining Digital Libraries* with **Martin Halbert**.

Katherine has a Ph.D. from **Emory University** and a B.A. from the **University of North Carolina at Chapel Hill**. **Katherine** worked in academic/non-profit marketing at **Mindpower, Inc.**, prior to starting graduate school in 1999. She is married and has a son.

PROFESSIONAL CAREER AND ACTIVITIES: Dr. Martin Halbert is Director for Digital Innovations at **Emory University**. **Martin** directs a variety of digital library research projects and services 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://www.MetaScholar.org>). The staff of Digital Innovations includes librarians, scholarly communication analysts, and systems programmers/support staff. He is the principal investigator for research projects totaling \$5M. He recently established the **MetaArchive Digital Preservation Network** (<http://www.MetaArchive.org>), a cooperative of ten university libraries and other cultural memory organizations acting in concert with the **Library of Congress** as part of the **National Digital Preservation Program**.

Martin has a Ph.D. from **Emory University**, an M.L.I.S. from the **University of Texas**, and a B.A. from **Rice University**. **Martin** has worked in library administration and systems positions at **Emory University** and **Rice University**. He served as an **ALA/USIA Library Fellow** in Estonia assisting with the automation of the **Tartu University Library**. He has also worked as a consultant for the **IBM Corporation**, and as a programmer for the **University of Texas**. He is married and has three children.

HOW/WHERE DO WE SEE THE INDUSTRY IN FIVE YEARS: We believe that academic libraries are poised at a fairly serious crossroads, particularly now that we are in the throes of a major recession. If librarians choose to function primarily as intermediaries seeking to coordinate the access and preservation functions for our campuses through outsourcing contracts to external entities, the value we add to the academic enterprise will greatly diminish and we will ultimately become a study hall / museum of aging physical media relics. If we alternatively choose to meet the digital information needs of our campuses ourselves, the value of research libraries will grow as vital and experimental arenas of scholarly inquiry and engagement with knowledge. The latter path is less charted and requires more experimentation, but is (we think) clearly preferable as a course for the future. 🌱

minimal storage fees cover the expense of replicated storage space for the network at cost. We believe that it is unlikely that any similar replicated digital preservation service can be established at lower costs.

The **Cooperative** membership structure is comprised of three tiers: Contributing Members, Preservation Members, and Sustaining Members. Contributing Member sites are smaller institutions interested in using the shared network infrastructure to preserve digital content but lacking the capacity to operate any technical infrastructure of their own. Preservation Member sites are responsible for the basic ongoing network activity of preserving digital content. At a minimum, every preservation site must include responsible staff and a minimally configured node server. Sustaining Member sites are responsible for hosting a preservation node and also for leading the **Cooperative** through steering committee participation and through developing the technical systems that enable the preservation network.

Decentralized Preservation Practices

A key strength in our approach to preservation is the distributed nature of both our technical and organizational infrastructures. **MetaArchive** centers on the principle, “lots of copies keep stuff safe.” We believe this to be true, not just in terms of replicating content and distributing it across a geographically dispersed network, but also in terms of replicating *knowledge* and distributing it among our members. To this end, major systems knowledge is not simply held by a central staff, but is deliberately spread out across our member institutions’ technical staff. Our sustainability is increased through this distributed knowledge in several important ways — we are not dependent on central staff members, but rather have shared expertise to draw on across all member institutions; we have a built-in system of checks and balances, as network monitoring is conducted by a committed core of Preservation and Sustaining members; and 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.

In keeping with this philosophy, we also do not rely on the **LOCKSS** team to administer our network. This is a major difference between the **MetaArchive Cooperative** and

other **PLNs**, which have largely opted to have **LOCKSS** manage and maintain their networks. We do *benefit* greatly from the **LOCKSS** team, both in terms of the regular updates they provide to the **LOCKSS** daemon and also in terms of the technical expertise they share with our central and distributed staff members, but we chose to build on an open source framework specifically because we believe this model offers the best odds for long-term sustainability. The overall **LOCKSS** community (including myriad **PLNs**) is already strong and it’s growing stronger. We believe that a solid open source development community could sustain and maintain the **LOCKSS** software if called upon to do so, and we have intentionally built a framework that relies only on this community, not on any one group within it.

Preservation and Institutionalization

Institutions form in order to address specific needs that are not already being met within the existing environment. This is to say that when dominant and traditional business practices (and libraries are a business, whether we think of ourselves as such or not) fail to meet community or market needs, it opens a space within which new institutions with new approaches may flourish. Witness **Google**, **Elsevier’s** journal services, and myriad other examples and exemplars that have already emerged to serve the information management and access needs of the digital age.

The library as an institution continues to serve many of the needs of its constituents — it is not in danger of perishing outright. However, it has not yet proven itself a serious contender in the digital realm. Scholars as well as the public are increasingly turning to companies such as **Google** to “to organize the world’s information and make it universally accessible and useful,” an access role that for centuries belonged primarily to the library field. To whom will these groups turn when they seek to preserve their digital assets, another core mission of the information science field? Will they turn to cultural memory organizations such as libraries and archives, or to corporations such as **Amazon** and **Google**? And should we, as cultural stewards, care so long as the preservation channel adequately provides for the needs of our institutions and our constituents? Is there a difference between commercially driven solutions and those created in the not-for-profit environment?

As libraries, our work is driven by the desire to maximize our stakeholders’ long-term access to materials, not by a desire to maximize profit for stockholders. This is a highly significant

distinguishing factor and one that we cannot afford to take for granted. Cultural memory organizations are not, on the whole, profit-making enterprises. They are funded by tax dollars, foundations, and parent institutions whose constituents we serve. Our reason for being is to serve a public good — making our cultural assets, from books to datasets, accessible to the public over the long haul. If we cease to perform that function, instead outsourcing it to external parties, we are putting both our field and our cultural resources in precarious positions.

This is not to say that we should not outsource *any* of the digital preservation work we undertake. Just as is true in the print and physical artifacts domain, there will doubtless be portions of this work that are well suited to the work of external parties and too expensive to build in house. But, in order to know which parts are suited to in-house work and which parts should be outsourced, we must begin to explicitly engage in our own digital preservation solutions.

So from where will the successful approaches to digital preservation emerge? Sociologists tell us to watch the fringes, not the center, for seismic changes.⁵ Today, one of these fringes might well *be* the library — or in this case, many libraries, banding together in collaborative, cooperative ways to accomplish the preservation of their unique resources in a communally owned network environment that they run for themselves. 🐼

Endnotes

1. **Library of Congress**. 2002. “Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program.” Accessed December 29, 2008, from http://www.digitalpreservation.gov/library/resources/pubs/docs/ndiipp_plan.pdf.
2. Consultative Committee for Space Data Systems. 2002. Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-R-1—Blue Book. Accessed December 28, 2008 from <http://public.ccsds.org/publications/archive/650x0b1.pdf>.
3. **David S. H. Rosenthal, Thomas Robertson, Tom Lipkis, Vicky Reich, and Seth Morabito**. 2005. “Requirements for Digital Preservation Systems: A Bottom-Up Approach.” *D-Lib Magazine* 11, no. 11. Accessed December 29, 2008 from <http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html>.
4. For more on the use of collaboration to achieve library goals, see **Michael Day**. 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*, 1, Vol. 3: 15-28. Accessed December 28, 2008, from <http://www.ijdc.net/ijdc/article/view/60/61>.
5. **Huseyin Leblebici, Gerald R. Salancik, Anne Copay, and Tom King**. 1991. “Institutional Change and the Transformation of Interorganizational Fields: An Organizational History of the U.S. Broadcasting Industry.” *Administrative Science Quarterly* 36: 333-363.

In Denver the exhibits and exhibitors were more upbeat. Though I did learn that **Michele Casalini** had broken his foot climbing uphill on ice and wasn’t able to travel to Denver, the smiling **Colleen Campbell** and the cheerful **Barbara Casalini** were positive about the future. Across from **Casalini**, I visited with **East View Information Services’ Zina Somova** (Director of Operations) and **David**

Ziembiec (Sales & Account Manager, Academic Markets). Watch for our interview coming soon in an issue of **ATG**!

And continuing to speak of Denver, I saw **Marc Lenzini** in person after his airplane ordeal. (see **ATG**, v.20#6, p.1). He is back from visiting daughter **Annie** in Honduras and seemed none the worse for wear. Whew!

Speaking again of Denver, **SAGE Reference** titles have been honored by both the **Reference and User Services Association (RUSA)** and its

continued on page 44