Distributed Digital Preservation: Private LOCKSS Networks as Business, Social, and Technical Frameworks

Victoria Reich, David Rosenthal

Library Trends, Volume 57, Number 3, Winter 2009, pp. 461-475 (Article)

Published by Johns Hopkins University Press
DOI: https://doi.org/10.1353/lib.0.0047

Distributed Digital Preservation: Private LOCKSS Networks as Business, Social, and Technical Frameworks

Victoria Reich and David Rosenthal

Abstract

The Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP) has helped underwrite the development of highly targeted collaborative preservation networks among content sites with common interests in specialized subject areas. Today, institutions worldwide have tapped the LOCKSS (Lots of Copies Keep Stuff Safe) Program's innovative approach to distributed digital preservation to accomplish a variety of business, social, and technical pursuits. LOCKSS technology enables nontechnical administrators to create and manage their own preservation network via a "Private LOCKSS Network" (PLN). A PLN is a scaled down version of the public LOCKSS network, which comprises two hundred library members and preserves over a thousand titles from more than three hundred publishers. PLNs enable like-minded institutions to shoulder the responsibility of preserving in perpetuity scholarly e-content of importance to the group. More and more research organizations are discovering the benefits of PLNs to address their collection and preservation needs.

Background

The Library of Congress has backed several significant collaborative preservation efforts including Stanford University Libraries' LOCKSS (Lots of Copies Keep Stuff Safe) Program. NDIIPP awarded LOCKSS with matching funds to launch a "Private LOCKSS Network" (PLN) called CLOCKSS. PLNs are a low cost, self-managed, distributed method of preserving e-content. Conceptually similar to the popular public LOCKSS network but on a much smaller scale, PLNs typically have seven or more like-minded institutions that work together to collect, preserve and provide perpetual

LIBRARY TRENDS, Vol. 57, No. 3, Winter 2009 ("The Library of Congress National Digital Information Infrastructure and Preservation Program," edited by Patricia Cruse and Beth Sandore), pp. 461–475

access to content of importance to them. The LOCKSS public network comprises two hundred library institutions and preserves content of interest to broad swaths of its membership. PLNs are community-run preservation efforts. The institutions running PLNs are motivated to protect similarly themed collections such as southern U.S. history, and/or content types such as state documents. They are among the best hope our society has to keep specialized materials available for future generations. But to understand PLNs, one must first understand LOCKSS.

LOCKSS

As one of the first employees of HighWire Press, Stanford University Librarian Victoria Reich, was concerned that an unforeseen consequence of the increased adoption of digital journals by libraries was putting their ability to serve society at risk. A library can only fulfill its role as a memory organization when it continues to build local collections and keeps those collections safe for future readers. To address this formidable challenge, Stanford University Libraries started the LOCKSS Program. LOCKSS replicates the traditional model held by libraries of keeping physical copies of books and journals in their collections. Today, ten years later, two hundred libraries worldwide use LOCKSS software to take local custody of the e-materials important to their community. Over three hundred publishers have granted permission to LOCKSS to allow its member libraries to collect, preserve, and provide access to the e-content. The content is automatically ingested and preserved on a basic PC running LOCKSS software called a *LOCKSS Box*. When the content is unavailable from the publisher, libraries seamlessly serve this preserved content to their readers. Hundreds of publishers and libraries have joined the LOCKSS community and are working together to ensure that libraries continue their important social role as memory organizations.

How LOCKSS Works

A library uses LOCKSS software to turn a low-cost PC into a digital preservation appliance called a LOCKSS Box that performs the following four functions:

- It collects content from the target websites using a Web crawler similar to those used by search engines.
- It continually compares the content it has collected with the same content collected by other LOCKSS Boxes, and repairs any differences.
- It acts as a Web proxy or cache, providing browsers in the library's community with access to the publisher's content or the preserved content as appropriate. It can also serve content by metadata (Open URLs) via resolvers.

- It provides a Web-based administrative interface that allows the library staff to target new journals for preservation, monitor the state of the journals being preserved, and control access to the preserved journals.

Before a LOCKSS Box can preserve e-content, the following must happen:

- The publisher has to give permission for the LOCKSS system to collect and preserve the journal. They can do this by adding a page to the journal's website containing a permission statement, and links to the issues of the journal as they are published.
- The LOCKSS Box has to know where to find this page, and how far to follow the chains of Web links. This is accomplished via a LOCKSS Plugin that is, for most content, supplied by the Stanford University LOCKSS team. The Plugin is distributed automatically to authorized LOCKSS Boxes.

LOCKSS Boxes at libraries around the world collect content directly from the publisher's website and then, among themselves, compare the collected content to what is available from the publisher in order to establish the content's authoritative version.

The LOCKSS Boxes then use the Internet to continually audit the content they are preserving. If the content in one LOCKSS Box is damaged or incomplete, that LOCKSS Box will receive repairs from the content based on other LOCKSS Boxes. This cooperation between the LOCKSS Boxes avoids the need to back them up individually. It also provides unambiguous reassurance that the system is performing its function and that the correct content will be available to readers when they try to access it. The more organizations preserve given content, the stronger the guarantee that they will all have continued access to it.

LOCKSS Boxes provide transparent access to the content they preserve. Institutions often run Web proxies, to allow off-campus users to access their journal subscriptions, and Web caches, to reduce the bandwidth cost of providing Web access to their community. Their LOCKSS Box integrates with these systems, intercepting requests from the community's browsers to the journals being preserved. When a request for a page from a preserved journal arrives, it is first forwarded to the publisher. If the publisher returns content, that is what the browser gets. Otherwise the browser gets the preserved copy.

Library staff administers their LOCKSS Box via a Web user interface. The interface enables new content preservation, monitors the preservation of existing content, controls access to the appliance, and a wide variety of other functions. The LOCKSS team at Stanford University provides technical support.

LOCKSS software is based on Association of Computing Machinery (ACM) award-winning technology (ACM, 2004). It provides an OAIS-

compliant, open source, peer-to-peer, decentralized digital preservation infrastructure. It is format-agnostic, preserving all formats and genres of Web-published content, provided the content has an authoritative version. The intellectual content, which includes the historical context (the look and feel), is preserved. Content preserved by libraries in their LOCKSS Box becomes a part of their collection, and they have perpetual access to all of it.

Public LOCKSS Network

The public LOCKSS network comprises nearly two hundred research institutions worldwide. Libraries can join LOCKSS at no charge; however, these members have limited access to publisher content. Libraries serious about building local collections join the LOCKSS Alliance for a nominal annual fee. Alliance members have access to all the content stored in LOCKSS to which they subscribe, along with other materials such as open access titles. Getting started is simple. Libraries download a CD image, burn it to a CD, and use the CD to boot a PC, turning it into a LOCKSS Box that automatically joins the international distributed preservation network. The LOCKSS system requires at least seven instances of any particular piece of content for preservation to be assured. Preservation redundancy is achieved because the content is of general interest. Most of the institutions participating in the public LOCKSS network preserve most of the content available to them, especially the content to which they subscribe. The average replication factor in the public LOCKSS network is approximately forty.

Access to content held in an institution's public LOCKSS Box is "light" to that institution's authorized users. In other words, LOCKSS members are permitted to access the content to which they subscribe whenever that content is unavailable from the publisher. Thousands of titles are available in the public LOCKSS network.

Private LOCKSS Networks (PLN)

A Private LOCKSS Network (PLN) offers institutions with synergistic collections a means to ensure the survival of their highly specialized content. Non-technical staffs are empowered to implement, manage, and govern their own distributed digital preservation network using LOCKSS software. Each PLN has different governance (formal vs. informal) and funding models. The PLNs in transition from soft money (grant funding) to becoming self-sustaining are charging their members nominal fees, and/or their members are contributing "in kind" to cover the overhead of PLN management (policies and practices concerning administration, collection development, etc.). All PLN members support the development and maintenance of the LOCKSS technology via the LOCKSS Alliance.

Private LOCKSS Networks generally preserve content that is outside the collection development scope of most institutions. In other words,

content preserved in a PLN is often more akin to a library's special collections, digitized images, local websites, etc. The partners in any one PLN explicitly agree to share in the preservation of each other's specialized content.

Private LOCKSS Network affiliates usually do not self-select. Current network members need explicit confirmation that a potential member institution will bring value to the network, for example, by contributing appropriate content and having a working LOCKSS Box online. Typically a PLN has seven to twelve institutional participants.

Access to PLN preserved content ranges from light, to dim, to dark. Each PLN community sets its own access policies based on local needs, resources, and the intellectual property rights associated with the content. Most PLNs are dark, however, with content access via a hosting platform such as ContentDM or DSpace.

Each PLN requires some technical administration, such as monitoring the network and maintaining the two databases associated with the ingest process. A community can manage its own technical infrastructure (e.g., the MetaArchive Collective) or the Stanford University-based LOCKSS staff can manage the infrastructure. Most often the LOCKSS staff helps with technical installation, and then a local community accepts as much responsibility as is appropriate and comfortable for them.

One PLN Story—CLOCKSS

The CLOCKSS Pilot Program was conceived and developed by top research libraries and scholarly publishers[1] who wanted to test whether a cooperative, decentralized, community-run dark archive could succeed. The founding libraries and publishers selected LOCKSS as the technology platform because of its proven track record for reliably archiving and preserving scholarly information with little administrative overhead. The pilot, which ended in May 2008, was so successful that its founders unanimously voted to transition the project to a full-scale geographically distributed archive and began accepting new members.

The pilot participants, eleven key industry publishers—who represent 60 percent of online journal content—and seven top academic libraries comprise the CLOCKSS Board of Directors. This board was tasked with developing a new model for archiving and as such spent most of its time debating, and setting policy designed to ensure unencumbered worldwide, long-term access to digital scholarly content.

CLOCKSS leverages the cost-effectiveness of libraries' existing networking, their technical infrastructure, and their historic mission as memory organizations. CLOCKSS Boxes form a PLN that is currently ingesting and preserving what will become comprehensive archival collections. CLOCKSS also taps publishers' technical know-how and experience with platform development and content distribution. As a result, the shared

goals and respective expertise of both groups is the catalyst for a robust, streamlined approach to meeting the preservation needs of the scholarly community.

The archived content is a valuable asset into which scholars, libraries, and publishers have made considerable long-term investments and as such must be protected from a wide variety of possible disruptions whether deliberate or accidental. The CLOCKSS Archive is made up of widely distributed "nodes" and "hosts," which span geographic, political, and legal boundaries. This global network is under the stewardship of the stakeholders who have invested so heavily in these assets.

Libraries and publishers (commercial and nonprofit), working together as equals, govern the CLOCKSS Archive. This unique governance structure is the linchpin that distinguishes this Archive from all others. The CLOCKSS board oversees the CLOCKSS Archive and is ultimately responsible for the safekeeping of its contents. An important strength and appeal of CLOCKSS is the assurance that all CLOCKSS supporters have a voice and an opportunity to advise the archive's governance. Every participating institution in the CLOCKSS Archive has one or more governing roles:

- Governing publishers have been elected to the board.
- Each supporting publisher holds a seat on the advisory board.
- Governing libraries have been elected to the board.
- Archive nodes house CLOCKSS Boxes; if not a board member, each node holds a seat on the advisory council.
- Supporting libraries hold a seat on the advisory council.
- Host institutions serve triggered open access content. If not a board member, each host holds a seat on the advisory council.

The participants are extraordinarily motivated to speak out on behalf of the archive and enact policy consistent with this responsibility. The collaborative nature of CLOCKSS and the working relationships among the directors have enabled consensus on a number of important policies and practices. The community with the most at stake, not a third-party, controls the generational persistence of this important material. Some example policies or practices include:

- CLOCKSS incorporates into its business model an endowment to be raised over a five-year period, which is expected to underwrite 80 percent of ongoing costs.
- CLOCKSS is not a lease or a subscription service.
- CLOCKSS is a global, geographically distributed dark archive.
- Either source files or presentation files are accepted for ingest and preservation.
- The original content, as supplied by the publisher, will be the "content of record for CLOCKSS," and will be preserved exactly as received.

- Triggered content will be made freely available to everyone.
- The community governs the archive.

The board has a strong motivation to keep costs low in order to achieve long-term sustainability. Low preservation costs will decrease resource competition, specifically between preserving today's materials for tomorrow, and acquiring and publishing new intellectual property. CLOCKSS is funded via sliding scale fees for libraries and publishers. Over the long term, the CLOCKSS board intends to raise an endowment to pay for most of the archive's ongoing expenses. Digital preservation requires continuous processes; when active preservation ceases, materials are lost. By building an endowment and becoming self-sustaining, CLOCKSS will ensure that the preservation processes continue over time, regardless of the availability of outside sources of revenue.

The CLOCKSS Pilot Program archive nodes (six in the United States and one in the United Kingdom) form the initial backbone for this comprehensive global preservation archive network. The CLOCKSS archive is now growing to a full network of nine to fifteen nodes. Each node is strategically located to ensure a geopolitical distribution of safely stored content. The CLOCKSS Boxes are at large libraries (or library organizations, such as OCLC) and are ingesting and preserving source files and presentation files, for comprehensive collections, materials to which each archive node institution does, and does not, subscribe. The CLOCKSS processes are continuously audited to ensure accurate and reliable preservation. The CLOCKSS Boxes are maintained at the archive nodes in secure computing environments with uninterrupted power and network connectivity. And, helping justify the decision to leverage existing technology and infrastructure, along with the cost-efficiency of the LOCKSS software, the costs associated with administration and maintenance of these CLOCKSS Boxes are proving to be negligible.

The board worked to establish the archive's policy on trigger events. The members defined what constituted a trigger event (for example, no publisher has current responsibility for, nor is providing electronic access to, selected content) and determined that materials affected by a trigger would be released to everyone, without regard for prior subscription/payment status (CLOCKSS, n.d.).

The process of triggering content from the dark CLOCKSS archive is different from the way content is accessed in the light archive provided by an institution's LOCKSS Box. And for CLOCKSS archived files, the process to make files accessible differs between presentation and source materials. An institution's LOCKSS Box preserves presentation files obtained by crawling the websites of the journals to which it subscribes. The box supplies individual files directly to a reader's browser as it requests them. When presentation file content is triggered from the CLOCKSS archive it

is copied in bulk and moved to a separate Web server. This process had been demonstrated privately to the CLOCKSS board using content from a SAGE Publication's journal published via the HighWire Press platform.

The recent decision by SAGE to discontinue its journal *Graft: Organ and Cell Transplantation*, also published via the HighWire Press platform, provided the CLOCKSS initiative with a public opportunity to demonstrate this process by offering continuing public access to all the SAGE-published volumes (three from 2001 to 2003) of *Graft* that are preserved in presentation file form by the CLOCKSS archive. The presentation files were extracted from the CLOCKSS archive, processed into a form suitable for use by the Apache open-source Web server, and deployed to two host institutions (Stanford University and University of Edinburgh). Both used Apache to make the *Graft* volumes freely available to everyone under a Creative Commons license. As can be seen, the re-published volumes appear identical to their previous incarnation at HighWire Press (CLOCKSS, n.d.). The details of the trigger process and the lessons learned were presented at the Coalition for Network Information (Reich & Rosenthal, 2008).

The process of triggering content that the CLOCKSS archive holds in source form is more complex. As with presentation file content, it happens in three phases. First, the content has to be copied from the archive. Second, it has to be processed into a form suitable for Apache. Third, it has to be deployed to re-publication sites. The first and third phases are identical to those for source form content; it is the processing that is more complex and publisher-specific. In general, it consists of rendering XML into HTML by applying suitable styles. The entire process had been demonstrated to the CLOCKSS board using Elsevier content.

A subsequent decision by SAGE to discontinue its journal *Auto/Biography*, which had been published on the Ingenta platform and provided to the CLOCKSS team in source file form, provided a similar public opportunity to demonstrate that the CLOCKSS archive could trigger source file content. The results are posted at two host institutions (Stanford University and the University of Edinburgh) (CLOCKSS n.d.). Note that the volumes look different from their previous incarnation at Ingenta. This is an inevitable result of preservation in source form and re-rendering.

Contributions

The CLOCKSS program has led to improvements and enhancements in LOCKSS technology that have benefited the general LOCKSS community in multiple ways. In addition to continuous development of the open source LOCKSS software, the CLOCKSS program specifically supported the following new features.

Enhanced Security for Closed PLNs

In the nature of an open, worldwide peer-to-peer network such as the public LOCKSS network, an individual site cannot place a high level of trust in any other individual site. Trust reposes in the consensus of a large number of other sites, mediated by the technology. PLNs, by contrast, have a much smaller number of sites but their trust in them is much higher and is mediated by social structures outside the technology itself. The CLOCKSS PLN pioneered the use of digital certificates to enable mutual authentication of peers in a PLN via mechanisms embedded in the secure socket layer of Web communication. This technology is being rolled out to other PLNs.

Handling Source File Content

Several of the larger publishers participating in the CLOCKSS program preferred to contribute their content in source file form, using the same process they use with other institutions maintaining local copies (for example, OhioLink, University of Toronto, etc.), rather than have the CLOCKSS Boxes crawl their websites to collect presentation files. The CLOCKSS team therefore built an ingest pipeline for this purpose that used FTP to collect daily packages of content from the publishers, verified, unpacked, and added them to a private Web server from which the CLOCKSS Boxes could collect them. Experiments with the feed from Elsevier showed that even a relatively small PC could perform these tasks at five times the rate Elsevier publishes content. Special kudos in this respect are due to Elsevier's technical team. Their secure distributed operating system specification is a model of how to deliver content as source files, meticulously documented and providing extensive integrity checks (Mostert, 2006). Even though the initial two-year CLOCKSS program was primarily devoted to organizational and social issues, it resulted in some significant changes to the technical requirements. To satisfy these, some evolution of the source file support is under way. The results will be made available for use by other PLNs.

The Library of Congress's NDIIPP support for the LOCKSS program has also contributed technology to the general LOCKSS community.

Support for ARC Files

The Internet Archive's ARC file format, and its enhanced WARC version, is becoming the standard for the interchange of collections of Web pages between Web crawlers and archives. With help from the Internet Archive, the LOCKSS team designed, developed, and demonstrated the ability to:

- Extract preserved content from a LOCKSS Box in ARC file form using a specialized configuration of the box and the Internet Archive's Heritrix Web crawler. This allowed the Heritrix crawl's output ARC files to

appear as if the content had been ingested directly from the original publisher, although in fact it had all come from the LOCKSS Box.
- Ingest content from a Heritrix crawl directly into a LOCKSS Box. The box unpacks the individual files from the ARC files and makes it appear as if they had been ingested directly from the original publisher.

Both technologies are now available to PLNs that wish to use them. This means, in OAIS terms, that LOCKSS boxes can use the ARC format for both their Submission Information Packages and Dissemination Information Packages. Further adoption of the WARC standard is now under way; a forthcoming revision of the repository in which LOCKSS boxes store their preserved content uses WARC as the format for the stored Archival Information Packages.

Solaris Support

Most libraries run their LOCKSS Box using the specially configured "network appliance" based on the OpenBSD operating system that the LOCKSS team supplies (Rosenthal, 2003). The NDIIPP-funded MetaArchive PLN (http://www.metaarchive.org) uses Linux, support for which was developed under LOCKSS alliance auspices. Much of the Library of Congress' infrastructure uses the Solaris operating system from Sun Microsystems. With help from the Library of Congress, the LOCKSS team ported the Linux version of the LOCKSS daemon to Solaris 9 and integrated Solaris 9 support into the regular six-weekly daemon release process. Based on this and with help from Penn State, the LOCKSS team added support for operating as a zone under Solaris 10. Both capabilities are now available to LOCKSS sites that would prefer to use Solaris.

Monitoring and Measuring Preservation Performance

Initially, the tools the LOCKSS team could use to monitor and measure how well the LOCKSS Boxes were working to preserve their content were oriented to debugging by developers. As the network and the content grew, they became unusable except for detailed debugging and were replaced by a first generation tool capable of looking at the network of boxes from a management point of view. As growth continued, this too became unusable, partly because it didn't scale up gracefully, and partly because security restrictions at some participating libraries meant that their boxes could be monitored only indirectly, by inferring information from their communication with boxes at libraries that did permit direct monitoring of their box.

It became apparent that a tool was needed that could collect and aggregate information from all the boxes in a network in order to

- compute and present statistics about the performance of the network as a whole;

- identify boxes that are outliers in performance terms and diagnose their problems;
- identify content that is not being preserved adequately and diagnose its problems.

In addition, the tool needed to be able to scale up to handle millions of instances of archival units (for example, journal volumes) in the boxes of the public network, while being easy to use for administrators of small PLNs. These typically have a relatively small number of archival units, albeit ones with many more bytes than the normal journal volume.

The LOCKSS team has developed a tool to automate the collection of status information from all the boxes in a network that allow direct monitoring, and store it in a MySQL database. A Ruby on Rails interface allows this database to be queried to present the latest available data from individual machines, and statistical summaries of current and historical data. The LOCKSS and MetaArchive teams have worked together to transfer the tool to the MetaArchive and assess its usefulness in the PLN context. Lessons learned in this process are being applied before the tool is made generally available.

Sustainability

Unfortunately, digital preservation cannot be accomplished in fits and starts; it must be protected from uncertain funding cycles. Keeping digital content static is an active process; the content must be continuously audited, repaired, and preserved if it is to remain accessible. Digital preservation is extremely vulnerable to funding disruptions; when money becomes scarce, other, more immediate and tangible needs take priority (*LC Hit by $47 Million Cut*, 2007). Thus the most intractable issue facing all digital preservation solutions is sustainability. This is vividly illustrated by the following recent developments. Repercussions from tight budgets caused the rescission of the Library of Congress NDIPP funds and cancellation of the U.K. Arts and Humanities Data Service (AHDS; http://ahds.ac.uk/).

Recognizing economic sustainability as key, the recently formed Task Force On Economic Sustainability of Digital Data is "charged with developing a comprehensive analysis of current issues, and actionable recommendations for the future to catalyze the development of sustainable resource strategies for the reliable preservation of digital information" (OCLC, 2007).

The LOCKSS program recognized from the beginning that "not enough money" is a key threat to digital preservation. Library budgets are, and have been for decades, under extreme pressure. Money spent on digital preservation is less money available for other important tasks, such as acquiring new materials. Digital preservation competes with the core

library functions of building a collection and providing services around that collection.

The LOCKSS program started in 1998 with a small grant from NSF (http://www.nsf.org). Development from the prototype that resulted, and initial deployment to a group of partner libraries was funded by major grants from NSF and the Andrew W. Mellon Foundation (http://www.mellon.org), together with support from Sun Microsystems (http://www.sun.com/), HP Labs (http://hpl.hp.com/), Intel Research (http://techresearch.intel.com/articles/index.html), and others. Since mid-2004, libraries participating in public and private LOCKSS networks have been asked to pay fees on a sliding scale to the LOCKSS alliance. In return they receive support from the LOCKSS team. All participants may preserve content whose release to the public network was funded by grants. Only alliance members may preserve content whose release was funded by the LOCKSS alliance. Institutions may only preserve content to which their institutions have authorized access (via subscription and/or open access).

The LOCKSS program's most recent grant was from the Andrew W. Mellon Foundation, the grant's purpose was to help the LOCKSS program transition from grant funding and become self-supporting. By 2007, the transition was complete. In that year, development of the technology, and support of public and private LOCKSS networks, accounted for two-thirds of the LOCKSS team's activities, and the LOCKSS alliance received fees amounting to two-thirds of the team's expenditures. The remaining one-third was accounted for by activities particular to the CLOCKSS program, funded by the CLOCKSS participants, and by activities on behalf of the Library of Congress, funded by NDIIPP. In other words, the public and private LOCKSS networks successfully reached the sustainability benchmark by being free of grant funding and depending entirely on their participants.

Enabling Sustainability

This successful transition to sustainability was made possible by the generous support of the Mellon Foundation, and adherence to three critical, synergistic features of the LOCKSS approach. These are to focus on:

- reducing costs;
- reusing existing technology, infrastructure, and institutions;
- involving the broadest possible range of libraries in the preservation network.

In the LOCKSS system, ingest, preservation, and dissemination are all highly automated, minimizing the staff time required of participating libraries. The system does not require expensive, enterprise-scale technology; it works well with low-cost consumer technology using the replication

and cooperation inherent in the preservation network to provide reliability. Since each LOCKSS Box serves a limited community, there is no need to re-create expensive, high-volume publishing platforms. Very little of the technology is new; the system is mostly a re-packaging of existing Web crawler, Web proxy, Web server, and peer-to-peer technologies. The system is entirely open source, both in terms of the components it reuses and the new technology the team developed to re-package them.

The system is designed for use by existing libraries leveraging their existing relationships with publishers. LOCKSS makes it economically and technically possible for even relatively small libraries to actively preserve their own content. They need not outsource their custodial role to a third-party service provider. The costs of development and support are spread widely, reducing the impact on individual sites, and improving sustainability by diffusing the impact of individual funding decisions. Note that the LOCKSS alliance fees have not increased for four years.

The LOCKSS approach to format obsolescence, detailed in a D-Lib paper, also minimizes costs by re-using existing technologies (Rosenthal, Lipkis, Robertson, & Morabito, 2005). As a recent Library of Congress report points out, Web formats become obsolete when the majority of browsers no longer render that format:

> If a format is widely adopted, it is less likely to become obsolete rapidly, and tools for migration and emulation are more likely to emerge from industry without specific investment by archival institutions. . . . Evidence of wide adoption of a digital format includes bundling of tools with personal computers, native support in Web browsers. (2007)

The LOCKSS approach stores only the original bits, avoiding the storage explosion required by preemptive format migration. It postpones migration until it is needed, when a reader is using a browser that cannot render the original bits. This approach exploits the time value of money and allows each reader to see the result of the state-of-the-art in migration at the time of his or her access. It leverages the pervasive adoption of open source browser technology for format migration and object rendering by providing a framework in which both open and closed source format converters can be deployed as they become available. As the report points out, converters for Web formats are unlikely to be needed and, if they are, they should not require much, if any, investment by the library community.

Conclusion

LOCKSS technology has proven to be fertile ground for collaborative distributed preservation networks. Its low barrier to entry and simplicity in use has allowed institutions to come together and focus on the organizational and collection development issues rather than the details of preservation technology. Although it is still early days, several business, social,

and technical models for these networks are emerging. The low cost of the technology and the way it leverages existing organizational structures are instrumental in helping these networks achieve sustainability.

Private LOCKSS Networks Affiliated with the Library of Congress NDIIPP

- PeDALS is preserving state public records. Participants are Arizona State Library, Archives, and Public Records, Florida State Archives, New York State Archives, Wisconsin Historical Society. http://rpm.lib.az.us/pedals/, http://www.digitalpreservation.gov/partners/states_az/states_az.html.
- Data-PASS is preserving Social Science data. Participants are Inter-University Consortium for Political and Social Research (ICPSR), University of Michigan; Roper Center for Public Opinion Research, University of Connecticut; Howard W. Odum Institute, University of North Carolina-Chapel Hill; Henry A. Murray Research Archive; NARA; Harvard-MIT Data Center http://www.icpsr.umich.edu/DATAPASS/, http://www.digitalpreservation.gov/partners/datapass/datapass.html.
- MetaArchive Cooperative is preserving Southern Culture. Participants are Emory University, Georgia Tech, Virginia Tech, Florida State University, Auburn University and the University of Louisville http://www.metaarchive.org/, http://www.digitalpreservation.gov/partners/metaarchive/metaarchive.html.
- CLOCKSS is preserving scholarly materials. Participants are Academic publishers and libraries who are cooperatively governing the archive www.clockss.org.

Notes

1. Top research libraries and scholarly publishers include:

Founding Publishers:	*Founding Libraries:*
Indiana University	
American Medical Association	New York Public Library
American Physiological Society	OCLC
Elsevier	Rice University
IOP Publishing	Stanford University
Nature Publishing Group	University of Edinburgh
Oxford University Press	University of Virginia
SAGE Publications	
Springer	
Taylor & Francis	
Wiley—Blackwell	

References

Association for Computing Machinery. (2004, January). SOSP "03 Award Winners" Abstracts. Retrieved June 23, 2008, from http://membernet.acm.org/public/membernet/general.January.2004.cfm?general=4&CFID=15151515&CFTOKEN=6184618

CLOCKSS. (n.d.). *Auto/Biography.* Retrieved June 23, 2008, from http://www.clockss.org/clockss/Auto/Biography

CLOCKSS. (n.d.). *Graft.* Retrieved June 23, 2008 from http://www.clockss.org/clockss/Graft

CLOCKSS. (n.d.). Triggered content. Retrieved June 23, 2008 from http://www.clockss.org/clockss/Triggered_Content

LC Hit By $47 Million Cut in Digital Preservation Funds. (2007, March 20). Library Journal. Retrieved June 23, 2008 from http://www.libraryjournal.com/article/CA6426077.html

Library of Congress. National Digital Information Infrastructure and Information Preservation Program. (2007, July). Sustainability for digital formats: Planning for Library of Congress collections. Retrieved June 23, 2008, from http://www.digitalpreservation.gov/formats/sustain/sustain.shtml

Mostert, P. (2006, March) Specifications for network delivery of datasets. Retrieved June 23, 2008, from http://www.sciencedirect.info/techsupport/sdos/netsdos15.pdf

OCLC. (2007, September 19). Panel to address economic sustainability of digital preservation. Retrieved June 23, 2008, from http://www.oclc.org/news/releases/200673.htm

Reich, V., & Rosenthal, D. (2008, Spring). Dark archive to open access: A CLOCKSS trigger event, project briefing, task force meeting coalition for networked information. Retrieved June 23, 2008, from http://www.cni.org/tfms/2008a.spring/abstracts/PB-dark-reich.html

Rosenthal, D. S. H. (2003, September). A digital preservation network appliance based on OpenBSD. Paper presented at the BSDCon 2003, San Mateo, CA, USA.

Rosenthal, D. S. H., Lipkis, T., Robertson, T., & Morabito, S. (January 2005). Transparent format migration of preserved Web content. *D-Lib Magazine, 11*(1), Retrieved June 23, 2008, from http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html

---

Victoria Reich is director and co-founder of the LOCKSS Program (http://www.lockss.org), Stanford University Libraries, and a founding member of the CLOCKSS Program (http://www.clockss.org). Prior to LOCKSS and CLOCKSS, Victoria helped launch Stanford University's HighWire Press, an online platform for the world's leading journals. Victoria also has extensive library experience in both public and technical services, having held positions at Stanford University Libraries, the National Agricultural Library, the Library of Congress, and the University of Michigan. A list of her publications and presentations can be found at: http://www.lockss.org/lockss/Vicky_Reich.

David Rosenthal is chief scientist and co-founder of the LOCKSS Program (http://www.lockss.org), Stanford University Libraries, and a founding member of the CLOCKSS Program (http://www.clockss.org). David joined Sun Microsystems in 1985 from the Andrew Project at Carnegie-Mellon University, departing in 1993 to become Chief Scientist and employee #4 at Nvidia, the leading supplier of high-performance graphics chips for the PC industry. In 1996 he joined Vitria Technology, a supplier of e-business infrastructure technology. After starting the LOCKSS Program at Stanford University, he spent 1999-2002 developing it at Sun Labs. From 2002 he has been working on LOCKSS at Stanford University Libraries. David received his MA degree from Trinity College, Cambridge, and PhD from Imperial College, London. He authors technical publications, holds twenty-three patents, and writes a blog http://blog.dshr.org/ about digital preservation.