Building and Preserving Library Digital Collections Through Community Collaboration

Victoria Reich
Executive Director, LOCKSS Program, Stanford University Libraries

Abstract
Traditionally, libraries in the paper world were responsible for the safekeeping of much of society’s cultural memory. They did this by building local collections, primarily to provide current local readers with rapid, convenient access to content. However, an important side effect of these local collections was a robust system for preserving the printed, and paper-based record for future readers. There were many copies of the content, on a durable, tamper-evident medium, under many separate administrations, with catalogue systems capable of directing a reader to a nearby copy. Over the last two decades, the transition from paper-based publishing to Web publishing changed models for access and preservation for libraries. Instead of purchasing a copy, libraries now lease access to the publisher's copy. This has made preserving the contemporary published record more difficult, in these ways: Unlike paper, the publisher maintains the only legal long-term copies, and the publisher can alter or destroy them at will; Copyright law makes it problematic, if not impossible, to maintain other copies without specific permission from the publisher; Unlike paper, the copies are on a fragile, easily alterable medium with a short service life (magnetic disk or tape); Unlike paper, the format of the copies is expected to become rapidly obsolete. Preparations for this eventuality are expensive. The paper reviews these challenges, describing how the collaboration between libraries underlying the LOCKSS program enables each challenge to be addressed successfully.

Author
Victoria Reich is Executive Director of the LOCKSS Program, Stanford University Library, [www.lockss.org]. LOCKSS empowers libraries to locally archive and control their collections, ensuring perpetual access, local control and ownership, and the health and strength of the library community. Over 500 publishers participate in LOCKSS. Victoria helped to launch the CLOCKSS Archive and HighWire Press. She has extensive library experience, having held positions at Stanford University Libraries, the National Agricultural Library, the Library of Congress, and the University of Michigan. Victoria is on several Advisory Boards, including “Stanford Copyright & Fair Use”. She received the 2008 Ulrich’s Serials Librarianship Award.
1. Print Libraries as Preservation Systems

“…the preservation of information resources is central to libraries and librarianship,” and is a core value of our profession (ALA 2012).

In the paper world libraries are responsible for the safekeeping of much of society’s cultural memory. They do this by building local collections, primarily to provide current local readers with rapid, convenient access to content. An important side effect of these local collections is a robust system for preserving the printed and paper-based record for future readers (Rosenthal 2005).

We tend to think of paper libraries as individual, stand-alone institutions. But if we think of the world's paper libraries as a network or distributed system of libraries, a different picture emerges. Viewed as a global system, libraries have attributes critical to their mission of preserving cultural memory that are not possessed by individual libraries. The important attributes of these paper library systems are:

- Highly replicated works have a better chance of surviving over many generations. The number of copies held by libraries reflects the importance of a work to various communities. More libraries hold the most important works, or most popular works implementing the meme; “lots of copies keep stuff safe”.
- Copies in libraries are distributed. The content is held by many libraries around the world, each under separate administrations, in countries with different laws, norms and customs. Some of these copies might be damaged by hurricanes, others by flood, earthquakes or fire. Others might be destroyed due to censorship, or accident. It is unlikely that all or even most copies of important works will be entirely destroyed, lost, or de-accessioned.
- Paper is a "tamper-evident" medium. It takes great skill to alter a printed book or journal in a way that does not provide obvious visual cues to the reader. Organizing personal visits to all the world's libraries to alter or destroy some book or journal article is difficult and expensive.
- Paper is a durable medium (Library of Congress 2012). Cotton papers and alkaline papers can last indefinitely.
- Paper libraries have local custody and control of their materials. Access is not dependent on continued payment and for as long as a library holds the title; their local clientele is guaranteed perpetual access. The purpose of preservation is to ensure content can be accessed.
- Libraries purchase and receive exactly what the publisher published. This is particularly true of content that’s considered part of a library’s “core collections”.

Despite the long and productive history of systematic cooperation among libraries, (an example is inter-library loan), the attributes described above that make the paper library system robust and effective in its role of preserving cultural memory were not deliberately created. Libraries did not have policies that mandated they cooperate to hold many replicas of publisher’s content, in a highly distributed network, on a durable medium.

A primary motivation for paper library collections was to provide easy access for readers; an important business outcome of this practice is that paper libraries keep what they buy. The resulting robust preservation system of replicated copies was a fortunate accident enabling them to fulfil their charge to preserve access to the scholarly record for future generations.
2. Moving Safely from Print to Web

Over the last two decades, libraries have transitioned from acquiring paper-based collections to primarily leasing access to web-published ejournals and ebooks. The shift from owning paper collections to renting digital collections has had several disturbing unintended consequences. An unfortunate, simplistic implementation of web technology put libraries in a precarious and weakened position. Most libraries in the web environment are not fulfilling a traditional core social value; preserving their core collections for their future readers.

The move from paper to web has made preserving the contemporary published record more difficult in these ways:

- Most content is held by centralized organizations with limited redundancy and geo-political diversity.
- Copyright law makes it problematic, if not impossible, to maintain copies without specific permission from the publisher.
- Unlike paper, it is easy to alter digital copies with no trace or evidence.
- Unlike paper, the digital copies are on fragile, easily alterable medium with a short service life (magnetic disk or tape).
- Unlike paper, digital file formats are expected to become rapidly obsolete and many approaches to prepare for this eventuality are expensive (Kejser 2011).

3. Restoring Library Collection Infrastructure

The LOCKSS program (Lots Of Copies Keep Stuff Safe) a trademark of Stanford University, shows that a collaborative approach to the economic, legal, technical and social challenges of building and preserving library collections of digital content is both practical and economically sustainable. The program, founded in 1998, has low overhead and flexible commitments. The “Red Hat” model of free, open source software and paid support from libraries worldwide that use the system has kept the program in the black for more than five years with no grant funding.

In the web environment, most libraries rent access to electronic content. Leasing access to content puts libraries at risk; it outsources their duties as information stewards to third parties. A unilateral change of policy by the publisher or third party provider, or a failure to renew a subscription, can result in loss of electronic access to past material.

The LOCKSS Program, led and supported by libraries, gives libraries a way to preserve and control their own collections. Librarians use their local LOCKSS box to take custody of subscription and open access ejournals, ebooks, digitized collections, government documents, etc. A LOCKSS box is a library’s “digital stacks”, analogous to a library using its own buildings, shelves and staff to obtain, preserve and provide access to paper materials. The LOCKSS model restores libraries’ ability to build local collections, bringing the traditional purchase-and-own model to electronic materials and strengthening the library’s role in the digital age.

When libraries take custody of content to which they subscribe, access is separated from payment and perpetual access assured. Material stored in a local LOCKSS box remains available to members of the library’s local community even when the publisher goes away due to merger, bankruptcy, subscription cancellation, network outage or for any other reason. The content is always available to the local
community, directly from the library, with no need to rely upon third parties. Locally owned collections guarantee 100% post cancellation access.

Copyright law makes it problematic, if not impossible, to maintain copies without specific permission from the publisher. Obtaining permission would be prohibitively expensive for libraries to do individually. When a publisher gives permission for the LOCKSS program to preserve their content, they give permission for all libraries that have authorized access to the content. There are many examples of a single library negotiating with a publisher for LOCKSS preservation, to the benefit of all libraries in the LOCKSS network. As with paper libraries, the collaboration is implicit, not explicit.

Paper library collections consist primarily of materials received directly from the publisher. The libraries are providing to the readers what the publisher published. LOCKSS preserves what the readers sees on the web. The LOCKSS preservation approach is unique in delivering the publisher’s original article, with the publisher’s original branding and imprimatur. LOCKSS technology preserves the publisher’s original content as of the date of web publication. The “look and feel” of the content, along with the publisher’s branding, is preserved, resulting in an authentic representation of the authoritative source file. Readers see and use the most trusted, authoritative and contextually accurate version of important cultural materials.

Because the permission to collect and preserve the content is automatically granted to all authorized libraries, there are many libraries to share the cost of the technical work to get the content from the publisher. Again, as with paper libraries, the collaboration is implicit, not explicit.

The LOCKSS Program approach re-implements the key attributes of the paper library system, making it robust and effective at long-term preservation of our cultural heritage. A bit of cooperation among libraries that choose to take leadership and control has transformed seemingly intractable problems into successful solutions. Libraries take custody of content; they maintain local control of collections and assets. They keep what they buy and fulfil their duty to preserve access to the scholarly record for future generations.

4. Guarding Against Known Threats

In the paper world, many libraries each get a copy of a book. It is hard for a book to be accidentally or purposely destroyed because these many libraries are geographically dispersed, they are operated by many different administrations, under diverse political and cultural conditions. A primary motivation for paper library collections is to provide easy access for readers, however an important business outcome of this practice is that paper libraries keep what they buy. When considered as a network, these independent paper library collections resist damage; there is no central point of failure. Paper libraries fulfil their charge to preserve access to the scholarly record for future generations.

For digital materials, evidence suggests that human factors (intentional and unintentional) are the greatest cause of loss or corruption (Rosenthal 2011). Technology failures, economic failures and social failures also pose threats to the protection of digital content. When content is held in a single centralized repository, it is easy for someone to tamper with a master copy without detection. A dark archive, into which content disappears, only to reappear in a future emergency, does not engender confidence in either its availability or its correctness.

The LOCKSS network has similar tamper evident properties to print libraries. In the LOCKSS system, it is very difficult and expensive for someone to find and tamper with a significant number of the preserved copies without being caught. The copies are geographically distributed and independently held
under many different administrations. Tamper-evidence engineering is a unique property of LOCKSS preservation and it is a keystone of our work (Rosenthal et al. 2005).

The LOCKSS Program is the only preservation approach that mitigates against the broad set of technical, economic and social threats to the security and long-term preservation of digital content. LOCKSS’ ACM award-winning open-source technology is built on a peer-to-peer software infrastructure (Maniatis 2003) (Stanford 2012c).

It is unwise to trust closed source preservation systems; closed source systems are written and designed by employees operating under nondisclosure agreements. This constricted review limits opportunities to reveal software weaknesses and vulnerabilities that may interfere with preservation processes. Open source software supports cooperative, collaborative preservation systems in the following ways:

- Open source software is in itself a collaborative enterprise.
- Open source software is itself much better preserved, and much less likely to go obsolete, than closed-source software among other reasons because its copyright license terms encourage "lots of copies keep stuff safe" (Rosenthal 2009).
- Collaboration requires a degree of mutual trust, which is hard to sustain if no one can be sure what software anyone else is running. Open source software is transparent, in that anyone can read and experiment with the code.
- Open source promotes the use of open standards, and thus interoperability.
- Open source prevents vendor lock-in and thus contributes to keeping the cost of preservation as low as possible.

The LOCKSS peer-to-peer infrastructure requires that each library’s LOCKSS box cooperate with other LOCKSS boxes in the system to determine if content being preserved is what was originally collected from the publisher. If there are differences, the system repairs the content back to the authoritative version. The LOCKSS preservation processes employ implicit community collaboration.

The Blue Ribbon Task Force on Sustainable Digital Preservation and Access identified economics as the major threat to preservation (Blue Ribbon 2010). Even with optimistic projections of future costs, there is not enough money to preserve everything that should be preserved (Rosenthal 2012). This places particular importance on minimizing the cost of digital preservation systems, which has been a major focus of the LOCKSS Program since it was founded in 1998. Every unnecessary dollar of cost means more content that will be lost.

The community collaborates and shares responsibility for reaching out to publishers. The software is open source and has been contributed to by many in the community over time.

Multiple copies of content at different libraries audit each other and repair any damaged or missing content, eliminating the need for costly back-ups and manual auditing processes. The total system cost never appears on any one single budget and is thus never at risk from a single red pencil.
5. Cooperating for Robust Infrastructure

The LOCKSS Program builds your institution’s preservation infrastructure by re-implementing characteristics of the paper library system that enable libraries to keep society’s cultural memory safe in the digital environment. Librarians use the LOCKSS system to preserve content and to build their library’s local collections in two distinct networking environments, the Global LOCKSS Network and Private LOCKSS Networks. Local collections give libraries control over perpetual access for current readers and provide a robust preservation approach for future readers.

Librarians participating in the Global LOCKSS Network build and preserve open access titles and subscription ejournals and ebooks from over 520 participating publishers (Stanford 2012a). The Global LOCKSS Network provides the digital equivalent to a library’s general collections. (As with paper collections the publishers hold the intellectual property or copyright of these materials and libraries take custody of the materials for long-term safekeeping).

The Global LOCKSS Network leverages implicit cooperation among libraries worldwide. Sufficient replication is ensured because the materials preserved in the public network are those that the wider community has agreed they wish to preserve. The business relationship between the library and the publisher in the Global LOCKSS Network mirrors the print environment.

Private LOCKSS Networks are the digital equivalent of a library’s special collections. The vast range of content genres preserved in Private LOCKSS Networks includes photo image collections, audio collections, government documents and databases. (For the most part, the libraries own the Intellectual Property or the copyright of the materials preserved in Private LOCKSS Networks).

Private LOCKSS Networks leverage explicit cooperation among a community of libraries that have common interests in specialized subject areas. These networks are highly targeted collaborative efforts among like-minded institutions sharing the preservation responsibility (including governance and sustainability) of e-content important to their institutions. Below are brief descriptions of several Private LOCKSS Networks, offered here as examples:

• Alabama Digital Preservation Network: Alabama libraries are collaborating to preserve a wide variety of historic archival materials, including image collections and databases (ADPN 2012).
• CLOCKSS Archive: Subscription and open access books, journals, and data are preserved in a network, spanning Europe, Asia and North America. When preserved content is no longer available from a publisher it is copied from the CLOCKSS Archive and made available for free, to everyone (CLOCKSS 2012).
• Council of Prairie and Pacific University Libraries: Consortium Canadian University libraries are collaborating to preserve collections important to the provinces of British Columbia, Alberta, Saskatchewan and Manitoba. This group has a particular focus on freely available born digital Web content including government documents, e-journals and small presses (COPPUL 2012).
• Data Preservation Alliance for the Social Sciences: Universities are collaborating to preserve social science data which include: opinion polls, voting records, surveys on family growth and income, social network data, government statistics and indices, and GIS data measuring human activity (Data-Pass 2012).
• MetaArchive: Cooperative Run by the non-profit Educopia Institute, this international membership organization coordinates cultural memory organizations that are collaborating to preserve very high value locally created digital materials (MetaArchive 2012).

6. Building Distributed Digital Preservation Networks

The LOCKSS system was the world’s first production quality Distributed Digital Preservation Network. Distributed Digital Preservation Networks are designed with intentional geographical and organizational distributed infrastructures as essential components of their preservation models. The LOCKSS Program set best practices for reliable preservation and persistent access to digital content via content replication, geographic distribution, infrastructure heterogeneity, modularity and organizational diversity.

Distributed Digital Preservation networks require community collaboration and cooperation. It is not possible to achieve the required robustness to ensure the long-term persistence of digital objects through centralized technical or organizational approaches. Beware organizations that claim distributed digital preservation infrastructure when their technical implementation is a few back-up copies around the globe.

Implementing the LOCKSS Distributed Digital Preservation approach requires three actions: a publisher to give permission for the target content to be preserved; for a library to bring online a LOCKSS box that has authorized access to the content; and for that LOCKSS box to be registered with one of a number of associated LOCKSS Alliance networks (Stanford 2012b).

Publishers grant the LOCKSS system legal permission to ingest, preserve and access their intellectual content by putting online a LOCKSS permission statement. The LOCKSS permission statement is bundled with the content so that the content and the legal rights are preserved together. Paper contracts are hard to track through time, preserving the legal agreement with the contract minimizes any future misunderstandings.

The Global LOCKSS Network implements an extremely effective form of cooperative collection development. When one library or library group secures a publisher’s agreement to participate in LOCKSS, all LOCKSS libraries with authoritative access to that content have permission to locally ingest and preserve the material.

A library uses the freely available, open source LOCKSS software to turn a mid-range PC, or virtual computing environment into a digital preservation appliance called a LOCKSS box (Stanford 2012).

A LOCKSS box performs five main functions:

• It **ingests** content from target websites using a web crawler similar to those used by search engines.
• It **preserves** content by continually comparing the content it has collected with the same content collected by other LOCKSS boxes, and repairing any differences.
• It **delivers** authoritative content to readers by acting as a web proxy, cache or via Metadata resolvers when the publisher’s website is not available.
• It provides **management** through a web interface that allows librarians to select new content for preservation, monitor the content being preserved and control access to the preserved content.
• It dynamically **migrates content** to new formats as needed for display.
LOCKSS Program staff at Stanford University analyses the target content’s URL structure, file formats and delivery mechanisms. They design, implement and update a tailored, content-specific preservation action plan that serves publishers, librarians and readers.

The publisher permits the LOCKSS system to collect, preserve and provide access to the content by putting a LOCKSS manifest page on the content’s website (Stanford 2012d). The manifest page contains a LOCKSS permission statement and links to the issues (or other parts) of the content as they are published. The required manifest page is ingested and preserved with the original content and negates the need for paper contracts.

Software called a LOCKSS plugin tells each institution’s LOCKSS box where to find the publisher’s LOCKSS manifest page, and how far to follow the chains of web links. A LOCKSS plugin encapsulates a publisher’s content model by listing parameters specific to each publishing platform. The LOCKSS team builds, tests and distributes plugins to LOCKSS boxes registered with the LOCKSS Alliance.

Every LOCKSS box is located at an IP address that falls within its parent University’s IP address range. Authorized LOCKSS boxes independently collect subscription or open access content directly from the publisher’s website. The publisher authorizes or denies a LOCKSS box’s access to content through their access control system. Publishers register LOCKSS activity on their web logs and have access to real time statistics through their own systems.

Once ingest is complete, the LOCKSS technology ensures that each LOCKSS box has collected all intended content, thus preserving the authoritative version. The LOCKSS software continually monitors the content in each LOCKSS box to ensure it is properly preserved though a cooperative preservation process that compares one LOCKSS box’s content with the same content on other LOCKSS boxes. When content is damaged or lost the system arranges for content repair from another LOCKSS box.

The administrator of each LOCKSS box can monitor the preservation status of the content in their box, by looking at delivered content and the management tools available through the LOCKSS box web administrative interface (Stanford 2012c).

7. Providing Continuous Access

An institution’s LOCKSS box can provide readers with continual, seamless access to branded publisher content (Stanford 2012f). The LOCKSS system preserves content at its’ original URL, critically retaining the content’s relationship to other web resources. An institution’s LOCKSS box delivers content to authorized readers only when the publisher’s website is unavailable (subscription cancelled, network traffic, publisher server down). The LOCKSS Program works to preserve and to deliver the publisher’s original artefact to readers, in other words - what the publisher published.

LOCKSS boxes provide four main ways for readers to access the content they preserve: by proxying (i.e. acting like a web cache), by serving (acting like a web server) or by serving through integration with an OpenURL resolver, or the emerging Memento standard.

- Proxying: Institutions often run web proxies to allow off-campus users to access subscription content. Libraries that integrate their LOCKSS box into a proxy (PAC Files, EZ Proxy, ICP, Squid) ensure a reader’s URL request is seamlessly fulfilled when the content is unavailable from the publisher’s website.
Basic Serving: In the basic serving model, articles are accessed using a local URL pointing to the LOCKSS box. The LOCKSS box checks if the publisher will provide content to fulfil a reader’s request. If the content is not available from the publisher, the LOCKSS box serves its own copy to the reader.

OpenURL Serving: Libraries can integrate their LOCKSS box with their library catalogue and OpenURL resolver by adding their LOCKSS box as a target to an OpenURL Resolver.

Memento: The Internet Engineering Task Force (IETF) is in the process of standardizing Memento, a mechanism created by Michael Nelson and Herbert van de Sompel by which browsers can access preserved versions of websites. With funding from the Mellon Foundation, the LOCKSS team is currently implementing Memento so that LOCKSS boxes will conform to this standard to access past content.

Three audit and verification tools detail what content is in a library’s LOCKSS box and the content’s preservation status.

- On demand, a LOCKSS box produces a KBART (Knowledge Bases And Related Tools) report of the locally preserved content.
- A LOCKSS box displays detailed preservation status for each Archival Unit. (An Archival Unit is typically a volume of a journal, or a complete book).
- A LOCKSS box administrator can use a properly configured web browser from an authorized IP address to view preserved content through an “audit proxy”. The viewer sees the content as it was collected by the LOCKSS system.

Librarians administer their institution’s LOCKSS boxes through a web browser that allows them to easily select new content for preservation, monitor content’s preservation status and a variety of other functions (Stanford 2012x).

Post cancellation access to all preserved content is ensured as the content is under the library’s local custody.

8. Migrating Obsolete Formats

LOCKSS preserves all web published formats (animations, datasets, moving images, still images, software, sound, text) and genres (journals, books, blogs, websites, scanned files, audio, video). The LOCKSS software is format-agnostic and preserves all content in its original format, as delivered from the publisher, including the format metadata that enables a browser to render the content.

The field of digital preservation has lavished much attention on the risk of format obsolescence (Rosenthal 2009). For web content, a format becomes obsolete when commonly available browsers and plugins can no longer render it. There is little evidence that this is happening to the widely used web formats in which the books and journals of the GLN or the special collections in PLNs are published.

Nevertheless, more than seven years ago the LOCKSS technology demonstrated its ability to handle format obsolescence if and when it occurs (Rosenthal et al. 2005). When a browser requests content from a LOCKSS box, it uses the part of the HTTP standard called "content negotiation" to specify the formats it can render. If the requesting browser cannot render the format of the preserved content, the
LOCKSS box invokes appropriate format migrators to create a temporary copy in a format the browser can render (Ockerbloom 1998). After use, the LOCKSS box discards the temporary access copy.

(The LOCKSS team demonstrated this process by modifying a browser to claim that it could not render the GIF format, whereupon the LOCKSS box created temporary copies of the GIF images in a histology paper in PNG format. An unmodified browser saw the same images as GIFs).

The LOCKSS Program’s “migration on access” approach has significant advantages over “format normalization” as it preserves the original artefact, uses much less overhead, saves money and takes advantage of the most up to date technology. Preserving the content in its original format satisfies archival requirements. It allows the LOCKSS system to be frugal with storage space. We know of no preservation system that discards the original bits after migrating them to a new format. Migrating and keeping both the original and the migrated copy multiplies the storage requirements for a preservation system by the number of migrations.

Preserved content is migrated by the most recent, and presumably best, technology available at the time the reader requests access. Preserved content is rarely accessed. Performing migration only when and if it is needed reduces the resource cost. Content can be migrated directly from the original to the current format, minimizing the effects of format conversion artefacts. The format converters, once developed, can themselves be preserved to document the original format.

9. Taking Action

Paper library systems, through implicit library collaboration and cooperation are robust and effective at preserving cultural memory. Libraries have an opportunity and a responsibility to acquire these characteristics in the digital environment.

The Stanford University LOCKSS Program provides libraries with the means to cooperate socially, technically and financially to build and preserve authoritative copies of subscription, open access and special digital collections. The open source distributed digital preservation approach builds tamper evident infrastructure. The time-tested robust and resilient paper world implementation of “lots of copies keep stuff safe” is easily implemented in the digital environment. Local infrastructure and collections ensure continual access and the more content is replicated, the greater its chances of surviving for the long-term. Collaborative, collective action at a local level yields highly leveraged results globally.

References


