May 2008

HELIN Task Force on Electronic Archiving Report

HELIN Task Force on Electronic Archiving

Follow this and additional works at: http://helindigitalcommons.org/task

Task Force on Electronic Archiving

May 2008

Task Force of the HELIN Serials Committee and the HELIN Collection Development Committee

Members: Judith Stokes, Andrée Rathemacher, Susan McMullen, Kathy Blessing

Overview

The charge of the Task Force on Electronic Archiving was to investigate electronic archiving technologies, specifically LOCKSS and Portico, and make a recommendation to the HELIN Chairs Council.   The purpose of both LOCKSS and Portico is to preserve digital scholarly material so as to ensure ongoing access.   Although their missions are similar, their methodologies for preserving electronic scholarly content are vastly different. LOCKSS (Lots of Copies Keeps Stuff Safe) is an open source software/ hardware system which enables a decentralized network of e-journal caches. Portico is a centralized archiving service which enables libraries to outsource preservation of electronic journals. Of particular note is the contrast in coverage and publisher participation. Even though LOCKSS and Portico are continuing to add new publishers, HELIN libraries should carefully analyze publisher participation against their own subscriptions.

Participating publishers in LOCKSS: http://www.lockss.org/lockss/Publishers_and_Titles
Participating publishers in Portico: http://www.portico.org/about/part_publishers.html

To use LOCKSS, a library buys its own server to download archival material, subscribes to the LOCKSS service (cost ranges from $2,160 to $10,800 annually depending on Carnegie classification), and installs the LOCKSS software. In order to preserve materials in their "LOCKSS box" a library must have an active subscription to electronic content and have permission from the publisher. The point of LOCKSS is to help libraries build and maintain their electronic collections. In short – it allows you to keep what you buy – as long as there is publisher participation within LOCKSS. The LOCKSS software helps you to make your institutionally owned content web accessible in the event that it is not available from the publisher's web site.   Here is how Victoria Reich, Director of the LOCKSS Program at Stanford University Libraries, describes the process*: Your LOCKSS box has an IP address. That IP address, via your subscriptions can access content. If the publisher of that content has made the content "LOCKSS compliant" then this content is automatically ingested into your LOCKSS box. Your LOCKSS box will automatically collect, preserve, and keep for you to serve to your readers all content that a) you subscribe to or is open access; b) the publisher has made LOCKSS compliant.*

Portico is an outsourced service that is available through subscription. Annual participation fees are determined by the Library's material budget and are available from the Portico web site. Portico originated in affiliation with JSTOR. Having concentrated on scanning technology in order to preserve the best possible digital copies of print back-files, JSTOR had no means of capturing digital originals and incorporating them in the collections until Portico was founded. Unlike LOCKSS, Portico devised a means to receive and normalize publishers' original *source files*, which are electronic files containing graphics, text, or other material that comprise an electronic journal article, issue, or volume. (Source files may differ from files presented online most typically by including more information or higher quality graphics.) Like JSTOR, Portico achieves redundancy by maintaining multiple servers and locations, and depends on library members for monetary support only. Unlike JSTOR, Portico seeks to preserve e-journal publishers' entire output, and charges them for the service. Libraries that subscribed to Portico are granted access to archived content only if a specific trigger event occurs. Trigger events include: 1) publisher stops operations; 2) publisher ceases to publish a title; 3) publisher no longer offers back files; 4) catastrophic failure of publishers platform.

Participation in LOCKSS or Portico can only be through an individual library. However, interested libraries may receive a 5% consortia discount, dependent on how many libraries decide to subscribe. *Upon signing the Portico journal archive license, a library belonging to a consortium that has agreed to promote Portico will receive a 5% reduction in their Annual Archive Support payment.* Because loss of journal content is often the result of cancellation, libraries should carefully look at the post-cancellation policies of each product as described later in this report.

Below is a chart outlining the differences, pros/cons, strengths/weakness of each product.

LOCKSS	Portico
- Libraries maintain their own server (LOCKSS box) and as such it is Institution driven and the library "owns" the content within their LOCKSS box. (Libraries maintain local control over their content) - Software is available for free download and weekly updates are easily done - Content remains in its original format - Ensures continual access to archived	- Subscription based – outsourced. If a library ceases participation in Portico they lose access to any content that had been opened up to them as members up to that time. However, if at a later day, the library re-joins, they would have access to all "liberated" or as they say "triggered" content. - Content is put into a standardized archival format – it will not look like the

content. If for any reason publisher content is not available via the web, users can automatically access it from the LOCKSS box. • Participating publishers are small and are not generally ones from which HELIN libraries receive content. • Builds local collections • Preserves all web formats	original • Ensures campus wide access to archived content when a trigger event occurs. • In this centralized solution, all library participants are granted access to an affected journal regardless of their subscription history. • Subscribers support a larger mission of helping to build a permanent, digital archive of scholarly materials.

Collection Analysis

Before making a decision to purchase either product, it is essential that libraries perform an analysis of the titles in their e-journal collections to ascertain what percentage of titles would be available via LOCKSS or Portico in the event that content is removed from the publisher's web site or is cancelled by the library.

Portico offers a Holdings Comparison Service http://www.portico.org/news/HoldingsCompService.html Given a list of ISSNs supplied by the library, the service reports back on a spreadsheet which titles are archived in Portico, which are committed to Portico (i.e., will be archived in future), and of those, which titles the publishers have contracted with Portico to provide post-cancellation access to former subscribers. RIC submitted a list of 2,132 ISSNs representing paid e-journal titles presently subscribed and later the same day received a spreadsheet indicating that 47% of the RIC titles are either archived or committed to be archived. The cost of preserving RIC's 1,001 Portico titles would be $4 to $7 per title at present and would decline as more journals are committed to the archive.

Using a different approach, RWU analyzed LOCKSS and Portico coverage of journal titles contained within subscribed journal packages. Below is a table showing the Participating Publishers in LOCKSS and Portico for RWU Electronic Journal Subscriptions

RWU Electronic Journal Publisher Subscriptions in Portico and LOCKSS		
	PORTICO	LOCKSS
BioOne	YES	YES -- 61% of BioOne Publishers
American Accounting Association	YES	
American Anthropological Association (Anthrosource subsription)	YES -- selected titles that are archived by JSTOR (59%)	YES -- all titles
American Chemical Society	YES - 60% of content. NO Post cancellation access	NO
American Psychological Association (access through PsycArticles)	YES	NO
Blackwell - Synergy	NO	NO
Elsevier (Science Direct)	YES -- majority of titles	NO
JSTOR	YES?	NO
Nature	YES	NO
Project MUSE	NO	YES
SAGE Premier	YES	SELECTED titles
Springer Link Contemporary	Selected titles -- NO Post cancellation access	NO

Post-Cancellation Access: LOCKSS vs. Portico

There are two main scenarios in which libraries could lose access to their online scholarly content: 1) if the content is no longer produced or hosted by the publisher or 2) if the library cancels its subscription. For libraries in the HELIN Consortium, the greatest threat of losing access to digital materials, most of which have been purchased or leased from publishers, comes from cancellation or non-renewal of these materials. In other words, the possibility of losing significant online content as a result of budgetary constraints is a greater concern than losing content due to publisher business decisions.

Both LOCKSS and Portico offer some form of post-cancellation access to digital materials.

LOCKSS

With the LOCKSS system, a library would maintain post-cancellation access to anything harvested by its LOCKSS box. Victoria Reich explains how this works:

> You have post-cancellation (access) to everything in your LOCKSS box. LOCKSS allows you to keep what you buy… All the publishers who participate in LOCKSS allow post cancellation access. Your LOCKSS box knows automatically what your institution can and cannot collect. If your LOCKSS box collects it – it's yours.
>
> Your LOCKSS box has an IP address. That IP address, via your subscriptions, can access content. If the publisher of that content has made the content "LOCKSS compliant" then this content is automatically ingested into your LOCKSS box.

> *Your LOCKSS box will automatically collect, preserve, and keep for you to serve to your readers all content that a) you subscribe to or is open access; b) the publisher has made LOCKSS compliant.*

According to Ms. Reich, libraries have successfully used LOCKSS for post-cancellation access. After cancellation, the LOCKSS box itself acts as a server for the content. According to tests, LOCKSS boxes and bandwidth are adequate to serve content to institutions "many times larger than any in Rhode Island." Libraries have integrated LOCKSS with proxy servers, and soon, LOCKSS will work with OpenURLs. Ms. Reich:

> *LOCKSS changes the model of access to content from rent to own. We want libraries to own the assets and stop paying rent – and to not buy rental insurance. If you stop paying for LOCKSS Alliance membership – you have your LOCKSS box, you have the content in your LOCKSS box. No one can take it away from you.*

It is important to understand that even though the library is taking local responsibility for preservation under the LOCKSS model, publishers must agree to allow the LOCKSS box to "harvest" their sites. They must ensure as well that their content is "LOCKSS compliant." Thus, LOCKSS only works with participating publishers. (See http://www.lockss.org/lockss/Publishers_and_Titles.) It is worth noting that many large publishers participate with LOCKSS minimally or not at all: Elsevier (no), Wiley-Blackwell (one title), Taylor & Francis (Project Muse titles only); Springer (no). **Thus at this time the usefulness of LOCKSS for post-cancellation access is limited.**

Portico

While the purpose of LOCKSS is essentially to allow libraries to retain ownership of digital content in the same way that they retain ownership of print books and journals in their collections, Portico has a broader mission:

> "*The mission of Portico is to preserve scholarly literature published in electronic form and to ensure that these materials remain accessible to future scholars, researchers, and students.*" (http://www.portico.org/about/)

Portico's goal is not service for individual libraries, but for the scholarly communication process as a whole. Portico works with both publishers and libraries to address long-term archival needs. Member libraries are not buying insurance for their own collections so much as they are participating collectively in developing a permanent, digital archive of scholarly materials. Ken DiFiore, Associate Director of Library Relations at Portico, explained that Portico exists so that libraries and consortia around the world don't have

to reinvent the wheel and figure out how to handle digital archiving on their own: "By centralizing preservation with Portico, libraries and consortia will save lots of money over the long haul. If everyone contributes, Portico can do the job. If everyone decides to wait, Portico won't survive."

Accordingly, post-cancellation access to content for individual libraries, while possible under Portico, is not Portico's primary purpose. Portico's agreement with publishers is essentially to serve up their content if, for business reasons, it is lost, orphaned, or abandoned. Publishers trust Portico not to release their content unless the publisher abandons the content.

On the list of publishers that participate with Portico (http://www.portico.org/about/part_publishers.html) an asterisk (*) indicates those publishers that "have chosen to name Portico as one possible mechanism to fill post-cancellation access claims by participating Portico libraries." Mr. DiFiore made it clear that in most cases, Portico would only provide post-cancellation access if the publisher chose not to do so or was unable to do so. And in such a case, Portico would work with the library and the publisher to establish access rights. The library would be asked to provide proof documenting their entitlement to the journal. So far, Portico has only satisfied one post-cancellation claim, and the details are not public. Of all the publishers that participate with Portico, only one (SIAM) has designated Portico as their official source of post-cancellation access.

Mr. DiFiore summed up by saying that Portico is taking a "wait and see" attitude on post-cancellation access. Portico wants to be "like Switzerland." They want to engender trust from publishers and do not want to cut into their business model.

Recommendation Re: Product for Post-Cancellation Access

For post-cancellation access, neither LOCKSS nor Portico are recommended at this time.

LOCKSS is the better product for providing individual libraries with post-cancellation access to purchased electronic resources. However, given the fact that there are currently a very limited number of publishers that participate in LOCKSS, its usefulness in this regard is limited. Should more publishers and titles be harvestable through LOCKSS in the future, participation should be revisited.

Portico is a worthwhile endeavor, but it is not primarily intended to be a mechanism by which individual libraries can access purchased content after they have cancelled their subscription with the publisher.

To ensure post-cancellation access, the best solution for HELIN libraries at this point in time is to push publishers to:

1. Allow perpetual access, through their sites, to content a library has paid for, even post cancellation, and to provide for this in the license agreement;
2. Participate in LOCKSS and allow Portico to provide post-cancellation access to their journals;
3. Convert the older volumes of their current journal titles to open access, which would make post-cancellation access less of an issue.

HELIN libraries should reinvestigate both LOCKSS and Portico in 2-3 years to see if post-cancellation access through either has improved.

Working Group Recommendation

From our study it appears that LOCKSS is truer to the mission of libraries – working together to collectively preserve access to scholarly output of all kinds. What you put in your LOCKSS box you own.   However, as noted above, there are currently a very limited number of publishers participating in LOCKSS which makes its usefulness limited at this time.   Portico covers many more of the publishers and journal titles that most HELIN libraries subscribe to, but does not ensure a reliable mechanism for post-cancellation access. The working group would recommend an individual HELIN library subscribe to Portico if publisher participation in Portico sufficiently matches the library's electronic journal holdings and the library is concerned about the loss of content from the publisher's web site.

A Note about CLOCKSS (Controlled LOCKSS)

CLOCKSS could possibly be the best solution for those who are most interested in continual access due to publisher failure to provide access. CLOCKSS' mission is to be *a community-governed partnership of libraries and publishers working to achieve a sustainable, globally distributed Archive and ensure reliable, long-term access to scholarly e-content.* Seven libraries and eleven publishers are involved in its Pilot Program. A list is provided at http://www.clockss.org.   Using LOCKSS software, they will continually monitor and preserve content over time. *If digital content becomes no longer available from the publisher, content will move to a hosting platform and the impacted content will be made available for free to the world.* CLOCKSS has intentionally brought together venerable publishers and libraries emphasizing a shared legacy of long-term sustainability. Unlike LOCKSS, library "nodes" (locations for LOCKSS boxes) are limited (they will be added by invitation only), allowing publishers collectively to exercise some control over the shared stewardship of their content. After the grant-funded pilot program, additional libraries and publishers will be asked to join the effort. The question remains to be answered as to whether or not this is a sustainable model.

Institutions supporting library "nodes" cannot bear all the costs of the program, so the libraries at large are being asked to join – in essence to help underwrite a platform for guaranteed "open access" to scholarly content after a publisher no longer is able or willing to host it. Library participation in CLOCKSS is based on Library Materials Budget. For example, for a library with a materials budget of $1–2 million, the annual membership costs $600.

APPENDIX:

**LOCKSS AND PORTICO USAGE
AT SIX REGIONAL INSTITUTIONS**

MAY 2008

LOCKSS
(Lots of Copies Keep Stuff Safe)

PRODUCT INQUIRY
RESPONSE SUMMARIES

Institution # 1

Dartmouth College Libraries

Contact: Jen Fritz, Library Systems Manager

Email: Jennifer.Fritz@Dartmouth.EDU

Phone: (603) 646.9394

- Began using in 2003.

- Has hundreds of titles archived in this product. Archives by package or by title. For example when OUP (Oxford Univ. Press) content is released, automatically adds titles to library's archive. Also archives on a title by title basis. If own data already (because bibliographer chose title) then want it cached.

- Maintaining/updating: 3 staff person all work in minimal capacity. Barring huge hardware or software problems, maybe an hour or two of work per month. Jen does all hardware work and receives software updates from Stanford's LOCKSS team. One of AULs gives go ahead to archive titles when know library owns package. Support person in Acquisitions looks up individual titles.

- Current set-up: Dell desktop machine (cost under $1k). Software: free from LOCKSS but library needs to burn to CD. Have two or three software updates that require CD-burning each year.

- Jen, as LSM, was staff person assigned to initially set-up LOCKSS.

- Initial set-up time: First LOCKSS box was easy.

- Another set-up time later: Library needed to purchase a second LOCKSS server. Purchased hardware that turned out to be too new. Had to fiddle a lot with second set-up. Tom Lipkis, Senior Software Architect at Stanford's LOCKSS, was extremely helpful. Set-up: Monitor, keyboard, desktop machine (ordered from Dell.com). Also needed to purchase different NIC (network interface card). Might have been able to avoid purchasing NIC if looked at Dell more closely.

- Best features: Likes all of LOCKSS. Able to use inexpensive hardware, has simple interface. Especially likes being able to restore archived titles (when upgrading hard drive) by uploading an archive. LOCKSS "crawls" other LOCKSS caches to put data back on library's hardware.

- ➢ Challenges/problems: Major drawback is can't buy hardware that's too new. OpenBSD OS doesn't have drivers
   to support.

- ➢ Also has Portico. Not have details about how library using it.

_____

Institution # 2

Marine Biological Laboratory and Woods Hole Oceanographic Institution Library

Contact: Matt Person, Technical Services Coordinator

Email: mperson@mbl.edu
Phone: (508) 289.7345

- ➢ Began participation in July 2005.

- ➢ Has 187 titles archived (as of May 8, 2008).

- ➢ Technical Services Coordinator initially set-up. (Matt not Coordinator at that time.)

- ➢ Initial set-up: Dedicated 2.4 GHz desktop. Still using same.

- ➢ Technical Services Coordinator (Matt) adds titles and an IT associate assists.

- Maintaining/updating time: 15 to 20 minutes per month.

- Best features: Participating in archiving of clean copies of library's subscribed content. Very little maintenance.

- Challenge (not problem): LOCKSS periodically emails archived titles announcements. Library has to cull out titles it subscribes to and add them to Box.

- Never had Portico.

- Additional comments: "The main idea is very little work involved - since you benefit from LOCKSS only when there is some calamitous event which triggers content being released, one does not really see the benefits on a daily basis; what you do have is a general peace of mind... it's all very simple."

_____

Institution # 3

University of Connecticut Libraries

Contact: Dave Bretthauer, ITS Enterprise Team Leader
Email: Dave.Bretthauer@uconn.edu
Phone: (860) 486.6494

- Began using in spring 2003.

- Has 4784 "Archival Units". Add titles as announcements new titles or journal years are available for caching.

- Availability for caching licensed journal content is time-sensitive.

➢ Currently cache has two 1 TB drives.

➢ Maintaining/updating work: Dave only one.

(University of Connecticut cont.)

➢ Time for maintaining/ updating: 15 minutes/ week. Cache new titles and save new copies of
configuration backup file in two separate locations. Every 4 to 6 months spend 2 hours or less updating

CD version.

➢ Hardware maintaining/updating: Over past two years hard drives in cache replaced twice. Originally installed one 80 GB drive. April 2007 replaced with two 250 GB drives. April 2008 replaced with two 1TB drives. Replacing hard drives takes few hours over several days. Time mainly spent planning then communicating with LOCKSS team to migrate data. Only other time: extending warranty on cache itself.

➢ Hardware: Dell GX270 (purchased August 2004); 1 GB RAM; two 1TB SATA drives; one external USB enclosure (for migrating data from old hard drives to new). This connected to KVM in server farm but in past had its own keyboard, video monitor, mouse.

➢ Initial set-up: Dave worked alone. Took approx. 2 weeks. Had not previously needed to obtain a static IP address or configured it on a server.

➢ Initial set-up: Used older PC. (Due to current storage needs requires PC that uses 1GB RAM; at least two 1TB hard drives; bootable CD ROM drive, and either floppy drive or USB port.)

- ➢ Best features: Very easy once set up. Have local copy of content library has licensed.

- ➢ Challenges/ problems: Because have not yet implemented EZProxy (currently in progress), providing user access to cached data not been easy. Will be useful for on-campus users in situations where campus loses connection to Internet.

- ➢ Also has Portico. Supports both efforts to hedge preservation bets. Far too early in digital preservation game to place all eggs in one basket.

_____

Note: The institutions provided above were selected from the list of participating LOCKSS libraries (http://www.lockss.org/lockss/Libraries). Victoria Reich (vreich@standford.edu), Director of LOCKSS, offered names of most knowledgeable individuals to contact at these institutions concerning product usage.

PORTICO

**PRODUCT INQUIRY
RESPONSE SUMMARIES**

Institution # 1

Brown University Libraries

Contact: Steven Thompson, Co-Leader Technical Services
Email: Steven_Thompson@brown.edu

Phone: (401) 863.2976

➢ Participated since December 2006.

➢ PORTICO site currently shows 4,078 titles archived.

➢ Attractive service aspect: No staff or hardware needed to set-up or maintain. All work done at Portico's end. In a sense, library outsources this service.

➢ Best feature: Library not need to do anything. If trigger event occurs Portico notifies. Library given access to affected titles. Nothing required except to insert URLs into records.

➢ No challenges/ problems.

➢ Never had LOCKSS.

➢ Designate 4 staff members for access to archive via username/ password. Access for auditing purposes only. If trigger event occurs affecting one of titles included in Portico then archive content of title made available to Brown community by IP authentication – e.g. last May first trigger event occurred when *Graft: Cell and Organ Transplantation* removed from Sage's platform (became unavailable). Portico notified its participants of event. Made content of archive available to member libraries. Library received information on URL to use, etc. and placed information in catalog. Title then accessible to university community.

What entry for *Graft* looks like in A to Z Ejournal List:

Graft (Georgetown, Tex.) (1522-1628)
from 01/01/2001 to 03/31/2003 in Portico (Triggered Content)

➢ Triggered event titles are available from off-campus as well as on campus. Content obtained via a proxy (EZ proxy) or VPN (client or Web based).

_____

Institution #2

Fairfield University
Contact: Keith Stetson, Collection Development Librarian
Email: Kstetson@mail.fairfield.edu

Phone: (203) 254.4000 ext. 2184

- ➢ Participation confirmed December 7, 2007.

- ➢ Contact person: Collection Development Librarian.

- ➢ Hosted service: no equipment, set-up, maintenance, or staff time.

- ➢ Best features: Portico's mission (http://www.portico.org/about/).

- ➢ No challenges/ problems.

- ➢ Never had LOCKSS.

_____

Institution #3

Lesley University
Contact: Marilyn Geller, Collection Management Librarian

Email: mgeller@lesley.edu

Phone: (617) 349.8859

- Signed contract December 2006.

- As of June 2007 (most current data) subscribed to approximately 450 individual journals and 118 titles were committed for inclusion in Portico. Publishers that produce titles Lesley subscribes to have joined project since last year.

- Believes Portico numbers will continue to increase and cover more of library's important titles.

- Product operates without library doing anything. Very minimal work. Portico switched on access to title in archive. Notification sent to all members. Library turned on access in Open URL Link Resolver.

- Best features: Portico guarantees archive if something happens to publisher. Small fee per title: insurance that materials library paid for (i.e. materials researchers need) remain available in readable format by current/ future standards.

- Cost per year for "archive insurance": Membership cost divided by number of titles subscribed to that are scheduled for inclusion. Very low cost per year: between $18.00 and $19.00 per title. Assumes cost per title will decrease as more publishers add their titles. Great price at present and probably even better later.

(Lesley University's response cont.)

- Portico also planning for "perpetual access after cancellation". If library needs to cancel a subscription (and if publisher license allows), will have access to what it

paid for in past through Portico without having to figure out how to serve content to users. Access situations: Publisher licenses may offer access after cancellation by sending data files. Creates library problems with providing reasonable access for users. Publishers may also say after cancellation will allow access on delivery platform but library may have access fee. Portico offers cleaner solution. Point to archive and get content library paid for in past but no longer subscribes to.

➢ No challenges/problems.

➢ Concern: Smaller publishers not included in Portico yet. Represents significant journals for Lesley's population. Smaller publishers lack resources that larger ones have. Currently smaller not high on list of potential Portico participants. Some small publishers even lack electronic versions of titles.

➢ Never used LOCKSS. Not have systems staff or money for dedicated machine and time to do work. Portico gathers, preserves, and migrates their titles.

➢ Access: Added to proxy server, and created access at title level through the OpenURL Link Resolver. Titles show up in A to Z list. If indexed in subscription databases, link resolver points to access. (No different than any other ejournal platform.) Portico no issues with use of a proxy server. Simply matter of registering address in proxy server configuration. Library routinely does this for any electronic resource. Portico is available to entire Lesley community from anywhere as long as authenticate through system. (Same as any other e-resource).

➢ Additional comments: "I love Portico. It's the right thing to do for the scholarly community in general and the right thing to do for the Lesley community specifically. Our job is to insure access to valuable research materials, and Portico does that for us. Also, I'm responsible for providing statistics for external review committees, and I often include number of titles preserved in Portico just to bring the issue to the forefront. Digital preservation sounds like a tenuous thing; Portico gives me a very high comfort level that we can protect our digital collections for the future in a like manner to how we protect our print collections."

_____

Note: The institutions provided above were selected from Portico's participating libraries list (http://www.portico.org/about/participating_libraries.html). Portico's Dawn Tomassi (participation@portico.org) offered names of most knowledgeable individuals to contact at these institutions concerning product usage.

K. Blessing

05/08