Final Report of the 2CUL LOCKSS Assessment Team
Cornell University Library & Columbia University Library

Report Completed: March 2011
Public Release: October 2011

Introduction

Although LOCKSS is considered a successful digital preservation initiative, neither of the CULs feels that they fully understand the potential of the system for their own settings and collections. There is a range of practical issues that need to be explored in order to leverage this preservation system. In support of this goal, a joint team was established in November 2010 to investigate a range of questions to assess how LOCKSS is being deployed and the implications of local practices for both CUL's preservation frameworks. This study was seen as a high-level investigation to characterize the general landscape and identify further research questions.

The team worked with a condensed timeline, November 2010-December 2011, and investigated the following questions:

1) To build a collection of preserved journals beyond the journals already preserved via the LOCKSS network, a library needs to select the titles it wants to preserve (subscription or open access). Working with the Stanford LOCKSS team, the next step is to obtain or confirm the publishers' permission to preserve the titles in the system. The Stanford team is responsible for building and testing the required publisher-specific "LOCKSS plugin." The size of the LOCKSS Alliance ensures each title chosen for preservation has a critical mass of preserving institutions. How does this process work for Columbia and Cornell? Who is involved in overseeing this process and tracking such collection decisions?

2) What needs to happen when a journal is canceled to have access to back issues? What kind of a mechanism needs to be put in place between the ERM license record for journal subscriptions (library management systems) and the local LOCKSS box to support uninterrupted access to digital content? What is the internal monitoring mechanism - all manual or can a part of it be automated?

3) Neither of the institutions has chosen to participate in CLOCKSS. Do we have a sufficient understanding of the difference between these two strategies? LOCKSS provides a community approach to long term preservation of a library's local collections while CLOCKSS aims to provide a long-term global archiving solution that will serve the joint library and publisher communities in the event of a long-term business interruption or in making orphaned or abandoned works readily available to the scholarly community.

4) How do we keep track of which e-subscriptions are represented in LOCKSS to understand their preservation status?

5) What are the tasks involved (and resources needed) in maintaining a LOCKSS box from the IT and collections perspectives?

6) Do we understand the difference between LOCKSS and Portico and have a clear sense of how these two approaches fit into our broader preservation objectives?

7) Two Cornell libraries maintain Digital Commons repositories (http://www.bepress.com/ir/). Now that Digital Commons has partnered with LOCKSS (http://www.lockss.org/lockss/News), what implications does this relationship have for our use of LOCKSS overall?

8) Have we taken advantage of LOCKSS so far by gaining access to a canceled subscription or a closed journal? Have we participated in a failure-recovery test?

9) Cornell and Columbia have received invitations to participate in the LOCKSS-USDOCS initiative. Subsequent to receiving the invitations, the libraries have received input both pro and con regarding the initiative. Can we place the various perspectives in context? Given what we know about the initiative, can we make a recommendation regarding the two CULs' participation?

10) Cornell has submitted its electronic and print serial holdings data to Portico and has received reports from Portico about Portico's coverage of its holdings. (Columbia's Portico analysis is forthcoming.) Cornell also has data about LOCKSS' coverage of its serial holdings. Can we do an analysis that compares Portico and LOCKSS coverage? And given a likely similarity between Columbia and Cornell's serial holdings, can we illuminate the Cornell Portico/LOCKSS data in ways that might guide the two CULs' ongoing participation in LOCKSS and Portico?

11) With the Portico/LOCKSS data in hand, do we know whether there is a significant body of material that Portico is not expected to cover, and whether these titles would be viable candidates for enrollment in LOCKSS? If so, what effort would be required?

12) Would it be possible for the two CULs to share a single LOCKSS box? If so, can we identify any risks or benefits associated with such an approach?

13) Given what we know about how the two CULs have approached LOCKSS participation to date and about how other libraries and library groups have participated in LOCKSS, can we make any recommendations regarding how best to position LOCKSS-related decision making within the CULs prospectively?

The research was coordinated by Marty Kurth (Cornell, Digital Scholarship Services) and the group members included Jeff Carroll, Columbia, Collections; Bill Kara, Cornell, Central Library Operations; Bill Kehoe, Cornell, Information Technology; and Breck Witte, Columbia, Library Information Technology. Jim Spear (Technical Services Assistant, Cornell) conducted the data analysis. The project was conceptualized and led by Oya Rieger (Associate University Librarian for Digital Scholarship Services, Cornell) and Patricia Renfro (Deputy University Librarian and Associate Vice President for Digital Programs and Technology Services, Columbia).

Background Information

Both CULs are members of the LOCKSS (Lots of Copies Keep Stuff Safe) Alliance. With 200+ participating libraries, LOCKSS provides libraries with open-source software to support the preservation of two kinds of content:

- Through a private LOCKSS network (PLN), institutional or consortial web-published materials. (For instance, the MetaArchive Cooperative uses LOCKSS for member institutions to identify collections that they want to preserve and ingest into a geographically distributed network of servers in multiple locations.)
- Through the global LOCKSS network, web-based electronic subscription, including e-journals, to support post cancellation access (~ 6600 committed journals for LOCKSS vs. ~12,000 titles for Portico).

Content preserved by libraries through LOCKSS becomes a part of their collections, and they have perpetual access to 100% of the titles preserved in their LOCKSS box. The box collects content from target web sites using a web crawler similar to those used by search engines and continually compares the content it has collected with the same content collected by other LOCKSS boxes (http://lockss.stanford.edu/lockss/How_It_Works).

Key Findings and Recommendations

This study was seen as a high-level report to understand the general landscape. Therefore the team worked with a condensed timeline, November 2010-December 2011, and presented the following interim recommendations.

- Both CULs have made the de facto decision not to serve from their LOCKSS boxes e-journal content that is preserved in LOCKSS and no longer available from the publisher, thus essentially treating LOCKSS as a dark archive.

- Our analysis of e-journal preservation coverage indicates that LOCKSS and Portico combine to preserve only a relatively small percentage of the CULs' e-journal holdings, for example, less than 15% of Cornell e-journal titles as a whole. There is overlap in coverage between the two services, but both services preserve titles uniquely. We recommend that the CULs continue to track and assess the coverage of LOCKSS and PORTICO to better understand the role of these two services in preserving each institution's collections.

- Collection development, technical services, and library IT staff skills all come into play in decisions related to e-resource preservation. Because many e-resources are not yet preserved and because responsibility in the CULs for e-resource preservation decision-making has historically been diffuse, we recommend that the CULs assign ongoing e-resource preservation responsibilities to designated groups that bring together skills from these three areas and that the groups have designated sponsors in library upper administration. We recommend that staff resources be formally reallocated to e-resource preservation in the context of a reduction in support allocated to other activities.

- The overall lack of e-journal publisher participation in preservation programs such as LOCKSS and Portico offers the two CULs an opportunity to use their individual or combined influence with publishers to improve the state of e-journal preservation as a whole.

Findings

As an outline for this report, the 2CUL LOCKSS Assessment Team has chosen to use a question-and-answer format as suggested by our charge and the subsequent questions asked of us. In addition to those thirteen questions we have added a fourteenth by way of conclusion.

1) To build a collection of preserved journals beyond the journals already preserved via the LOCKSS network, a library needs to select the titles it wants to preserve (subscription or open access). Working with the Stanford LOCKSS team, the next step is to obtain or confirm the publishers' permission to preserve the titles in the system. The Stanford team is responsible for building and testing the required publisher-specific "LOCKSS plugin." The size of the LOCKSS Alliance ensures each title chosen for preservation has a critical mass of preserving institutions. How does this process work for Columbia and Cornell? Who is involved in overseeing this process and tracking such collection decisions?

Both institutions (Columbia and Cornell) participated in the early pilot LOCKSS Humanities Project. At Columbia, the process during the pilot involved a selector identifying titles that she wished to see preserved via LOCKSS. She then worked with the Library Technology Office to contact the publishers for permission to archive the titles in LOCKSS and to create plug-ins for the publishers' platforms. Only a handful of titles were identified at Columbia and added to the archive in this way during the pilot and few, if any, have been added in this way since. Cornell's process and experience in the pilot was substantially the same. Neither Columbia nor Cornell has formal processes currently in place to identify titles for preservation or to follow the steps necessary to preserve them via LOCKSS.

Regarding such processes, a 2006 NYU LOCKSS Task Force found that two external dependencies impacted the process. First, some journals do not pass the technical evaluation of required elements for LOCKSS preservation. Second, some journal editors or publishers prove unresponsive to library preservation requests. As a result, the NYU Task Force was able to preserve 8 of the 22 titles it had identified as candidates for LOCKSS preservation.

Going forward, one can envision selectors identifying publishers whose titles have not yet been preserved during the selection or renewal process. The selectors would contact the Stanford LOCKSS Team and the journal publishers to pursue necessary permissions. Once permissions were obtained, the technical staff on the Stanford LOCKSS Team would create the plug-ins needed.

Alternatively, endeavors such as the PEPRS Project (Piloting an E-journals Preservation Registry Service http://edina.ac.uk/projects/peprs/index.html) might help facilitate the process of identifying publishers whose titles are not archived in any of the recognized archival/preservation services. The PEPRS project anticipates providing online functionality,

provisionally called "Holdings Comparison" in the PEPRS documentation, which would process a delimited list of an institution's holdings and run a match (e.g., on ISSN and eISSN) against the registry. The resulting report would identify those publishers and titles that have archival arrangements with recognized preservation agencies. Publishers listed in the report with no archival arrangements could be filtered out for further investigation by subject specialists. The subject specialists could then refer to the LOCKSS Alliance and Portico publishers and titles prioritized for further action based on criteria such as risk in order to enlist a "critical mass" of institutions that would be willing to cache those publishers' titles

2) What needs to happen when a journal is canceled to have access to back issues? What kind of a mechanism needs to be put in place between the ERM license records for journal subscriptions (library management systems) and the local LOCKSS box to support uninterrupted access to digital content? What is the internal monitoring mechanism - all manual or can a part of it be automated?

Both Columbia and Cornell have stand-alone ERMsand these ERMs are from different vendors. Populating the ERMs with the diverse and detailed information needed for managing, implementing and maintaining access to large numbers of e-resources is complicated and it differs at each institution. It is automated in some ways but also relies on much manual input and maintenance. Even higher priority potential uses of the ERMs, for example, importing usage and cost data for further analysis of our e-resources collections, have not been successfully implemented. Although each ERM has flexibility and potential for coding information and importing files, neither Columbia nor Cornell currently uses its ERM to record and manage details related to potential LOCKSS or Portico access. In 2008 a library intern worked with Cornell's e-resources staff to code details related to LOCKSS and Portico preservation into the ERM. This coding is inadequate, is only at the publisher or resource level (not at the individual title level), and is now significantly out of date.

The broader question of "what needs to happen when a journal is canceled..." is addressed in our responses to Questions 5 and 8. Essentially, both CULs have treated LOCKSS as a dark archive, with the assumption that titles could be made accessible on an as-needed, though not uninterrupted, basis.

3) Neither of the institutions has chosen to participate in CLOCKSS. Do we have a sufficient understanding of the difference between these two strategies? LOCKSS provides a community approach to long term preservation of a library's local collections while CLOCKSS aims to provide a long-term global archiving solution that will serve the joint library and publisher communities in the event of a long-term business interruption or in making orphaned or abandoned works readily available to the scholarly community.

As a general comparison of LOCKSS and CLOCKSS, LOCKSS enables libraries with licenses for content preserved in LOCKSS to have continued access to that content when events such as cancellations or publisher cessations occur, while CLOCKSS is a top-down effort among libraries and publishers to serve specified publisher content as open access should the publisher no longer make it available. For a brief, objective introduction to and comparison of LOCKSS and CLOCKSS (and Portico), see *Ensuring that 'e' doesn't mean ephemeral: a practical guide to*

e-journal archiving solutions. One important distinction between LOCKSS and CLOCKSS is that LOCKSS supports post-cancellation access, whereas CLOCKSS does not.

As reported in *Ensuring that 'e' doesn't mean ephemeral*, JISC Collections maintains a table entitled "Which NESLi2 and NESLi2 SMP publishers are participating in e-journal archiving programmes?" http://www.jisc-collections.ac.uk/archiving/participation, comparing publisher participation in LOCKSS, CLOCKSS, and Portico.

An at-a-glance comparison of the core strategies underlying LOCKSS, CLOCKSS, and Portico can be found in the table below from *Ensuring that 'e' doesn't mean ephemeral:*

Trigger Event	LOCKSS	CLOCKSS	PORTICO
Library cancels subscription and needs access to past issues to which they subscribed	Yes	No	Post-cancellation access can be provided as a service to participating publishers and participating libraries. If a library chooses to discontinue Portico participation, then they will no longer be able to get post-cancellation access to content through Portico.
E-Journal or its past issues are no longer available from the publisher	Yes	Yes. The title would be made openly accessible to all.	Yes. The title would be opened up to all active participants, regardless of whether they previously subscribed to the content.
Publisher has ceased operation and e-publication is no longer possible.	Yes	Yes. The title would be made openly accessible to all.	Yes. The title would be opened up to all active participants, regardless of whether they previously subscribed to the content.
Catastrophic failure of publisher's operations/servers	Yes	Yes, providing the publisher is unable to provide a service.	Yes, providing the publisher is unable to provide a service.
Temporary failure of publisher's operations/servers	Yes	No	No

4) How do we keep track of which e-subscriptions and e-books are represented in LOCKSS to understand their preservation status?

Identification of titles for which access has been triggered is not handled through the ERMs at either Columbia or Cornell. Each institution has relied on LOCKSS, Portico or the publisher to notify us of changes in access. The number of such titles has been so small (and relatively recent) that standard, more systematic procedures and policies have not been developed.

At LOCKSS Alliance membership renewal time, Cornell's head of collection development customarily asks the library IT person in charge of the LOCKSS box, "How many titles are we preserving?" The programmer queries the LOCKSS box, massages the resulting list of current archival units (journal volumes), and produces a list of current titles. Columbia uses the same method, and it is worth noting that this approach has been validated by LOCKSS technical staff. The number of preserved titles, the amount of disk storage used, and the number of system maintenance hours inform the decision to renew. Similarly, at renewal time Columbia verifies

that the technical and staff investment continue to be minimal before deciding to renew. A formal assessment of value has not yet been a factor in Columbia's renewal decision.

With regard to possible approaches to tracking, as mentioned in our response to Question 1, the two CULs might use "Holdings Comparison" functionality in a registry such as PEPRS to track the preservation status of e-resources held. Lacking such functionality, the libraries would regularly need to conduct analyses themselves similar to the one described in our response to Question 10 below.

5) What are the tasks involved (and resources needed) in maintaining a LOCKSS box from the IT and collections perspectives?

LOCKSS server architecture is in a transition phase from FreeBSD to Linux (CentOS distribution). (Columbia, for example, completed this transition in mid-February.) The use of Linux is expected to improve support for server virtualization, though some sites have reported success in running FreeBSD on virtual machines. FreeBSD server maintenance has historically involved very low overhead. Over the course of the last ten years maintaining the LOCKSS server at Columbia has required no more than 4-6 hours per year, including years in which the server itself was replaced. Columbia decided against virtualization of its LOCKSS server due to the unusual requirements of a LOCKSS cache, namely low CPU and memory demands but large storage requirements (local storage in the 2-4 TB range), which is precisely the opposite of the kinds of resources well suited to virtualization (high CPU and memory demands, with low storage). Cache content at both Columbia and Cornell is administered via a web interface. Both institutions' approach has been to follow the approach recommended by the LOCKSS Alliance, that is, to cache all available content, following the logic that the cost of disk space is low enough that this is an economically expedient alternative to deciding on a title-by-title basis which titles to cache. Eliminating title-by-title decisions also means that the time commitment required of collection development staff is minimal. Because neither Columbia nor Cornell has maintained links for cached LOCKSS content in its proxy server, there has been no overhead to date in terms of link maintenance. Thus, both institutions have essentially treated their LOCKSS caches as dark archives that provide potential sources of content that can be activated when needed, for example by proxying access to individual titles from their caches should the need arise.

More broadly with regard to LOCKSS from an IT perspective, CRL expressed the concern in its LOCKSS audit that a viable open source community has yet to form surrounding the LOCKSS software itself. This makes LOCKSS vulnerable with regard to succession planning should the Stanford LOCKSS Team no longer be able to manage the central technical support and development of the software.

That said, the CRL audit reported that in 2007 the LOCKSS Alliance generated enough income to cover the costs of the LOCKSS Team, which appeared to CRL to be a favorable sign regarding LOCKSS' prospective independence from soft-money sources, though CRL stopped short in its audit of saying anything definitive about LOCKSS' fiscal viability. Our LOCKSS contacts have informed us that since 2007 LOCKSS has been supported entirely through LOCKSS Alliance fees and contracts for service.

6) Do we understand the difference between LOCKSS and Portico and have a clear sense of how these two approaches fit into our broader preservation objectives?

As noted in our response to Question 3 above, for a brief, objective introduction to and comparison of LOCKSS and Portico (and CLOCKSS), see *Ensuring that 'e' doesn't mean ephemeral: a practical guide to e-journal archiving solutions*.

E-Journal Archiving for UK HE Libraries: A Draft White Paper has cases studies of the University of Glasgow and the London School of Economics and Political Sciences (LSE) e-journal archiving practices. The Glasgow case study notes that LOCKSS offers local control and access to archived e-journal content while Portico is a fully outsourced service. Glasgow's perspective is that it is important not to rely on only one e-journal archiving approach. The LSE case study notesthat there are overlaps in content between LOCKSS and Portico but generally that content differs significantly between the two services. LSE holds that no one e-journal archiving service fully meets its needs, so it subscribes to both LOCKSS and Portico as "insurance policies" to protect a portion of its e-journal collections.

Technically LOCKSS and Portico represent two different approaches to digital preservation. Portico ingests data files in whatever format the publisher uses. These could be database files, XML, HTML, or files in an unspecified proprietary format. Then Portico normalizes the files to a standard archival format which it can subsequently manage over time. As noted on the Portico website (http://www.portico.org/digital-preservation/glossary/#sfiles), "Portico's primary preservation methodology is *migration*, which involves transitioning content from one file format to another as technology evolves and file formats become obsolete."

LOCKSS collects 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. Formats that are collected and preserved include: spreadsheets, XML, HTML, PDF, video, and sound. LOCKSS defines "obsolete content" as "when a reader's web browser does not display the content." The reader's browser determines this based on the preserved format metadata, LOCKSS detects it and invokes an "on the fly" process that creates an access copy by migrating the preserved content and format metadata so that it displays properly in the reader's web browser. (For more information on the processes involved, see David S. H. Rosenthal et al, "Transparent Format Migration of Preserved Web Content," *D-Lib Magazine* 11:1 (January, 2005). http://dx.doi.org/10.1045/january2005-rosenthal)

There are strengths and weaknesses to each of these approaches: Portico does not require manifests and crawlers that need to be updated to accurately describe content boundaries, while LOCKSS does not require a comprehensive understanding of underlying data formats and structures. Another distinction involves the distribution of content. Portico is a centrally administered archive with content stored in multiple locations, while management and storage of LOCKSS content is, by design, geographically dispersed among Alliance member sites.

In its LOCKSS audit, CRL highlights the importance of locally managed content in the LOCKSS model. In particular, CRL points to the dependency in Portico's centralized model on publisher

agreements that involve waiting periods and publisher notification that could prohibit user access to archived content for extended periods of time after a trigger event occurs.

Finally, Michael Seadle, Dean and Director at the Berlin School of Library and Information Science, authored an article for Library Hi Tech in 2011 that came to the attention of our team too near the submission of this report for us to be able to analyze it fully. We provide a citation at the end of the report. According to the published abstract, the article's "findings show a significant overlap among the archiving systems. They also show that Portico has no prejudice against small publishers and that large publishers are as willing to choose the LOCKSS software as to choose Portico. LOCKSS does, however, archive many more small and arguably endangered publishers and may be the only economically viable choice for them."

7) Two Cornell libraries maintain Digital Commons repositories (http://www.bepress.com/ir/). Now that Digital Commons has partnered with LOCKSS (http://www.lockss.org/lockss/News), what implications does this relationship have for our use of LOCKSS overall?

The Cornell Catherwood Library (serving the Industrial & Labor Relations School) manages a Digital Commons repository instance (hosted by bepress) containing approximately 13,000 documents. The Cornell Law Library manages a Digital Commons instance (also hosted) containing between 400 and 500 documents. Both libraries have participated with six other higher education institutions that manage Digital Commons instances in preliminary discussions surrounding the implementation of a Digital Commons Private LOCKSS Network (PLN). Bepress (which markets Digital Commons) and the Stanford LOCKSS Team have taken some steps toward making a Digital Commons PLN possible: 1) Bepress has implemented permissions to enable the LOCKSS crawler to collect and preserve content; 2) Bepress has defined boundaries of LOCKSS-available content via a manifest page to include all Digital Commons repositories; and 3) the Stanford LOCKSS Team has written and tested a LOCKSS plugin for Digital Commons.

The remaining work to establish a Digital Commons PLN needs to be undertaken as a grassroots effort by institutions with Digital Commons instances. A minimum of seven institutions would need to join the PLN for it to be viable as a preservation network. To date, planning discussions about a PLN have occurred, but the eight institutions involved have yet to take concrete steps toward implementation. Costs to participating institutions would include staff time associated with getting the PLN off the ground, ongoing costs associated with PLN administration (aside from the costs involved in administering the individual LOCKSS boxes in the PLN), and the typical IT staff and hardware costs associated with administering a LOCKSS box. At this time it seems most likely that a Digital Commons PLN would require a LOCKSS instance at a participating institution that is architecturally separate from that institution's LOCKSS Alliance instance. Some staff-time savings might accrue if the same IT staff from an institution (e.g., in Cornell's case, staff in CUL-IT) administered its Digital Commons PLN LOCKSS instance and its LOCKSS Alliance LOCKSS instance.

While the LOCKSS decentralized model can be a strength in cases such as journal cessation, to our teamthe LOCKSS/Digital Commons experience demonstrates the weakness of the LOCKSS

decentralized model. Bepress and the Stanford LOCKSS Team have apparently taken the Digital Commons preservation effort as far as they are going to. The remaining work to establish a Digital Commons LOCKSS network is now up to the Digital Commons participating institutions. That work has yet to take place, most likely because competition for resources in participating institutions is currently high and the critical mass of effort required to launch a coordinated inter-institutional effort such as this exceeds the resources that are available.

Note: Columbia discontinued its use of Digital Commons in 2008, moving its institutional repository first to DSpace, then in late 2009 to Fedora.

8) Have we taken advantage of LOCKSS so far by gaining access to a canceled subscription or a closed journal? Have we participated in a failure-recovery test?

CRL reported in its LOCKSS audit that LOCKSS core functionality had been tested and validated through many anticipated failures such as network interruptions, data corruption, and LOCKSS box failures. In all instances known to the auditor, the LOCKSS boxes performed as expected and access to content was restored.

A scan of the LOCKSS technical mailing list from the beginning of the public LOCKSS Alliance in 2004 through October 2010 showed the withdrawal of one web-based reference book (not subscribed to by Cornell). On November 18, 2010, Cornell,Columbia and all LOCKSS Alliance members received notification that twelve titles would be withdrawn from their current publishing platforms, two from HighWire and ten from SpringerLink, at the end of 2010. Interestingly, when library IT staff at both Columbia and Cornell began to add the twelve titles to their LOCKSS box caches, they found that their disks were full. Both turned to the Stanford LOCKSS staff for assistance, because of the full disks and because both institutions were due to migrate their LOCKSS boxes. LOCKSS staff helped Columbia move stored content and ingest the 12 titles by the end of 2010 deadline.

LOCKSS staff tried unsuccessfully to help Cornell resolve a hardware issue with its old storage disk, so LOCKSS and Cornell set up a temporary LOCKSS server to store the 12 titles before the deadline. They then migrated that content back to Cornell’s permanent LOCKSS box once the hardware issue was resolved.

Neither institution yet has a plan in place for serving the 12 titles should they not be available from the new publishers.

In February 2005, Stanford programmers served up a page they called the "LOCKSS Card 2005," to be used in a demonstration of how LOCKSS works. Many LOCKSS participants preserved the page in their local LOCKSS boxes. At the end of March 2005, Stanford took the page down, simulating a failed journal. The LOCKSS boxes performed as promised by serving the site from the local boxes instead of from the original location at Stanford. Although LOCKSS worked as it should, in order to demonstrate the test, staff at Cornell had to simulate a link from the Voyager catalog to the page, because its LOCKSS box is not in the network pathway between its proxy server and the outside world. The address of the LOCKSS Card was put into both Cornell's and Columbia's proxy servers for testing purposes.

As noted in our response to Question 5 above, Cornell's and Columbia's de facto decision has been not to serve lost journals from their LOCKSS boxes themselves. This approach had been predicated on the expectation that journal failure would occur slowly enough for each institution to replace the original journal with the copy preserved in its LOCKSS box, to serve it from a local production server, and to point to it directly from the catalog. In this way, the library would avoid the need to address concerns about reconfiguring the catalog to serve individual volumes, adding all the preserved volumes' addresses to its proxy server, and maintaining a LOCKSS server and associated storage with sufficient CPU capacity, availability, and redundancy to allow the routing of traffic through its LOCKSS configuration without creating a single point of failure.

Contrary to this de facto "re-route" strategy (i.e., transferring selected content from a LOCKSS box to another location for access purposes after a trigger event), however, our team learned from Vicky Reich, Director of the Stanford LOCKSS Program, that such a strategy violates the Alliance's agreement with publishers. In other words, Alliance members do not have the legal right to move content out of the LOCKSS system. Therefore, the two CULs will need to address their concerns about catalog configuration, content proxying, and server/storage vulnerability by adopting another strategy for activating content preserved in LOCKSS.

Detailing such a strategy is beyond the scope of our assessment, so we recommend that CUL groups charged with e-resource preservation oversight undertake this task as soon as it is feasible do so. Indeed, this effort calls into question the CULs' ongoing commitment to LOCKSS participation. Thus, before embarking on an effort to detail a content activation strategy, we recommend that the CULs' leadership ask and answer fundamental questions about their institutions' LOCKSS participation such as: Are the CULs confident in LOCKSS as a preservation platform? Are the CULs able to demonstrate a level of commitment to LOCKSS that justifies ongoing reliance on the geographically (and institutionally) dispersed LOCKSS preservation model, which is ultimately only as strong as the combined commitment of its participants?

Though our team cannot recommend a detailed strategy for activating LOCKSS content as such, here is some relevant information that can serve as a starting point for the CUL groups investigating one. LOCKSS boxes are capable of both proxying content (at the publisher's original URL) and serving content (at a URL pointing to the LOCKSS box). A library that enables their LOCKSS box to proxy content need do no other configuration work. The publisher's URLs will continue to work. If the content is available from the publisher, it will be obtained from the publisher. If not it will be obtained from the LOCKSS box. A library that chooses to have its LOCKSS box serve content needs to configure its OpenURL resolver to point to the LOCKSS box for the volumes that are no longer available from the publisher. The LOCKSS Alliance has been collaborating with Ex Libris and Serials Solutions in the context of the KBART initiative (http://www.uksg.org/kbart) to allow LOCKSS to work with the SFX and 360 Link OpenURL resolvers. (To date, the LOCKSS Alliance has had no contact with Innovative Interfaces concerning its WebBridge OpenURL resolver that Cornell uses.) Configuring LOCKSS to work with an institution's OpenURL resolver does not require LOCKSS content to be proxied, and for load, fault-tolerance and other reasons is now the preferred technique. With regard to server capacity, a production-grade server is likely not

needed for LOCKSS because a single institution's traffic to specific pieces of older e-journal content at any given time should not be great enough to necessitate one. With regard to storage redundancy, the new Linux LOCKSS platform can accommodate RAID disks, so this appears to address one area of concern. (In fact, Columbia's new LOCKSS server uses RAID disks.) And, with regard to server availability, though many LOCKSS trigger-event scenarios are not likely to involve high-profile, high-use titles, the CULs' IT staff may still want to consider whether the lack of LOCKSS server fail-over and redundancy is a risk that the institutions should address.

9) Cornell and Columbia have received invitations to participate in the LOCKSS-USDOCS initiative (http://lockss-usdocs.stanford.edu/). Subsequent to receiving the invitations, the libraries have received input both pro and con regarding the initiative. Can we place the various perspectives in context? Given what we know about the initiative, can we make a recommendation regarding the two CULs' participation?

The two CUL university librarians received the initial invitations to join the initiative. The university librarians then circulated the invitations via email to other library administrators as well as to staff with known expertise in this area. These staff made inquiries about the initiative and reported back to correspondents on the email thread. The university librarians also received email input from administrators at other libraries who had received invitations. All these email threads were shared with our team.

The cost of LOCKSS-USDOCS participation for Cornell and Columbia would be relatively low. It would require no additional participation fee beyond the annual LOCKSS Alliance dues that the two institutions already pay. Participation requires a separate LOCKSS box and disk storage, which would amount to an investment in the low thousands of dollars for commodity hardware. Ongoing costs would parallel those for the LOCKSS Alliance boxes, and the ongoing staff time involved in managing two LOCKSS boxes would very likely be less than double the cost of managing one.

From what we have learned about the LOCKSS-USDOCS initiative and about preservation of U.S. government information generally, we have found it helpful to view the initiative in the context of two perspectives that have emerged regarding the preservation of U.S. documents. The first perspective holds that it is GPO's responsibility to archive U.S. documents and to enlist the aid of third parties such as NARA or the Hathi Trust to do so. Further, federal depository libraries, which already devote considerable resources in support of the Federal Depository Library Program, should not have to step in to provide this function. Cost concerns are particularly acute for regional depository libraries because their investments are greater and their ability to cut FDLP-related costs is less than for selective repositories. (Neither Columbia or Cornell are regional depositories.) An example of this perspective is the October 2010 ARL position on digital preservation and the FDLP: "Federal Depository Libraries are not required by law to provide long-term storage for digital Federal documents. GPO should identify and have certified one or more trusted third party repositories that are not part of the Federal government for preservation of and, when necessary, access to digital Federal documents." NARA is not currently involved in preservation of digital U.S. documents. Nor is the Hathi Trust though we have learned through informal sources that depository libraries in the CIC that are Hathi Trust

members are coordinating the digitization of subsets of their print U.S. document collections, which will allow them to withdraw print copies selectively.

The second perspective contends that it is very much within the mission of depository libraries to preserve digital U.S. government documents, and that this is in fact an extension of the role that depository libraries have always played with the print. documents. In addition, depository libraries have a responsibility to participatein ensuring that digital U.S. documents are widely available for the long term because there is no guarantee that the U.S. government will provide the resources to support this effort centrally. An example of this perspective is stated quite strongly in Ithaka S+R's *Modeling a Sustainable Future for the Federal Depository Library Program in the 21st Century: Recommended Direction*, a draft report commissioned by GPO and released for public comment on 4 Feb 2011: "Due to the critical importance of these materials, relying exclusively on GPO (or any government agency, for that matter) to preserve digital FDLP materials constitutes an unacceptable risk to the long-term survival of these materials. Multiple copies of digital FDLP materials must also be preserved independently of GPO according to community best practices, to provide critical assurance of their long-term availability to address user needs" (p. 5). (Note: GPO has retained Ithaka S+R to lead a project (http://fdlpmodeling.net) to develop a model for the Federal Depository Library Program (FDLP) to more efficiently accomplish its mission in a digital environment. Ithaka S+R's final report on this effort is due in March 2011, in time for the April 2011 Federal Depository Library Council Meeting .) The LOCKSS-USDOCS initiative falls directly in line with the second perspective and, indeed, the rationale for the initiative mirrors the concern expressed in the Ithaka S+R recommendation.

Our team consulted U.S. documents selectors from the two CULs concerning the initiative. These selectors share concerns about leaving the preservation of U.S. documents to GPO and ultimately to federal government funding. The selectors note the relatively low cost of LOCKSS-USDOCS participation and one selector pointed out that because Cornell and Columbia are selective depositories at least some of the cost of participating in LOCKSS-USDOCS could be offset by reducing ongoing costs associated with managing print U.S. document collections, for example, by tightening selection profiles or by selectively withdrawing print documents.

With regard to a recommendation from our team on LOCKSS-USDOCS participation, we find the last-mentioned selector's view compelling and in line with best practices of libraries that have considered LOCKSS or Portico participation in the overall context of collection management strategies, operations, and costs. We think that the CULs should consider LOCKSS-USDOCS participation in the context of their participation in the FDLP generally. Just as the CIC depository libraries are using availability of digital U.S. documents as an opportunity to reduce the cost of managing print U.S. document collections, so could the CULs audit and reduce their print document collections in light of the availability of digital U.S. document content. Regardless of the decision concerning LOCKSS-USDOCS participation, U.S. document preservation-related costs should be considered part of FDLP participation costs rather than as stand-alone initiatives or programs.

Finally, we think there is no reason for the CULs to make a rush decision on LOCKSS-USDOCS participation . The initiative already has enough participants for the LOCKSS-USDOCS PLN to

be viable as a preservation network. The Ithaka S+R report to GPO is due in March 2011. Depository library representatives will discuss the Ithaka report at the April 2011 depository council meeting. We think there is no risk for the CULs to wait to make a decision until after GPO and the FDLP community formally receive and react to the Ithaka report.

As an example of community response to the Ithaka draft report, see the Free Government Information (FGI) group's comments at http://freegovinfo.info/node/3193. Note that Stanford's James R. Jacobs is a co-founder of FGI. Clearly opinions vary regarding FDLP libraries' roles in digital preservation of U.S. documents, which our team feels argues for a wait-and-see approach to LOCKSS-USDOCS participation.

[Note (August 2011): The LOCKSS Team's response to this question was written in early March 2011, before Ithaka S+R submitted its report on the FDLP to GPO. On August 5, 2011, the following statement concerning the Ithaka report was posted on the *FDLP Desktop* site:

> In September 2010, the U.S. Government Printing Office (GPO) contracted with Ithaka S + R (Ithaka) to develop practical and sustainable models for the Federal Depository Library Program (FDLP) to continue to fulfill its mission in a changing information environment now dominated by digital technology. These models were intended to serve as a guide in planning the future direction of the Program. After careful review it was determined that the models presented by Ithaka are not practical and sustainable to meet the mission, goals, and principles of the FDLP. These models have some value as we move forward together with the library community to develop new models based on a shared vision which will increase flexibility for member libraries and ensure the vibrant future of the Program in the digital age. (http://www.fdlp.gov/component/content/article/184-gpoprojects/1006-future-direction-of-the-fdlp; accessed August 22, 2011)]

10) Cornell has submitted its electronic and print serial holdings data to Portico and has received reports from Portico about Portico's coverage of its holdings. (Columbia's Portico analysis is forthcoming.) Cornell also has data about LOCKSS' coverage of its serial holdings. Can we do an analysis that compares Portico and LOCKSS coverage? And given a likely similarity between Columbia and Cornell's serial holdings, can we illuminate the Cornell Portico/LOCKSS data in ways that might guide the two CULs' ongoing participation in LOCKSS and Portico?

(2CUL and our team owe Cornell's Jim Spear our gratitude for his yeoman's work on this analysis. He showed great skill and patience as he compiled his findings and explained them to us.)

We want to preface our comparative analysis of LOCKSS and Portico coverage of Cornell e-journals by mentioning this cautionary opening line from Portico's report on its analysis of Portico coverage of Cornell's journals: "Journal holdings are complicated."

With that warning firmly in mind, we begin with general statements concerning the boundaries of the analysis: 1) First, the analysis covers e-journal holdings only. Portico did supply coverage

numbers for Cornell print journals, but we don't have similar numbers for LOCKSS coverage of print journals, nor did the Portico print numbers account for holdings in both electronic and print formats. 2) Second, the analysis is limited to matches of Cornell e-journal holdings with LOCKSS and Portico holdings for e-journals with valid ISSN numbers, valid eISSN numbers, or both. Titles preserved in LOCKSS and Portico are almost exclusively limited to titles that have one or both of these identifiers. (98+% of ~6,600 titles in LOCKSS and 99+% of ~12,000 titles in Portico.) Though the analysis is restricted to titles with standard serial identifiers, such identifiers are found in only 50% of Cornell e-journal records. We did some random sampling of Cornell e-journals without identifiers to see whether LOCKSS and Portico covered them. They did not, which isn't surprising given that so few titles in LOCKSS and Portico (just over 100 titles in each) lack identifiers. 3) Finally, journal holdings figures are a moving target. Journal counts change monthly given Cornell's use of batch loads of e-journal records from its record supplier. Counting holdings and determining matches are further complicated by the occurrences of both ISSN and eISSN numbers in some records. .

Given these caveats, we found:

- At the time of thisanalysis, Cornell held 45,602 e-journal titles with an ISSN, eISSN, or both.
- Of those 45,602 titles, LOCKSS preserves 5,245 titles or 11.5%.
- Of those 45,602 titles, Portico preserves 10,114 titles or 22.2%.
- Of the 45,602, both LOCKSS and Portico preserve 3,476 or 7.6%.
- Given the overlap of 3,476 titles, LOCKSS uniquely preserves 1,769 titles or 3.9%.
- Given the overlap of 3,476 titles, Portico uniquely preserves 6,638 titles or 14.5%.
- LOCKSS and Portico combined preserve 11,883 titles, or 26.1%, of Cornell's 45,602 e-journal titles with an ISSN, eISSN, or both.

Subsequent comparison of Columbia’s e-journal holdings with Portico revealed similar results: of the e-journal titles with an ISSN or eISSN, 17% are preserved in Portico. Also, we can share some preliminary information on the types of Cornell and Columbia e-journals with ISSNs or eISSNs that did not match Portico holdings. These titles roughly break out into the following categories (which we acknowledge are in some cases arbitrary and in others not necessarily mutually exclusive):

- Available through aggregators: 25-30%
- Miscellaneous freely accessible: 22-25%
- Newsletters: 10%
- East Asian: 10%
- Publishers who otherwise participate in Portico: 8-9%
- Non-participating publishers: 4-5%
- Digitized collections with e-journals (commercial): 5%
- Digitized collections, library based (e.g. Hathi Trust): 4%
- Government, IGO (e.g. OECD): 3-4%
- Book series, conference proceedings: 2-3%
- Data errors (e.g., ISSN mismatch: 2%

Regarding implications of these findings for the two CULs' ongoing participation in LOCKSS and Portico, we note, as did the London School of Economics, that while there is overlap in coverage between LOCKSS and Porticoboth services nevertheless preserve titles uniquely. We also echo LSE's observation that neither service preserves a large percentage of our e-journal holdings. If one were to factor in the 50% of Cornell's e-journal holdings that lack standard identifiers, LOCKSS and Portico combine to preserve only about 13% of Cornell's e-journals. Given this lack of preservation coverage overall, we think it is important for the CULs to do what they can to improve the state of e-journal preservation as a whole. With this coverage analysis as a baseline, we recommend that the CULs track LOCKSS and Portico coverage prospectively in order to hold ourselves and the services we use accountable for preserving e-journal content.

Further, it is important to note that while these findings reflect a quantitative analysis of e-journal preservation coveragethey do not reflect any qualitative analysis of the importance to the CULs' research communities of preserving the particular titles that LOCKSS and Portico preserve., A preliminary review of the titles suggests to us that both Portico and LOCKSS do indeed preserve publisher and society titles that are important to the two CULs' users, but transforming such subjective impressions into a meaningful qualitative analysis is beyond the scope of our charge. It does, nevertheless, merit more investigation.

11) With the Portico/LOCKSS data in hand, do we know whether there is a significant body of material that Portico is not expected to cover, and whether these titles would be viable candidates for enrollment in LOCKSS? If so, what effort would be required?

We are not able to do this kind of gap analysis at this point. More study of local holdings for publishers not covered by LOCKSS or Portico would be required, as would study of Portico's preservation patterns. This would certainly be a worthwhile undertaking and we recommend that CUL sponsors hand this assignment to the CUL group(s) tasked with ongoing e-resource preservation responsibilities. Regarding the effort required, given that such a small percentage of e-journal content is currently preserved, the demand for e-journal preservation effort far exceeds the resources that either or both of the CULs could supply. We suggest that the CULs allocate some e-journal preservation effort to work such as this using the best tools available.

12) Would it be possible for the two CULs to share a single LOCKSS box? If so, can we identify any risks or benefits associated with such an approach?

According to Vicky Reich, the two CULs sharing one LOCKSS box would only be technically possible if the CULs' current e-journal subscriptions and e-journal backfile holdings were identical. This is the case because the LOCKSS system determines what an institution is entitled to access based on the IP address of the LOCKSS box. For example, if Columbia were to maintain a LOCKSS box on behalf of Cornell and itself, that box would use its Columbia IP to determine the content it was authorized to serve. If Columbia's and Cornell's e-journal holdings were the same, the box would ostensibly serve only content which both institutions were entitled to serve. That said, Vicky Reich expressed concern that such an approach could erode publisher commitments to LOCKSS because of potential backfile variations between institutions.

13) Given what we know about how the two CULs have approached LOCKSS participation to date and about how other libraries and library groups have participated in LOCKSS, can we make any recommendations regarding how best to position LOCKSS-related decision making within the CULs prospectively?

Our experience suggests that collection development, technical services, and library IT staff skills all come into play in decisions related to e-journal preservation via LOCKSS. Of these three groups, serials/e-resource staff appear to be most intimate with the intricacies of the data that would drive additional e-journal preservation efforts. That said, using data to make decisions about where to focus e-journal preservation efforts would fall to collection development staff. Thus a partnership between serials/e-resources and collection development, with involvement from IT staff from the outset as well, seems like a viable scenario. Having said this, we are concerned that this work could easily become an unfunded mandate for all units involved. If CUL groups are tasked with expanded roles in e-journal preservation, we recommend that staff time be formally reallocated to this work in the context of reducing other commitments and responsibilities. Finally, because the approach to oversight we envision would divide responsibility for e-journal preservation among organizationally distinct units, we think it is important that e-resource preservation have a clearly designated sponsor in library upper administration who can provide direction and advocate for support.

14) By way of a conclusion and overall recommendations, what is the value proposition for our two institutions in maintaining our LOCKSS memberships? Overall, do our present levels of engagement make the best use of our investments in e-journal preservation, and if not, what else could the two CULs be doing to get better value from them?

By inertia more than design, both Columbia and Cornell have to date used a minimum-cost, just-in-time approach to their participation in the LOCKSS Alliance. The two institutions have maintained their LOCKSS boxes via a least-effort, dark-archive strategy and only minimally, if at all, have they integrated preservation considerations via LOCKSS or Portico in e-journal license management or in e-journal collection management as a whole. This strategy may be justified in the near term in order to keep costs down, but the approach masks and defers costs that would be required to1) provide library access to LOCKSS-held content should publishers cease to provide access, 2) better understand the coverage in LOCKSS and Portico of library e-journal holdings, and 3) pursue preservation of e-journal content not currently preserved in either LOCKSS or Portico. Our responses to Questions 8, 10, 11, and 13 point to ways in which the CULs could make better use of their e-journal preservation investments.

We end our report by recalling our response to Question 10, in which we noted that LOCKSS and Portico combine to preserve roughly 26% of Cornell e-journals with standard identifiers and roughly 13% of all Cornell's e-journals. This overall lack of publisher participation in either of the leading e-journal preservation programs offers the two CULs an opportunity to use their individual or combined influence with publishers, to whom they pay substantial licensing fees, to improve the state of e-journal preservation as a whole.

References

- Association of Research Libraries, ARL Statement of Principles on the Federal Depository Library Program (October 2010) http://www.arl.org/bm~doc/fdlpprinciples14oct10.pdf
- Burnhill, Peter and Guy, Fred (2010) 'Piloting an E-journals Preservation Registry Service (PEPRS)', *The Serials Librarian*, 58: 1, 117-126. http://dx.doi.org/10.1080/03615261003622742; (accessed on November 19, 2010)
- Center for Research Libraries, Auditing and Certification of Digital Archives Project, *LOCKSS Audit Report* (November 2007) http://www.crl.edu/sites/default/files/attachments/pages/LOCKSS_Audit_Report_11-07.pdf
- *Digital Commons Private LOCKSS Network Planning* (closed wiki page made available to our team; accessed December 13, 2010)
- *E-journal archiving for UK HE libraries: a draft white paper* (JISC Consultation Draft 1 Oct 2010) http://www.jisc.ac.uk/whatwedo/programmes/preservation/2010ejournalwhitepaper.htm.
- *Ensuring that 'e' doesn't mean ephemeral: a practical guide to e-journal archiving solutions* (Version 1.1, Feb 2010) http://www.jisc-collections.ac.uk/E-journal-archiving-solutions/
- Ithaka S+R, *Modeling a Sustainable Future for the Federal Depository Library Program in the 21st Century: Recommended Direction* (Public Draft for Comment 4 Feb 2011) http://fdlpmodeling.net/wp-content/uploads/2011/02/FDLP-Direction-Draft-2-4-2011.pdf
- LOCKSS: Selecting and Building the Collection. http://lockss.stanford.edu/lockss/Selecting_and_Building_the_Collection (accessed on November 29, 2010)
- *[NYU] LOCKSS Task Force--Final Report*, August 4, 2006 (offline copy made available to our team)
- Reich, Vicky (February 23, 2011 phone call with Marty Kurth)
- Seadle, Michael (2011) " Archiving in the networked world: by the numbers", *Library Hi Tech*, 29:1, 189 - 197. http://dx.doi.org/10.1108/07378831111117001 (accessed on March 14, 2011)