

April 12, 2007

I. Introduction

The mission of the Government Printing Office (GPO) is to provide permanent public access to official Federal Government publications in print and electronic formats through the Federal Depository Library Program (FDLP). In addition, the International Exchange Service (IES) operates according to 44USC1719 which states, “there shall be supplied to the Superintendent of Documents...all Government publications...for distribution to those foreign governments which agree, as indicated by the Library of Congress, to send to the United States similar publications of their governments for delivery to the Library of Congress.” Various distribution systems (e.g. tangible depository distribution, *GPO Access*, etc.) exist to facilitate dissemination via the FDLP and IES. A federated dissemination system exists for the management and distribution of Government journals in print format, but a similar federated system does not exist for journals in electronic format (e-journals).

While GPO’s mission includes permanent public access, many Federal agency publishers’ missions do not. In order to make room for new issues and volumes, agency publishers often overwrite or remove old e-journal content. This practice frequently leads to the disappearance of Government e-journals from the Web.

GPO has received requests from research institutions, universities, depository libraries, and other Federal Government agencies to investigate using LOCKSS (Lots of Copies Keep Stuff Safe) in conjunction with existing harvesting technologies as a means to collect, manage, disseminate, and safeguard access to Web-based Federal Government e-journals that are within the scope of the FDLP.

As agency publications are becoming increasingly Web-based, LOCKSS presented a unique opportunity for GPO to investigate digital distribution. GPO agreed to work with Stanford and the participating FLDP partner libraries, to test the LOCKSS technology as a potential precursor to GPO’s Future Digital System (FDsys).

GPO, Stanford, and the participating FDLP libraries partnered to conduct the test pilot. GPO then analyzed the LOCKSS technology and compiled this report of findings and future recommendations, thus completing the pilot project.

II. Pilot Description

A. LOCKSS Technology

LOCKSS is a decentralized system of persistent digital caches of HTTP delivered content (e.g., HTML, JPEG, PDF, WAV, etc.). A LOCKSS machine runs on free open-source software that is supported by Stanford University. A LOCKSS cache collects content and metadata by slowly crawling a publisher's Web site and harvesting static content. At least six caches must crawl and harvest the same content. In order to maintain content integrity, caches in a LOCKSS system engage in a mutual audit and repair mechanism operated through a peer-to-peer sampled voting protocol.

LOCKSS uses a distributed repository model for conservation of e-journal content. In the online environment, libraries typically pay for access to electronic journals, in effect leasing content from publishers. Libraries lose access to back issues when they cancel a subscription or when a publisher revokes access. When a library cancels a subscription to an e-journal, they lose access to current and back issues. By caching journal content as it is published, the LOCKSS model allows for local ownership of e-journal content.

LOCKSS provides robustness through redundancy. Multiple installations crawling and auditing the same content means there is no single point of failure. The auditing process maintains content integrity by repairing damaged content.

LOCKSS runs on inexpensive consumer hardware, keeping costs of installations low. Presently, one PC can hold approximately 3000 e-journal years of content. Costs are shared widely among LOCKSS institutions, reducing the costs to individual institutions. System administration requirements of maintaining a LOCKSS cache are low. Once the LOCKSS software and plug-ins are loaded, the LOCKSS crawls automatically with little or no intervention.

B. Goals and Objectives of Pilot

Phase 1 Goal: Establish and test a LOCKSS cache at GPO.

Objectives:

- Evaluate the security of a LOCKSS cache at GPO.
- Address any security problems revealed by the security test and reevaluate security as many times as necessary until the system is deemed secure by GPO IT.
- Establish a LOCKSS cache at GPO.
- Crawl and harvest two freely available e-journals for one month.

Phase 2 Goal: Make Federal Government e-journals available to select Federal depository libraries (FDLs) and International Exchange Service (IES) pilot partner libraries that are operating LOCKSS caches.

Objectives:

- Compile a list of Government journals that are in-scope for the Federal Depository Library Program (FDLP), distributed in tangible (print) format via the IES, and are available in electronic format via Federal agency Web sites.

- Manually harvest ten e-journals from Federal agency Web sites.
- Add the e-journals to a secure LOCKSS Web directory on the Federal Depository Library Program Electronic Collection (FDLP/EC) Archive server <<http://permanent.access.gpo.gov>>.
- Create a publisher manifest for each e-journal and add the manifests to the secure LOCKSS Web directory.
- Develop a plug-in for each e-journal and coordinate with Stanford on the testing and dissemination of each e-journal plug-in.
- Establish IP based authentication to allow pilot partner LOCKSS caches to have access to the e-journals in the secure LOCKSS Web directory.
- Develop quality control (QC) policies and procedures to ensure that the content in the secure LOCKSS Web directory is identical to the source content on agency publisher Web sites.
- For one year, coordinate with FDL and IES pilot partners to use LOCKSS to crawl and harvest e-journals from the secure LOCKSS Web directory on GPO's FDLP/EC Archive server.
- Execute real world scenarios and evaluate the pilot.

C. Pilot Scope

The scope of this project includes Federal Government e-journals that are within the scope of the Federal Depository Library Program (FDLP) and disseminated in tangible format via the IES.

D. Pilot Assumptions

- GPO will be able to establish a secure cache within GPO's internal network infrastructure.
- At least five pilot library partners will be interested in participating in GPO's LOCKSS pilot.
- Stanford will continue to support the LOCKSS software.
- The Superintendent of Documents' current harvesting workflow (manual harvesting) will accommodate the harvesting required for the LOCKSS project.

E. Pilot Risks

- LOCKSS could pose a security threat to GPO's network.
- There could be discrepancies between content on agency Web sites, in the FDLP/EC Archive, and in the secure LOCKSS directory.
- The demand for the LOCKSS pilot could exceed the resources available to harvest and oversee the quality control of e-journal publications.

F. Approach to Pilot

The LOCKSS pilot project was divided into two phases. The purpose of the first phase was to test the security and network integration of a LOCKSS cache at GPO with the goal of establishing a LOCKSS server on GPO's network. The purpose of the second phase was to harvest Government e-journals from Federal agency Web sites and add the e-journals to a secure directory on the FDLP/EC Archive server. GPO and the pilot partners then used LOCKSS to crawl and harvest

the e-journals from GPO. GPO and the pilot partners executed real world scenarios to help assess the effects of using LOCKSS as a FDLP/IES dissemination tool.

III. Analysis

The following analysis examines the processes used for harvesting content and the creation of plug-ins and the LOCKSS cache.

A. Harvesting e-Journals from Federal Agency Web Sites

Requirements

- E-journal content must be harvested from source agency Web sites.
- Harvested content directory structures must be modified to accommodate plug-in development.
- A Publisher Manifest Web page conforming to LOCKSS specifications must be created for each archival unit.
- Link integrity must be verified for each issue of harvested content. Metadata must be added to harvested content identifying it as a harvested copy.
- Harvested content in the modified directory structure must be available on a Web server.

Pilot Process

Manually Harvest E-Journals

As part of the GPO LOCKSS Pilot Project, GPO harvested the 2005 Archival Unit (volume) of ten Government e-journals from ten Federal agency Web sites.

GPO did not obtain prior permission from the agencies to harvest the e-journals because all e-journals were in the public domain, and the following criteria were used to select titles for inclusion in the pilot:

- Disseminated via the FDLP and IES
- Selected by over 600 Federal depository libraries
- Available in electronic format from Federal agency Web sites.

GPO manually harvested the pilot e-journals issues. Manual harvesting helped GPO identify various characteristics of the e-journals, such as boundary issues or what was included in the e-journal and what was not included. It also facilitated GPO's ability to modify a local copy of the e-journals' directory structure as harvesting was taking place. Furthermore, it allowed GPO to examine each e-journal's interface to identify e-journal content versus presentation.

A chart, accessible via Appendix A, provides information about the number of issues, harvested formats, and size of the 2005 archival unit for each pilot e-journal.

Modify E-Journal Directory Structure

In order to store the harvested content in the modified directory structure, GPO created a local or internal development site for the project that contained separate folders for each individual e-journal, each archival unit, and each e-journal issue.

Create Publisher Manifest Web Page

It was necessary for GPO to create a Publisher Manifest for each of the ten harvested e-journals. A LOCKSS Publisher Manifest page provides a set of links that the crawler will follow to find content to collect and a permission statement saying that the LOCKSS system has permission to collect and safeguard access to the content found by following the links. An example of a Publisher Manifest page for the GPO LOCKSS pilot project can be viewed via Appendix B.

Adjust and Validate Links within Individual E-Journal Issues

HTML links in harvested content were manually adjusted via DreamWeaver MX 2004 software to verify that link integrity had been maintained within each individual issue.

Add Metadata to Harvested e-Journals

The harvesting process did not include the acquisition of descriptive metadata from source agency Web sites or GPO's ILS. Despite this, after the e-journals were harvested, GPO added the following metadata to the HTML and PDF files: "This copy was downloaded from the agency's Web site by the U.S. Government Printing Office under Title 44 USC, MM/DD/YYYY. External links, forms, and search boxes may not function within this collection."

In an early "Real World Scenario," GPO worked with Stanford to add metadata to issues of an e-journal (i.e. *Treasury Bulletin*) that had already been harvested from GPO's Web server by the pilot partner LOCKSS caches. The caches detected that a new "version" of the content was available from GPO and it was added to pilot partner caches along with the previous version.

Establish a LOCKSS Directory on a Web Server

GPO established a secure LOCKSS directory on a publicly accessible Web server at <http://permanent.access.gpo.gov/lockss/>. The LOCKSS directory on the publicly accessible Web server mirrored the LOCKSS directory on the development server, and the development server was used to update the live publicly available server.

Analysis

One of the most challenging and time consuming aspects of the pilot was the manual harvesting process. Prior to harvesting an archival unit, each e-journal was individually analyzed in order to make decisions related to how e-journal boundaries, formats, and Web interfaces would be impacted by harvesting. Determining the boundary for each issue was challenging because many e-journals linked to supplemental content that was not part of the official printed version. Almost all e-journals were available in multiple formats, and when feasible, all formats were harvested. For some e-journals with complex Web interfaces and navigational elements, it was not feasible to manually harvest the entire issue. Instead, a PDF was harvested for each issue, and when available, "Web Only Updates" were harvested in HTML.

GPO selected a mixture of basic and complex e-journals for the pilot. Characteristics of a basic e-journal include individual articles linked off of an HTML index page, limited file formats, or individual issues contained in a single file. Characteristics of a complex e-journal are a robust Web interface, complex directory structure, or multiple file formats. Appendix C provides time averages for harvesting tasks performed by a GS-12 with intermediate Web development skills.

Outstanding Issues

Policy

- GPO needs to be assured that content can be removed from caches in the event of an agency recall, in accordance with SOD 110. Additional work needs to be done to streamline content removal to enable efficient removal of recalled content.
- IP authentication for over 1260 depositories would be cumbersome, and may not be cost effective in relation to the benefit received.

Technical

- Automated harvesting tool that will harvest entire contents of archival unit regardless of structure of source agency Web site.
- Automated harvesting tool that will capture any associated descriptive and technical metadata from source agency Web site.

Recommendations

- Explore the effect of using automated harvesting tools on the plug-in creation process.
- Explore automated methods to add metadata to harvested content that enables it to be easily identified as a harvested copy.
- Explore partnerships with agencies to add mandatory LOCKSS metadata to Web sites so Federal depository libraries could use LOCKSS to crawl and harvest directly from agency sites.
- Explore using tools such as Google sitemaps with open APIs for harvesting rules and instructions.
- Explore other user authentication options or consider making content available without user authentication.

B. LOCKSS Plug-in Development and Distribution

Requirements

- Harvested content must be reformatted into packaged structure.
- Harvested content directory structures must be reformatted to Journal Name with a Journal year/number subdirectory.
- Plug-in must conform to valid XML structure.
- Plug-in must be verified by Stanford.

Pilot Process

In order to tell the pilot partner LOCKSS boxes to harvest content from the secure LOCKSS directory on the GPO Web server at specific intervals, it was necessary for GPO to create an XML based plug-in for each e-journal archival unit (e.g. 2005 archival unit). GPO used Stanford's plug-in development tool to create the plug-ins. The plug-in software can be found at <http://www.lockss.org/plugin/plugin-tool.html>. Appendix D describes the steps for the *Amber Waves* journal plug-in to be uploaded and checked on <http://permanent.access.gpo.gov/lockss/>, and they are consistent with Stanford's example.

Analysis

Plug-in Development

There are two ways to implement the plug-in in order to achieve the same results for the LOCKSS Pilot. In the first option, GPO could harvest the content for a journal, add metadata, and place it directly on GPO's Permanent server in its current structure. Once there, the plug-in would be written to conform to the directory structure created by the originating agency.

The second option still requires GPO to harvest the journal and add the metadata. The journal's directory layout would be changed in order to fit all related files in a subdirectory named for the journal's number or year, with a parent directory named for the journal. Although it would be significant work to harvest and reformat the journal content and update links appropriately, it would drastically simplify the plug-in writing process. If all of the journal content was contained within a package, the plug-in could be limited to the directory for that journal, making the contents of plug-in only a few short rules. All plug-ins were created using the plug-in tool provided by Stanford, making the process much quicker than creating the XML structure manually. Once defined, the development process took approximately three hours.

Plug-in Testing

The crawl rules were tested using the Test CrawlRules option in the plug-in tool, which simulates how the crawler would crawl through the journal. The results Test CrawlRules shows which files were fetched, and whether those files were included or excluded based on the rules from the plug-in. GPO would test the crawl rules twice – once on a local copy of the harvested journal, and again after it was put on the live server on Permanent.

Plug-in Distribution

Once a plug-in had been written and tested by the plug-in tool, GPO sent the XML file to Stanford. Stanford would verify that the plug-in was valid and awaited GPO's determination for a release date. Overall, the process for distribution was transparent to GPO. Pilot partner LOCKSS boxes automatically downloaded the plug-in from Stanford. This action also initiated the initial harvest and subsequent crawls of the e-journal issues that were available in the secure LOCKSS directory on the GPO Web server.

Plug-in Technical Support

The LOCKSS Team at Stanford was responsible for all technical support queries from GPO. Stanford was able to address all of GPO's issues and questions associated with setting up the Pilot quickly and expertly.

Outstanding Issues

- Automated harvesting would make the process of capturing a journal much easier, but it must be determined if the automated harvesting would restructure the journal content, or if a more complex set of crawl rules would need to be written.

Recommendations

- Explore writing plug-ins for non-harvested journals.
- Explore writing plug-ins for auto-harvested journals.
- Explore methods for determining journals to harvest into LOCKSS cache.

C. LOCKSS Cache Set-up and Operation (GPO and Partners)

Running on standard desktop hardware and requiring almost no technical administration, LOCKSS creates a low-cost, persistent, accessible copy of e-journal content as it is published. Accuracy and completeness of LOCKSS appliances is assured through a robust and secure, peer-to-peer polling and reputation system.

Requirements

- Enough CPU and memory. 1GHz VIA CPUs, a 2.4GHz Celeron is lavish. 512MB is ample memory.
- CD drive and optionally either:
 - A floppy disk drive
 - A USB flash memory drive with hardware write-protect switch
- Enough disk. 250GB is enough to get started. The current CD supports both parallel ATA (PATA) drives and serial ATA (SATA) in native mode.

Pilot Process

Creating Publisher access directory:

1. Create a secure folder on the permanent.access.gpo.gov server to house the e-journals selected for the GPO LOCKSS pilot project.
2. Provide IP authentication to allow access to IP addresses approved by GPO to the secured directory.
3. Request/receive IP addresses for all pilot project participants.
4. Submit request to add IP addresses to access list.

GPO/IT&S Security Scan on the LOCKSS cache:

1. Complete LOCKSS sensitivity assessment questionnaires from IT&S
 - a. GPO IT Security Program sensitivity assessment questionnaire (Part I)
 - b. GPO IT Security Program, sensitivity assessment questionnaire (Part II application)
2. Configuration of the LOCKSS cache
 - a. The LOCKSS OS is a downloadable CD-ROM that provides the software and OS.
 - b. GPO specific IT data elements were required (DNS, IP address, gateway, firewall port openings, and cache deployment in the DMZ).
3. Security scan of the LOCKSS software
 - a. Required setting the LOCKSS cache on a network platform outside of the GPO main network
 - b. Temporary IP and DNS information were configured to perform the IT security scan
4. Deployment of the LOCKSS software
 - a. The security cleared low-end PC was placed in the DMZ.
 - b. The GPO LOCKSS system was added to the GPO server farm for deployment.
5. Operation
 - a. Perform routine monitoring of the LOCKSS cache

Analysis

The Network Vulnerability Assessment Summary Report explains how susceptible the organization could be to an attack based on the number and the severity (or risk level) of vulnerabilities detected by Internet Scanner after scanning the network.

The security scan report identified:

- One vulnerability
- Security risk as low

The conclusion of the analysis is that the supported operation enables the system to recover from several events that cause data loss or damage.

Outstanding Issues

- The technical sustainability and longevity of the LOCKSS platform as a long-term archiving solution needs to be assessed.
- Are there any potential long-term risks? Is the approach of packaging LOCKSS with its own operating system environment (OpenBSD) sustainable and viable in the long term?
- How might the expertise of the current LOCKSS development team be maintained and disseminated to the users' community after the termination of the pilot project?
- Administration over HTTP:
We've built the Web UI with the premise that nothing that is hard to reverse should be possible using it. The most severe thing that can be done is "deleting" an AU, which actually just unconfigures it (leaving all the content intact). Re-adding that AU will restore everything. Permanently reconfiguring the machine or changing any passwords all require physical access to the machine. Permanently deleting content requires at least root SSH access to the machine (and this is firewalled off except to a controlled set of IP addresses). The issues of Confidential, Integrity, and Availability remain until the application UI is moved to SSL, and it is a possible enhancement in the next development cycle of future improvements.
- Inability to push system logs to a log server:
Enabling the system to send a SYSLOG message is a critical element of auditing the system to ensure integrity of the data.

Recommendations

- Compare LOCKSS with other caching and replication technologies as a part of the appraisal.

D. Library Program Issues

Requirements

General

- E-journal content must be complete.
- Depository and IES libraries must be able to access the content of e-journals.
- Content must be available for permanent public access.
- Technology must provide a cost efficient distribution mechanism for electronic content.

FDLP

- Depository libraries must be able to make content available to users in multiple locations, including onsite users and remote users accessing content through library services.
- Depository libraries must be able to remove content through routine “weeding” or in response to agency recalls.
- Installation support must be provided to depository libraries.
- Ongoing technical support must be available to depository libraries.
- Depository libraries must be able to receive depository content distribute through LOCKSS free of charge.

IES

- The Library of Congress must be able to select which titles it wishes the IES libraries to receive via LOCKSS.
- IES libraries must be able to make back-up and local electronic copies of this material.
- Electronic copies must be furnished to the IES libraries no later than the day of publication of the print version.
- Installation support must be provided to the IES libraries
- Ongoing technical support must be available to the IES libraries.
- There must be the ability to withdraw material from the IES program.

Pilot Process

The pilot plan was to test the technology with a limited number of pilot partners accessing and caching a limited amount of content. Potential FDLP library partners were identified that met the following criteria:

- Library currently held depository status.
- Library had an existing LOCKSS installation.
- Library was a member of the LOCKSS Alliance.

Additionally, the Library of Congress identified three German libraries to invite to participate in the pilot, in order to test the technology for use in the IES program. Including the Library of Congress, 20 depository libraries agreed to participate in the pilot. In order to access content, partner libraries had to supply GPO with the IP addresses of their LOCKSS caches, as well as any administrative computers that they wished to use to view LOCKSS content, including pilot documentation. GPO granted read access to the LOCKSS directory on the Permanent server to the IP addresses supplied. GPO also supplied these IP addresses to Stanford so that Stanford could grant permission to access plug-ins for GPO LOCKSS titles on their site.

Three “real world scenarios” were included in the project plan to test issues that could arise in the real world and see how LOCKSS responded to them. Real World Scenario #1 involved removing content from GPO’s server and determining if libraries could access their cached content seamlessly. Topics for Real World Scenario #2 and Real World Scenario #3 would address any issues that arose during the pilot. For Real World Scenario #2, damaged content was placed on a partner’s cache and access to that cache was provided to other participants so that all could watch the content repair itself. Real World Scenario #3 was cancelled due to lack of testable ideas and to time constraints.

Analysis

Online versions of the e-journals selected for this pilot included complete content. However, in some cases, online content was altered when a new issue was released. If a title such as this were to be distributed through LOCKSS only, it would be important to harvest each issue at the time of release to ensure that monthly data were captured. A temporary backlog in harvesting these or similar titles would result in loss of content. Depository libraries participating in the pilot were able to access the content of the e-journals provided.

While content can be removed from a LOCKSS box, the process is complicated. GPO needs to be assured that content can be efficiently removed from LOCKSS boxes in accordance with SOD 110, in the case of an agency recall. More work would be required to streamline content removal to ensure that content can be removed efficiently in order to comply with a recall.

LOCKSS technology in itself appears to be relatively cost efficient as a distribution mechanism. Costs appear to be a bigger issue in relation to staff time required to harvest content, alter directory structures, write plug-ins, and administer IP authentication. Additionally, the cost of technical support is an issue. Stanford provides technical support to members of the LOCKSS Alliance, and GPO is not in a position to provide technical assistance to depository libraries in support of LOCKSS caches. If LOCKSS were to become the only distribution method for e-journals distributed to the FDLP, all depositories would have to join the LOCKSS Alliance.

LOCKSS technology could be used to enable the Library of Congress to select any title available electronically through the FDLP for inclusion in the IES program. However, titles outside the scope of the FDLP or titles with incomplete online content might not be available as GPO LOCKSS titles. IES libraries could make back-up copies of this material by establishing a backup cache.

As with FDLP libraries, GPO is not in a position to provide installation support or ongoing technical support to IES libraries. The Library of Congress could be required to provide technical support or make arrangements for another institution to do so on their behalf.

Outstanding Issues

Some libraries may want to “weed” publications from their caches in order to regain disk space at times. GPO requires the ability to remove content from caches in the event of an agency recall. Additional development work would be required to streamline removal of content to ensure that libraries and GPO can comply with recall requirements.

FDLP

- It is unclear what percentage of FDLP libraries want to utilize LOCKSS for e-journal content.
- It is not clear whether libraries that do want LOCKSS want it as an exclusive service to depositories, or whether they simply want to enable libraries to archive content locally.
- The staff time required to harvest content from agency Web sites and republish it on GPO servers may be prohibitive if it is a duplication of effort because FDLP libraries are split on whether they want e-journal content through LOCKSS.
- Many agencies will complain if they believe GPO is taking business away from their Web sites by republishing content on GPO Web sites.
- Some libraries may not have the technological capabilities to support LOCKSS.

- The ability to efficiently remove content is absolutely necessary.
- LOCKSS may have format migration issues similar to CD-ROMs and other tangible electronic media in depository libraries.

IES

- It is unclear if all the libraries that participate in the IES have the technological capabilities to support LOCKSS.
- Since the Library of Congress participated on behalf of the German libraries, the pilot was not able to determine what, if any, problems or difficulties might arise from the participation of foreign libraries.
- The providing technical support outside of the U.S. is an issue.
- There are concerns from IT Security regarding foreign libraries having access to a GPO server to download material.
- The Library of Congress has not surveyed the IES libraries to determine interest in possible participation in LOCKSS.
- It is unclear how many of the IES libraries may decide to participate in LOCKSS. Would it be acceptable for only a portion of the libraries to participate?
- Any changes to the IES program must be made in cooperation with the Library of Congress.
- Will the Library of Congress or GPO pay for IES membership in the LOCKSS Alliance if that is needed?

Recommendations

- Explore options for streamlining removal of content from LOCKSS boxes.
- Explore options for making content available from a single site that would allow LOCKSS libraries and non-LOCKSS libraries to access content from the same source. This would eliminate duplication of effort required to make content available to both groups of libraries.
- If GPO chooses to implement LOCKSS and require depository libraries to receive content through LOCKSS, options for a “consortium membership” to the LOCKSS Alliance should be investigated.

IV. Lessons Learned

The model used in this pilot would force GPO to choose between the following options:

- Duplicating effort to provide access to Federal e-journal content through LOCKSS for libraries that wish to use LOCKSS and separate access for libraries that do not wish to use it.
- Requiring all libraries to use LOCKSS for e-journal content.

The depository library community has not been surveyed to determine whether there is wide enough support for LOCKSS to use it as the only e-journal delivery mechanism to justify requiring all libraries to receive e-journal content through LOCKSS. It is unclear if the libraries that do wish to utilize LOCKSS want it as an exclusive service, but there may be options that would enable libraries that wish to utilize LOCKSS to do so without requiring libraries that do not wish to archive e-journals to utilize LOCKSS and without forcing GPO to duplicate effort to provide both options. The most straightforward model might be to provide access to both groups of libraries through a LOCKSS enabled site that LOCKSS libraries could cache and other libraries could access through a URL or PURL. However, formatting the content to enable

LOCKSS use could require more clicks than the current model for non-LOCKSS users to access the content.

If GPO chooses to provide e-journal content through LOCKSS only, membership in the LOCKSS Alliance becomes an issue. Stanford provides technical support to members of the LOCKSS Alliance, and GPO is not in a position to provide technical support to libraries that choose not to join. GPO would either need to require libraries to join the LOCKSS Alliance or would need to negotiate with Stanford to pay the dues for FDLP libraries to ensure that depository libraries had technical support.

A search of the online List of Classes

<<http://fedbbs.access.gpo.gov/library/download/CLASS/listclas.txt>> found 592 serial titles currently disseminated to FDLP libraries with frequencies of quarterly, bimonthly, or monthly¹. Based on the data outlined in Appendix C of this report, a basic e-journal requires 3 hours of GPO staff time per issue, plus 1 additional hour of GPO staff time per archival unit, to harvest and republish on GPO's Web site. Republishing a complex e-journal requires an average of 5 hours of GPO staff time per issue, plus 1 additional hour of GPO staff time per archival unit to harvest and republish the content.² The table below assumes the same ratio of basic (60%) and complex (40%) e-journal sites as the titles in the pilot.

Frequency	Number / Complexity	Staff time required per archival unit (hrs)	Total hours annual staff time required
Quarterly	178 basic	13	2314
	119 complex	21	2499
Bimonthly	39 basic	19	741
	26 complex	31	806
Monthly	138 basic	37	5106
	92 complex	61	5612
Total			17,078

The titles in the pilot were selected based on inclusion in both the FDLP and the IES and the item selection rates of those titles. As such, they may not be a representative sample. Based on this data, it would require approximately 9 FTEs to harvest and re-publish content to GPO servers in a LOCKSS-friendly format.³

Additionally, 3 hours are required to write a plug-in for each archival unit.⁴ Assuming one archival unit per year, writing plug-ins for 592 titles would require 1776 hours of GPO staff time, approximately 1 FTE.

¹ List of Classes downloaded from the FDLP Desktop on 11/06/2006 contained 297 quarterly EL titles, 65 bimonthly EL titles, and 230 monthly EL titles. More frequently (daily, weekly) or less frequently (annual, biennial) published serials were not considered in this report.

² For a basic journal, manually harvesting (.5 hr.) + adjusting HTML links (.5 hr.) + adding metadata (2 hrs) = 3 hrs. per issue. For a complex journal, manually harvesting (2 hrs) + adjusting HTML links (1 hr) + adding metadata (2 hrs.) = 5 hrs per issue. Both require creating a publisher manifest (.5 hr) + creating an index page (.5 hr.) = 1 hr. per archival unit, p. 18.

³ For this report, 1 FTE was considered to be 48 weeks at 40 hours per week, or 1920 hours.

⁴ See p. 7.

V. Final Recommendations

GPO's emerging enterprise architecture requires that new applications be compatible with FDSys or face the risk of near-term obsolescence. Based on an extensive analysis of all possible options, GPO has decided to devote its resources to the development of FDSys, combined with a concurrent updating of several of its legacy systems. Further implementation of technologies for digital distribution in conjunction with FDSys will not only be more productive, but also more efficient. The most feasible path is to provide for digital distribution as a function of the FDSys for the libraries that wish to utilize it. This would allow GPO to devote more staff time to preparation for FDSys and would reduce the risk of developing something that might not be continued after FDSys implementation. As GPO continues to evaluate technologies for digital distribution, GPO staff plan to survey libraries for more information on their specific needs in this area.

The LOCKSS pilot was successful in providing insight into Web-based content distribution. The goals of the project were met, and useful information was gained for furthering GPO's commitment to permanent public access.

Information on the LOCKSS pilot will be archived on a Web page, including the journal content used in this pilot. The official version of the content, like other official content of the FDL P, will be available permanently via the Catalog of U.S. Government Publications. Additionally, this report and other comments received via forums and future surveys will be used to further refine requirements for the FDSys.

Appendix A:

Issues, Harvested Formats, and Size of the 2005 Archival Unit for Each Pilot E-journal

E-Journal Title	Issues	Harvested Formats	2005 Archival Unit Size (MB)
<i>Treasury Bulletin</i>	4	<ul style="list-style-type: none"> • HTML homepage • Articles in Word • Issue in PDF 	118
<i>Social Security Bulletin</i>	4	<ul style="list-style-type: none"> • HTML homepage • Articles in PDF • Articles in HTML, if available • Issue in PDF 	37
<i>Journal of Research of the National Institute of Standards and Technology</i>	6	<ul style="list-style-type: none"> • HTML homepage • Articles in PDF 	88
<i>Humanities</i>	6	<ul style="list-style-type: none"> • HTML homepage • Articles in HTML 	2
<i>Survey of Current Business</i>	12	<ul style="list-style-type: none"> • HTML homepage • Articles in PDF 	57
<i>Monthly Labor Review</i>	12	<ul style="list-style-type: none"> • HTML homepage • Articles in PDF • Abstracts in HTML • Excerpts in HTML 	12
<i>Monthly Energy Review</i>	12	<ul style="list-style-type: none"> • Issue in PDF 	32
<i>FBI Law Enforcement Bulletin</i>	12	<ul style="list-style-type: none"> • Issues in HTML • Issues in PDF 	63
<i>Amber Waves</i>	5	<ul style="list-style-type: none"> • Issues in PDF • Web only updates in HTML 	28
<i>Environmental Health Perspectives</i>	12	<ul style="list-style-type: none"> • Articles PDF 	395
Totals	115	PDF, HTML, Word	820

Appendix B:

An Example of a Publisher Manifest Page for the GPO LOCKSS Pilot Project

The screenshot shows a Mozilla Firefox browser window displaying the GPO LOCKSS Pilot Project website. The address bar shows the URL: <http://permanent.access.gpo.gov/lockss/Humanities/2005/index.html>. The page features a blue header with the GPO Access logo and navigation tabs for LEGISLATIVE, EXECUTIVE, JUDICIAL, HELP, and ABOUT. A search bar is located in the top right corner.

The main content area is titled "GPO LOCKSS Pilot Project: Humanities - 2005 Volume 26 Archival Unit". Below the title, there is a paragraph explaining the purpose of the page as a Publisher Manifest for the 2005 Volume 26 Archival Unit of Humanities. The page lists the following issues for 2005:

- [2005-01](#)
 - January/February 2005
- [2005-03](#)
 - March/April 2005
- [2005-05](#)
 - May/June 2005
- [2005-07](#)
 - July/August 2005
- [2005-09](#)
 - September/October 2005
- [2005-11](#)
 - November/December 2005

At the bottom of the page, it states: "A service of the U.S. Government Printing Office and the Stanford University LOCKSS Program." and "Last updated: December 5, 2005". The page name is also provided: <http://permanent.access.gpo.gov/lockss/Humanities/2005/index.html>.

The Windows taskbar at the bottom shows the Start button, several open applications including "Inbox - Microsoft Out...", "LOCKSS comments(ch...", "LOCKSS Report Draft...", and "GPO LOCKSS Pilot Pro...", and the system tray with the time 11:20 AM.

Appendix C:

Time Averages for Harvesting Tasks Performed by a GS-12 with Intermediate Web Development Skills

Task	Average Time
Manually harvest 1 issue (basic)	.5 hour
Adjust HTML links using DreamWeaver for 1 issue (basic)	.5 hour
Manually harvest 1 issue (complex)	2 hour
Adjust HTML links in DreamWeaver for 1 issue (complex)	1 hour
Create table of contents and link articles for 1 issue	3 hours
Add metadata to 1 issue	2 hours
Create a publisher manifest for an archival unit	.5 hour
Create index page for each e-journal	.5 hour
Web administration per week	1 hour

Appendix D:

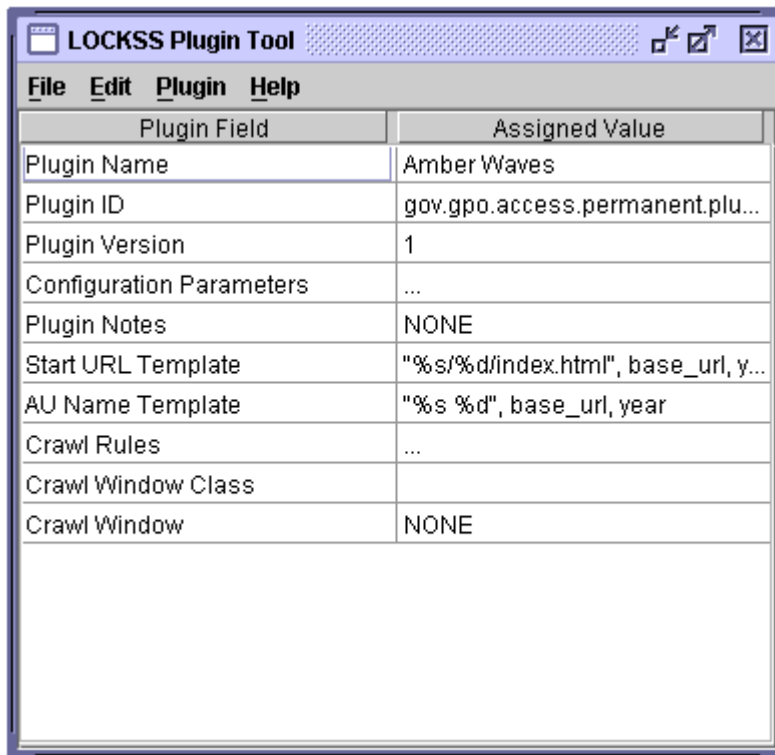
Please note that <http://permanent.access.gpo.gov/lockss/> is the secure directory used for the pilot. It will only be accessible to pilot partners and GPO. It is cited for here for descriptive purposes.

Steps for the *Amber Waves* Journal Plug-in to be Uploaded and Checked on
<http://permanent.access.gpo.gov/lockss/>

1. Download the e-journal content.
2. Restructure the content to fit into a directory for the journal year or number, and place that directory in another directory for the name of the journal. For example, all of the content for the 2005 *Amber Waves* journal will be located at http://permanent.access.gpo.gov/lockss/Amber_Waves/2005/.
3. Download the plug-in software, and run “runtool.bat”
4. Double-click the Assigned Value for Plug-in Name and change it to the name of the journal. For example, *Amber Waves*.
5. Double-click the Assigned Value for Plug-in ID, and change it according to the following template: “gov.gpo.access.permanent.plugin.Amber_WavesPlugin”. The first few terms are the reverse of the domain name of the location of the plug-in (gov.gpo.access.permanent.), followed by “plugin.”, followed by the plug-in name (“Amber_Waves”, replacing spaces with _’s), followed by “Plugin”. The main point here is to be unique from other plug-ins; there is no domain translation being performed.
6. Most plug-in versions will stay at 1, so leave the Plugin Version field as 1.
7. Click the Assigned Value for Configuration Parameters, and a new window will appear. Since we placed the e-journal content in a directory inside a directory (the journal year/id inside of the journal name), we will have to add those two fields to the Parameter List. Click “base_url (Base URL)” under Available Parameters, and click “Add”. If you placed your content in a directory titled by the year (ex: 2005), click “year (Year)”, otherwise click “journal_id (Journal Identifier)” and click “Add”. Now, both “base_url (Base URL)” and “year (Year)” will appear under “Plugin Parameters”. Click OK.
8. Click the Assigned Value for Start URL Template, and a new window will appear. Here, we give the starting URL for the plug-in. In our example, the starting URL is actually http://permanent.access.gpo.gov/lockss/Amber_Waves/2005/index.html, but we need to insert the Template into this window, not the actual link. The end result will look like this: “Base URL/Year/index.html”. In order to input this into the window, click on the drop-down menu at the bottom, choose “Base URL”, and click “Insert Parameter”. Click in the text window, and add a “/” after “Base URL”. Click on the drop-down menu again, choose “Year”, and click “Insert Parameter”. Click in the text window, and add a “/index.html” at the top. Click “Save”.
9. Click the Assigned Value for AU Name Template, and a new window will appear. The end result will be “Base URL Year”. In order to input this into the window, click on the drop-down menu at the bottom, choose “Base URL”, and click “Insert Parameter”. Click in the text window, and add a single space after “Base URL”. Click on the drop-down menu again, choose “Year”, and click “Insert Parameter”. Click “Save”.
10. Click the Assigned Value for Crawl Rules, and a new window will appear. Multiple rules can be added, but we only need to add one rule. Click “Add”, and a new rule will appear. Click on “NONE” and a new window will appear to edit the Crawl Rule template. The end result will look like this: “Base URL/Year/*”. In order to input this

into the window, click on the drop-down menu at the bottom, choose “Base URL”, and click “Insert Parameter”. Click in the text window, and add a “/” after “Base URL”. Click on the bottom drop-down menu again, choose “Year”, and click “Insert Parameter”. Click in the text window, and add a “/” after “Year”. Click on the top drop-down menu, choose “Anything”, and click “Insert Match”. Click “Save”. Click “OK”.

11. Save the plug-in by clicking “File” and choosing “Save as”.
12. Test the plug-in by clicking “Plugin” at the top of the screen, and choosing “Test Crawlrules”. In the new window, insert the year (2005) and the Base URL (http://permanent.access.gpo.gov/lockss/Amber_Waves/), and click “Check AU”. In the new window, change the Test Depth to 9, and click “Check URL”. It may take a few minutes to complete the entire run of the test, as each link has a 6 second fetch delay.



The screenshot shows a window titled "LOCKSS Plugin Tool" with a menu bar containing "File", "Edit", "Plugin", and "Help". Below the menu bar is a table with two columns: "Plugin Field" and "Assigned Value".

Plugin Field	Assigned Value
Plugin Name	Amber Waves
Plugin ID	gov.gpo.access.permanent.plu...
Plugin Version	1
Configuration Parameters	...
Plugin Notes	NONE
Start URL Template	"%s/%d/index.html", base_url, y...
AU Name Template	"%s %d", base_url, year
Crawl Rules	...
Crawl Window Class	
Crawl Window	NONE