GPO LOCKSS Pilot: Final Analysis, Executive Summary

U.S. GOVERNMENT PRINTING OFFICE

KEEPING AMERICA INFORMED

April 12, 2007

GPO LOCKSS Pilot: Final Analysis, Executive Summary

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.

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.

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.

Phase I of the pilot called for GPO to establish and test a LOCKSS cache at GPO. Phase II entailed making Federal Government e-journals available to select Federal depository libraries and International Exchange Service (IES) pilot partner libraries that are operating LOCKSS caches.

Lessons Learned

The depository library community has not been surveyed to determine whether there is sufficient interest to use LOCKSS as the only distribution mechanism for e-journal content, thereby requiring all libraries to receive e-journal content through LOCKSS. It is unclear if the libraries that do wish to utilize LOCKSS want it to be 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. This would most likely involve posting content in a LOCKSS-enabled format where both groups of libraries could access the content. However, this could increase the number of clicks required 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 supply technical support. 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 access to technical support.

While it is possible to remove content from a LOCKSS cache, it is a complicated process. GPO requires the ability to remove content from caches in the event of an agency recall. More work would be required to streamline removal of content to ensure that libraries and GPO can comply with recall requirements.

Recommendations & Next Steps

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 FDLP, 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.