

Lots of Copies Keep Stuff Safe (LOCKSS) Government Documents SGER # 0245231
Chuck Eckman; Victoria Reich; Thomas Robertson, David SH Rosenthal

Background The Government Printing Office (GPO) and the 1200 libraries of the Federal Depository Library Program (FDLP) have been collaborating for 140 years to ensure that the citizens of the United States have access to their government's information. They now face a transition from a paper-based to a web-based system. A recent joint study by the GPO and a group of FDLP librarians, funded by a NSF SGER, identified the LOCKSS (Lots Of Copies Keep Stuff Safe) technology as a potential basis for this transition. The system may restore to librarians the ability to take custody of government document materials, to continue to build collections, and to ensure government document products will be available for the medium and very long term. The LOCKSS Government Documents program is an official GPO Partnership project

The LOCKSS program (<http://lockss.stanford.edu>) aims to provide affordable tools by which access to web delivered information can be preserved over generations. The LOCKSS software creates low-cost, persistent digital "caches" of authoritative versions of http-delivered content. All file formats delivered through HTTP are included. The LOCKSS system enables institutions to build, preserve and provide access to local collections. An institution's LOCKSS cache *collects* material as it is published by crawling the publisher's web site, *preserves* it by engaging with other LOCKSS caches holding the same material in a continual, slow, automatic process of mutual audit and repair, and *distributes* it to the institution's readers by acting as a proxy web cache, transparently delivering the preserved material if, for any reason, it is not available on request from the publisher (2). The mutual audit and repair mechanism operates through a novel peer-to-peer sampled voting protocol, which tolerates both failures and attacks (3, 4). It provides assurance that the material collected is complete and authentic, and that its preservation has been successful. If a cache loses or corrupts material the failure will be detected by the audit mechanism and it will be repaired automatically from the publisher or from another LOCKSS cache. A principle of the LOCKSS design is that the most significant threats to digital preservation are economic. Thus the LOCKSS software is free and open source, it runs on generic low-cost PC hardware, and great attention has been paid to reducing the need for skilled system administration (5). The LOCKSS project is hosted at <http://www.sourceforge.net>, from where the software is freely available.

Partners GPO; California Digital Library; US National Agricultural Library; U of Minnesota; U of North Texas; CSU San Bernardino; U of Colorado; U of Nevada Reno; Georgetown; Yale; North Carolina State.

Findings Federal websites occupy approximately one-half to one percent of the 'surface web'. The dot gov domain is largely opaque, and occupies as much as 85 percent of the 'deep web'. Between 1992 and 2003 the 'dot gov' domain has increased thirteen-fold. This content is highly volatile; the average online Federal web resource exists for four months (1). Web government documents are marked by filetype diversity but with a preponderance of two filetypes (html and pdf). There is great genre diversity (publications, documents, databases of all types including numeric, GIS, transaction records, etc.); and many complex file structures. Key problems for preservation of dot gov domain websites are: tracking changes in versions and editions; assuring readers regarding the 'official-ness' of information; ensuring integrity of information.

The distributed, open model represented by LOCKSS appeals strongly to partners because of its direct conceptual relationship to the principals of the print depository program where long-term preservation and access is assured by virtue of a highly distributed system based on maximum redundancy and the high volume of direct user access to content (6).

Possible roles for library partners: Affirming as memory organizations as partners in maintaining file integrity and authenticity; Providing possibility of incorporating government content within local digital libraries and allowing cross-fertilization of information access with non-government content; Creating possibility for new knowledge creation through the development of new interfaces to government content at the local level.

Possible roles for the US GPO: Leverage its relationship with various Federal agencies and partnerships with depository libraries to develop agency-specific LOCKSS technology; Crawl agency web sites looking for publications that GPO had not been alerted to by agencies; Use as a tool for monitoring agency site changes, new editions, new issues, etc.; Normalize formats across agencies because as GPO came across divergent formats, it could contact the agency and request another official version in a different format; Authenticate captured content; Apply as appropriate digital signatures; Disseminate the content and associated metadata through the LOCKSS network.

Building a community of LOCKSS partners from the existing FDLP program raises several issues. Why might libraries join a LOCKSS-based network? These may be the same general reasons that they joined the depository program originally. Namely: Good citizenship: it is a benefit to all to collect U.S. public documents for current and future access; Authenticity: participation helps to assure that the content is authentic (as assured by the imprint on paper in the past) for local users; Preservation: participation in the network ensures the integrity of the files; Access: file redundancy provides the opportunity to develop direct local access to cached content not subject to the bandwidth limitations posed by remote access to resources.

Sustainability is a key economic aspect of collection development: (a) How to best support LOCKSS development for diverse government agency web-publishing platforms? (b) How to ensure that a sufficient number of institutions develop and maintain caches for a given array of government web publications to ensure file integrity? (c) How to make active participation in a LOCKSS documents program attractive to institutions that withdrawal would be very rare?

The legal requirements governing a potential LOCKSS government documents application are contained in Chapter 19 US Code Title 44. LOCKSS would need to comply with Title 44 requirements pertaining to retention periods and withdrawals of deposited materials, and provision of access to deposited materials. It appears that a LOCKSS implementation would be consistent with existing law and GPO's Legal Counsel is affirming this assumption.

References 1. "Web-based Government Information: Evaluating Solutions for Capture, Curation, and Preservation". Project Report – 8/7/2003. Unpublished. California Digital Library. 2. David S. H. Rosenthal and Vicky Reich. "Permanent Web Publishing". In Proceedings of the USENIX Annual Technical Conference, (Freenix 2000), pages 129-140, San Diego, CA, 6/2000. 3. Petros Maniatis, Mema Roussopoulos, TJ Giuli, David S. H. Rosenthal, Mary Baker, and Yanto Muliadi, "Preserving Peer Replicas By Rate-Limited Sampled Voting", *Proceedings of the nineteenth ACM symposium on Operating systems principles, Bolton Landing*. 10/2003, p. 44. 4. Petros Maniatis, Mema Roussopoulos, TJ Giuli, David S. H. Rosenthal, Mary Baker, and Yanto Muliadi, "Preserving Peer Replicas in the LOCKSS System", *ACM, Transactions on Computing Systems*. To be published. 5 David S. H. Rosenthal. "A Digital Preservation Network Appliance Based on OpenBSD". In Proceedings of BSDcon 2003, San Mateo, CA, 9/2003. 6. Chuck Eckman et. al., "LOCKSS for Government Documents: A Needs Assessment for the Federal Depository Library Community", <http://lockss-docs.stanford.edu/NAfinal.pdf>, 1/2004.