An Open-Source System for Automatic Policy-Based Collaborative Archival Replication

Thu-Mai Christian
& Jonathan Crabtree
University of North Carolina
Odum Institute
jonathan_crabtree@unc.edu

Nancy McGovern
University of Michigan
ICPSR
nancymcg@umich.edu

Micah Altman
Harvard University
IQSS
micah_altman@harvard.edu

ABSTRACT
In this poster, we provide an overview of the SafeArchive system and describe how a curator can use the tools to generate an archival policy schema and monitor compliance. Also, the poster details the technical implementation of the SafeArchive system including the policy schema, how information used in the auditing process is obtained from a set of LOCKSS peers without modifying the LOCKSS trust model or configuration, and the organization of SafeArchive software components.

Categories and Subject Descriptors
H.3.4 [**Systems and Software**]: Distributed Systems Audit

General Terms
Management, Measurement, Documentation, Performance, Reliability, Legal Aspects, Verification

Keywords
Audit, Open-Source, Policy, LOCKSS, TRAC, Preservation, Archive

1. INTRODUCTION
Verified geographically-distributed replication of content is an essential component of any comprehensive digital preservation plan. This requirement has emerged as a necessity for recognition and certification as a trusted repository. As embodied in Trustworthy Repositories Audit & Certification (TRAC) [1] and the subsequent TRAC-based ISO 16363 Audit and Certification of Trustworthy Digital Repositories, and in other best practices, an organization must have a managed process for creating, maintaining, and verifying multiple geographically distributed copies of its collections in order to be fully trusted.

The LOCKSS (Lots of Copies Keep Stuff Safe) [2] system has been widely adopted by libraries and archives for replication and preservation. As a collaborative effort of Data-PASS partners (ICPSR, Roper Center, University of Connecticut, Odum Institute and IQSS), the SafeArchive system has been developed to extend LOCKSS capabilities by making distributed replication easier for curators and automating compliance with formal replication and storage policies. This innovation provides the auditability and reliability of a top-down replication system with the resilience of a peer-to-peer model.

iPRES2011, Nov. 1–4, 2011, Singapore.

2. THE SYSTEM
2.1 Overview of the SafeArchive System
SafeArchive is described in more detail in [3] and is based on a prototype [4] developed by the Data-PASS partners [5, 6], and funded by the Library of Congress. This prototype established feasibility and the core operational use cases for the system. The SafeArchive system has been completely rewritten and redesigned for production use.

Abstractly, the system is designed to create a virtual overlay network on top of a peer-to-peer replication network that supports provisioning, monitoring, and TRAC/ISO 16363-based auditing.

Operationally, users of the system can perform the following functions, as illustrated in figure 1:

- Analyze any LOCKSS network;
- Check that collections are replicated, valid, and up-to-date;
- Create formal replication policies;
- Replicate content from web sites or digital repository systems;
- Audit the network for current and historical TRAC/ISO 16363 compliance; and
- Automatically manage and repair a LOCKSS network based on a specified replication policy.

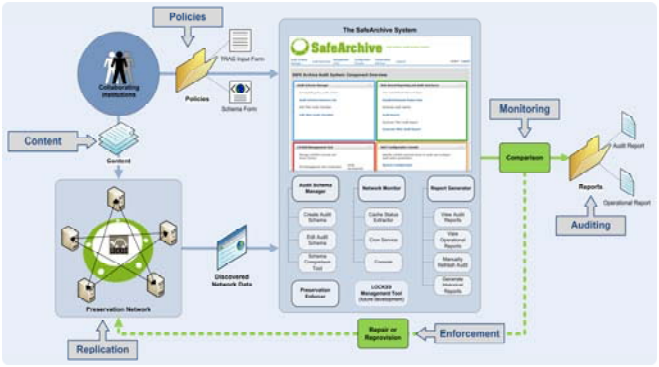

Figure 1. Abstract diagram of system functions and roles

The SafeArchive system is designed to collaborate with the Dataverse Network® [7] system. Curators who store content in a Dataverse can easily expose content for replication by LOCKSS and SafeArchive through a simple graphical interface.

Institutionally, the SafeArchive enables memory institutions and preservation collaborations to formalize their replication policies and inter-archival replication commitments; represent these replication policies in machine-readable form; and to continuously audit any set of public or private LOCKSS hosts for policy compliance. The SafeArchive system is open source and available at: http://www.safearchive.org.

2.2 Using the SafeArchive System

Generally speaking, the system coordinates six activities:

1. Collaborating institutions agree on a replication policy. This records the resource commitments, descriptions of the collections to be preserved, and desired replication guarantees.
2. Institutions make collections of content ("archival units") available through the web (e.g., as web pages or through the Dataverse Network®).
3. LOCKSS caches harvest the collections from their original source repositories using standard protocols such as HTTP or OAI-PMH.
4. SafeArchive monitors the network, assesses it against the stated replication policy, and produces an audit trail. The system also alerts collaborators when formal policies are not being met.
5. SafeArchive produces an audit trail of operational and audit reports.
6. The SafeArchive will also coordinate harvesting of the LOCKSS caches by "inviting" members of the network to harvest content that is under-replicated. This will be used to automatically configure a network based on a policy schema to reconfigure and repair the network as the number of participating caches, collections and institutions changes intentionally or unintentionally.

The SafeArchive system is designed to give curators the ability to easily define preservation policy, examine the content of the preservation network, and generate regular audit reports that support TRAC/ISO 16363 compliance. All changes to the policy schema instance and the machine-readable audit reports are versioned and stored permanently—so that a complete history of compliance is preserved.

3. SUMMARY

The SafeArchive system provides a way to ensure that replicated collections are both institutionally and geographically distributed while allowing for the development of increasingly measurable and auditable trusted repository requirements. Designed as a virtual overlay network on LOCKSS, the system provides the auditability and reliability of a top-down replication system with the resilience of a peer-to-peer model. This enables any library, museum, or archive to audit the replication of their collections across an existing LOCKSS network in compliance with documented archival policies. It also allows groups of collaborating institutions to automatically and verifiably replicate each others' content consistent with a set of expressed commitments stored in machine readable XML based policies. The result is that archives can more easily collaborate to preserve content through geographically and institutionally distributed replication, which mitigates technical and organizational threats to preservation.

The project is in its second year of development and the first official version 1.0 of the system has been released. The system is being field-tested, and optimizations from those experiences are being incorporated into version 2.0 that is slated for release in early 2012.

4. ACKNOWLEDGEMENTS

The project is a collaborative effort of the Data-PASS Partners: The International Consortium for Political and Social Research, University of Michigan; The Roper Center for Public Opinion Research, University of Connecticut; the Howard W. Odum Institute at the University of North Carolina at Chapel Hill; the National Archives and Records Administration; and the Institute of Quantitative Social Science, Harvard University. It is managed through the Institute of Quantitative Social Science, and works in collaboration with the LOCKSS project at Stanford University.

The project is sponsored by the Institute of Museum and Library Services (IMLS), under award #LG-05-09-0041-09.

5. REFERENCES

[1] RLG-NARA Task Force on Digital Repository Certification. (2007). Trustworthy Repositories Audit and Certification (TRAC): Criteria and checklist (version 1.0). Chicago, IL: Center for Research Libraries. Retrieved from http://www.crl.edu/PDF/trac.pdf

[2] Reich, V. & Rosenthal, D. (2001). LOCKSS: A permanent web publishing and access system. *D-Lib Magazine*, *7*(6). Retrieved from http://www.dlib.org/dlib/june01/reich/06reich.html

[3] Altman, M., & Crabtree, J. (2011). Using the SafeArchive System : TRAC-based auditing of LOCKSS. *Archiving 2011* (pp. 165-170). Society for Imaging Science and Technology. doi:ISSN 978-0-89208-294-0

[4] Atman, M., Beecher, B., Crabtree, J., Andreev, L., Bachmann, B., Buchbinder, A., Burling, S., King, P., & Maynard, M. (2009). A prototype platform for policy-based archival replication. *Against The Grain*, *21*(2), 44-47.

[5] Altman, M., Adams, M., Crabtree, J., Donakowski, D., Maynard, M., Pienta, A., & Young, C. (2009). Digital preservation through archival collaboration: The Data Preservation Alliance for the Social Sciences (Data-PASS). *The American Archivist*, *72*(1), 169-182.

[6] Gutmann, M., Abrahamson, M, Adams, M.O., Altman, M., Arms, C., Bollen, K., Carlson, M., Crabtree, J., Donakowski, D., King, G., Lyle, J., Maynard, M., Pienta, A., Rockwell, R, Timms-Ferrara L., & Young, C. (2009). From preserving the past to preserving the future: The Data-PASS project and the challenges of preserving digital social science data. *Library Trends*, *57*(3), 315-337.

[7] Crosas, M. (2011). The Dataverse Network®: An open-source application for sharing, discovering and preserving data. *D-Lib Magazine 17*(1/2). Retrieved from http://www.dlib.org/dlib/january11/crosas/01crosas.html