AN OPEN-SOURCE SYSTEM FOR AUTOMATIC POLICY-BASED COLLABORATIVE ARCHIVAL REPLICATION

Website: http://safearchive.org
Contact: Micah Altman <micah_altman@harvard.edu> | Jonathan Crabtree <jonathan_crabtree@unc.edu>

The Need for Policy-Based Replication

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

Overview of the SafeArchive System

SafeArchive automates high level replication policies (e.g., TRAC), and helps institutions to collaborate in preserving digital content. GUI-based tools are designed for librarians and archivists–not systems administrators.

The system coordinates and audits existing groups of public or private LOCKSS peers. Without requiring a single authority, this allows a group of institutions to establish actionable and mutually verifiable policies governing the replication of content and interest to those institutions.

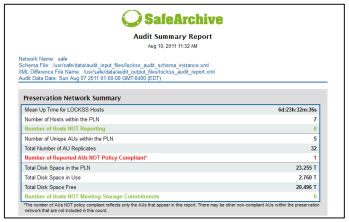
Operationally, system users can:

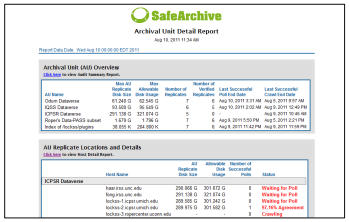
- Analyze any LOCKSS network
- Check that collections are replicated, valid, and up-to-date
- Create formal replication policies
- Replicate content from web sites or digital repository systems, such as *The Dataverse Network®*
- Audit the network for current and historical TRAC compliance
- Automatically manage and repair a LOCKSS network based on a specific replication policy

SafeArchive provides the reliability of a top-down replication system with the resiliency of a peer-to-peer model.

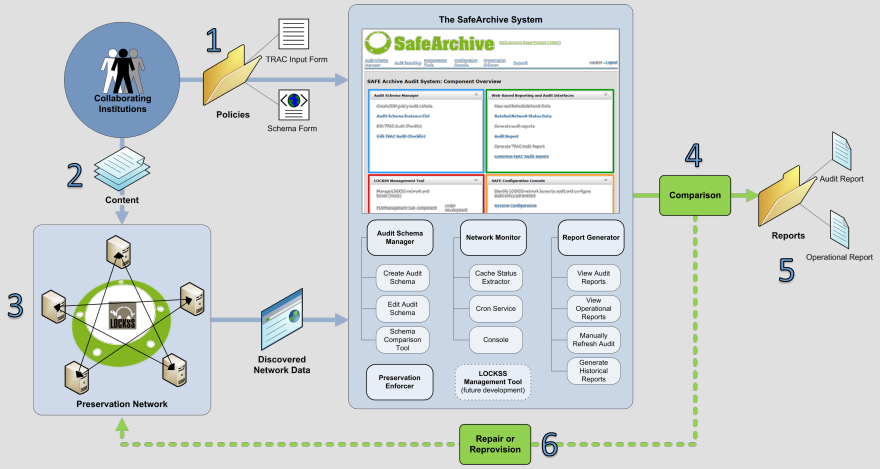
How the SafeArchive System Works

- The **Network Monitor** gathers the information on each cache necessary to support policy reporting and auditing.
- Curators specify replication policies for LOCKSS networks, which they input into a web-based questionnaire using the SafeArchive user interface. The **Audit Schema Manager** outputs the policies into a machine-readable XML-based schema. A comparison tool then produces a machine-readable diff.XML difference report that enumerates discrepancies between the actual and desired states. All changes to the policy schema instance and the diff.XML reports are versioned and stored permanently to provide a complete history of compliance.
- The **Report Generator** uses network and diff.XML data to create formatted reports containing both "audit" summaries that reflect policy compliance, and "operational" information used for diagnostics and performance analysis.





- The **Preservation Enforcer** allows the system to make "adjustment" requests to individual LOCKSS caches to initiate or discontinue content harvesting.
- If content needs to be retrieved from the system, the **LOCKSS Management Tool** coordinates the location of the appropriate cache and restores the content (future development).



Using the SafeArchive System

The SafeArchive System coordinates six (6) primary activities to give curators the ability to easily define preservation policy, examine the content of the preservation network, and generate regular audit reports that support best practices.

1. **Collaborating institutions author a replication policy with SafeArchive tools**.
2. **Institutions make collections of content available** through the web.
3. **LOCKSS caches harvest the collections** from their original source repositories and coordinate peer-to-peer to monitor and maintain network integrity.
4. **SafeArchive monitors the state of the network** and compares it to the stated replication policy.
5. **SafeArchive produces an audit trail** of operational and audit reports, and alerts collaborating institutions when formal policies are not met.
6. **SafeArchive provisions replication** as necessary to enforce policy compliance.

The SafeArchive project is a collaborative effort of the Data-PASS Partners: The Inter-university Consortium for Political and Social Research, University of Michigan; the Roper Center for Public Opinion Research, University of Connecticut; the Howard W. Odum Institute for Research in Social Science, University of North Carolina at Chapel Hill; U.S. National Archives & Records Administration (NARA); 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.