



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

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), the subsequent TRAC-based ISO 16363 Audit and Certification of Trustworthy Digital Repositories, and in other best practices.

Overview of the SafeArchive System

SafeArchive automates high level replication policies (e.g., TRAC/ISO 16363), 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.

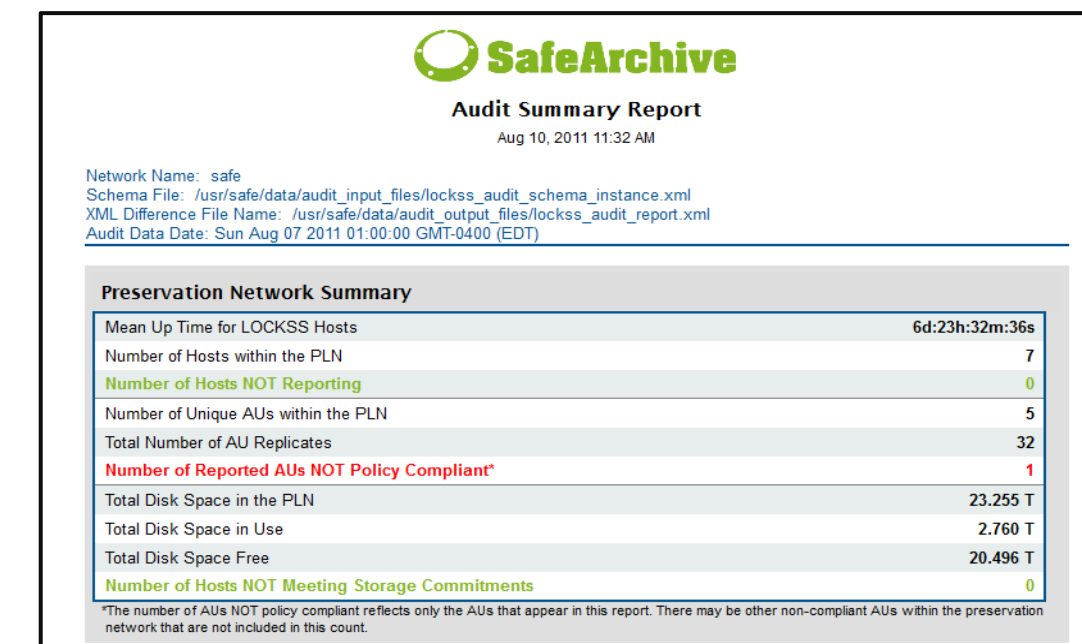
Operationally, system users can:

- 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/ISO 16363 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 analyses.

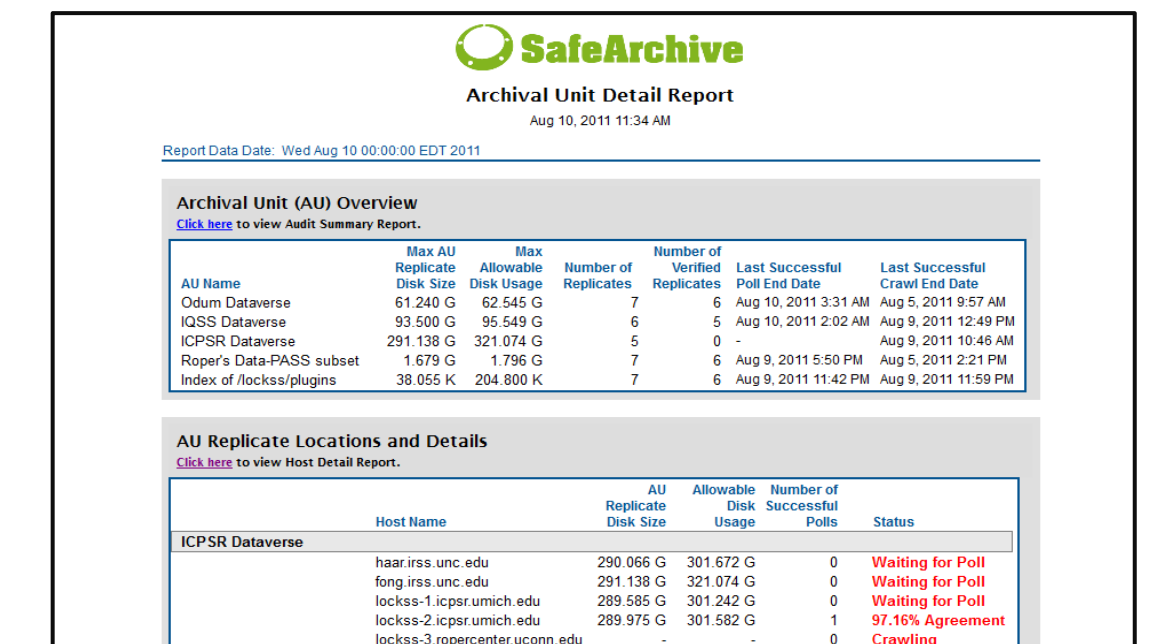


SafeArchive Audit Summary Report
 Aug 10, 2011 11:32 AM

Network Name: safe
 Schema File: /usr/local/data/audit_input_files/locks_audit_schema_instance.xml
 XML Difference File Name: /usr/local/data/audit_output_files/locks_audit_report.xml
 Audit Data Date: Sun Aug 07 2011 01:00:00 GMT-0400 (EDT)

Preservation Network Summary	
Mean Up Time for LOCKSS Hosts	6d:23h:32m:36s
Number of Hosts within the PLN	7
Number of Hosts NOT Reporting	0
Number of Unique AUs within the PLN	5
Total Number of AU Replicates	32
Number of Reported AUs NOT Policy Compliant*	1
Total Disk Space in the PLN	23,255 T
Total Disk Space in Use	2,760 T
Total Disk Space Free	20,496 T
Number of Hosts NOT Meeting Storage Commitments	0

*The number of AUs NOT policy compliant reflects only the AUs that appear in this report. There may be other non-compliant AUs within the preservation network that are not included in this count.



SafeArchive Archival Unit Detail Report
 Aug 10, 2011 11:34 AM

Report Data Date: Wed Aug 10 00:00:00 EDT 2011

Click here to view Audit Summary Report.

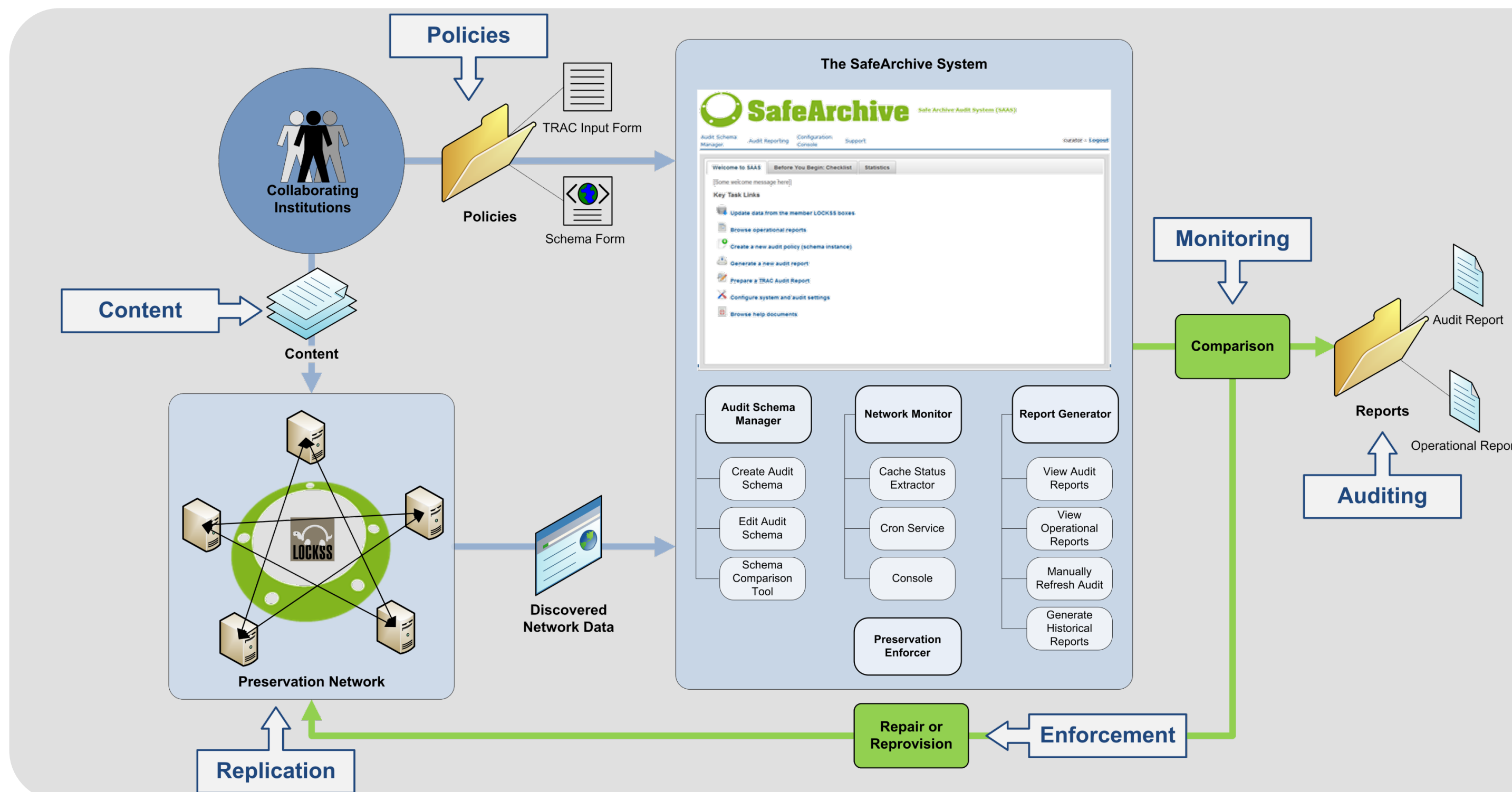
Archival Unit (AU) Overview							
AU Name	Max AU Replicate	Max Allowed	Number of Replicates	Number of Verified Replicates	Last Successful Poll End Date	Last Successful Crawl End Date	
ICPSR Database	61,240 G	62,545 G	7	6	Aug 10, 2011 3:31 AM	Aug 5, 2011 8:47 AM	
ICPSR Database	93,500 G	95,549 G	6	5	Aug 10, 2011 2:02 AM	Aug 9, 2011 12:48 PM	
ICPSR Database	291,138 G	321,074 G	5	0	-	Aug 9, 2011 10:48 AM	
Roper's Data-PASS subset	1,879 G	1,795 G	7	6	Aug 9, 2011 5:50 PM	Aug 5, 2011 2:21 PM	
Index of locks/plugins	38,055 K	204,800 K	7	6	Aug 9, 2011 11:42 PM	Aug 9, 2011 11:59 PM	

Click here to view Host Detail Report.

AU Replicate Locations and Details					
Host Name	AU	Replicate	Allowed	Number of	Status
		Disk Size	Usage	Successfull	
icpsr1.unc.edu	icpsr	290,066 G	301,672 G	0	Waiting for Poll
icpsr2.unc.edu	icpsr	291,138 G	321,074 G	0	Waiting for Poll
icpsr3.unc.edu	icpsr	289,586 G	301,282 G	0	Waiting for Poll
icpsr4.unc.edu	icpsr	289,975 G	301,582 G	1	97.16% Agreement
icpsr5.unc.edu	icpsr	289,975 G	301,582 G	6	Crawling

- The **Preservation Enforcer** initiates any of four actions to enforce compliance when policies are not met:
 - Initiate a poll
 - Initiate a crawl
 - Add content copy
 - Change crawl frequency

Enforcement actions are logged by both the SafeArchive systems and the participating LOCKSS nodes. These actions become part of audit trails and reports.



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.

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