A Platform for Auditable, Distributed, Asymmetric Archival Replication[*]
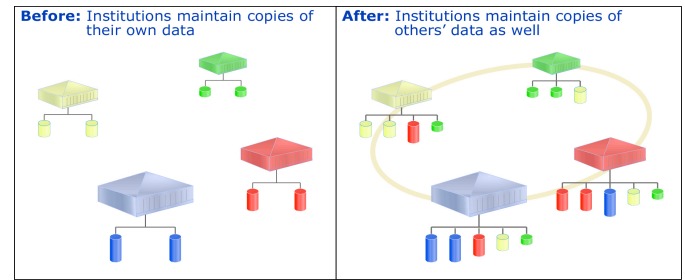
Abstract

Distributed digital preservation networks have
been growing in popularity among institutions
that wish to collect and preserve materials of
high importance to specific communities. We
describe a multi-institutional project to develop
an open-source platform for distributed multi-
institutional asymmetric replication of archival
content. Our goal is to support policy-driven
replication of large collections of research data,
where the partners in the replication network
differ significantly in the size of their collections
and the resource commitments they are able to
contribute to the network.

We have built a prototype system around a core
of Private LOCKSS Networks (PLN)
technology; developed a schema to encapsulate
inter-archival replication commitments; built an
automated schema-driven service that audits
PLN's; and adapted OAI harvesting clients to
harvest data collections from the Dataverse
Network (DVN) and other repositories using the
Data Documentation Initiative (DDI) schema.

This work is conducted by Data-PASS, a
partnership of five major archives, and
supported by the Library of Congress NDIIPP
program.

Features

- Policy-driven:
  o Institutional policy establishes replication
     commitments
  o Resource commitments may be
     heterogeneous
- Uses open, standard, technology
  o Services built on top of LOCKSS
  o No patches or code alterations
- Central auditing, distributed management
  o Each host is independently managed, no
     global super-user
  o Centrally audited to detect divergence from
     commitments

Before: Institutions maintain copies of their own data **After:** Institutions maintain copies of others' data as well

How it Works

1. Commitment schema describes:
   a. Desired replication guarantees
   b. Resources offered by hosts participating
      in the replication network
   c. Collections to be replicated, and limits
      on resources to be allocated to them
2. Network is audited
   a. Network is monitored by LOCKSS
      "cache manager"
   b. SSP tool queries cache manager
      database to determine state of network
   c. State of network is compared to
      commitment schema
   d. A report of all differences between
      actual and desired state is produced
3. Change network state
   a. SSP tool processes "difference report"
   b. Generates sets of requests of form:
      HOST_ID [start|stop]
      COLLECTION_ID
   c. Sends requests to hosts
   d. Host administrators approve requests

For more information

For official releases see the Data-PASS web site:

   http://data-pass.org

For questions about the project contact:

   lc-operations@icpsr.umich.edu

---

[*] Micah Altman, IQSS, Harvard. U. <Micah_Altman@harvard.edu>; Bryan Beecher, ICPSR, U. of Michigan,
<bryan@umich.edu>; Leonid Andreev, IQSS, Harvard U.; Adam Buchbinder, Roper Center, U. of Connecticut; Ed

Bachmann, Odum Institute, U.N.C.; Steve Burling, ICPSR, U. of Michigan; Jonathan Crabtree, Odum Institute, U.N.C.; Marc Maynard, Roper Center, U. of Connecticut