A Community Approach to Preservation: “Experiences with Social Science Data”

DigCCurr Spring 2009

Jonathan Crabtree
April 3, 2009
The Odum Institute

• Oldest Institute or Center at UNC-CH Founded 1924
• Mission: Teaching, research, & service for social sciences
• Cross-disciplinary focus
The Partners

- ICPSR
- Odum Institute
- Roper Center
- Henry A. Murray Research Archive
- Harvard-MIT Data Center
- National Archives and Records Administration
The Plan

• Identify significant data collections (classic)

• Identify important contemporary data ("at risk")

• Develop common standards and procedures across partnership
Partnership Goals

• Develop common standards and activities
• Determine how the partnership can expand
• Use technological advances to encourage metadata standards and a shared catalog
Dataverse Network

- Open source platform
- OAI server
- DDI metadata standards
- Federated Approach
Overview of How Catalog Works

1. Search Shared Catalog
2. View Information on Data Through Catalog, Link to Data at Partner Site
3. Access Data With Extraction and Analysis, Through Catalog Direct to Partner Sites

- Online Catalog
- Metadata Catalog
- Data Mirror
- Online Analysis
- Harvester

Data-PASS
Data Preservation Alliance for the Social Sciences
Multi-Archival: Syndicated Storage Platform
Nexuses for Preservation Failure

- Technical
  - Media failure: storage conditions, media characteristics
  - Format obsolescence
  - Preservation infrastructure software failure
  - Storage infrastructure software failure
  - Storage infrastructure hardware failure
- External Threats to Institutions
  - Third party attacks
  - Institutional funding
  - Change in legal regimes
Replication as Part of a Multi-Institutional Preservation Strategies

There are potential single points of failure in both technology, organization and legal regimes:

- Diversify your portfolio:
  - multiple software systems, hardware, organization
- Find diverse partners – diverse business models, legal regimes

*Preservation is impossible to demonstrate conclusively:*
- Consider organizational credentials
- No organization is absolutely certain to be reliable
- Consider the trust relationships across institutions
Data-PASS Requirements for SPP

- **Policy Driven**
  - Institutional policy creates formal replication commitments
  - Replication commitments are described in metadata, using schema
  - Metadata drives
    - Configuration of replication network
    - Auditing of replication network
- **Asymmetric Commitments**
  - Partners vary in storage commitments to replication
  - Partners vary in size of holdings being replicated
  - Partners vary in what holdings of other partners they replicate
- **Completeness**
  - Complete public holdings of each partner
  - Retain previous version of holdings
  - Include metadata, data, documentation, legal agreements
- **Restoration guarantees**
  - Restore groups of versioned content to owning archive
  - Institutional failure restoration – support transfer of entire holdings of a designated archive to another partner
- **Trust & Verification**
  - Each partner is trusted to hold the public content of other, not to disseminate improperly
  - Each partner trusts replication broker to add units to be harvested
  - No partner is trusted to have “super-user” rights to delete (or directly manipulate) replication storage owned by another partner
  - Legal agreements reinforce trust model
  - Schema based auditing used to verify replication guarantees are met by the network
Syndicated Storage Platform (SSP)

- Start with LOCKSS
- Lots of Copies Keep Stuff Safe
- But used in a closed network
  - Private LOCKSS Network (PLN)
  - A few of them out there
    - MetaArchive perhaps the best known
- Biggest selling point was independence of each node in the PLN
PLNs

• LOCKSS is really easy to setup
  – PLNs can be more difficult
• Other differences between traditional PLN and our needs
  – Our content isn’t harvestable via HTTP
  – Our PLN nodes are different sizes
  – Our trust model requirement prevents a centralized authority controlling the network
SPP Commitment Schema

- **Network level**:
  - Identification: name; description; contact; access point URI
  - Capabilities: protocol version; number of replicates maintained; replication frequency; versioning/deletion support
  - Human readable documentation: restrictions on content that may be placed in the network; services guaranteed by the network; Virtual Organization policies relating to network maintenance

- **Host level**
  - Identification: name; description; contact; access point URI
  - Capabilities: protocol version; storage available
  - Human readable terms of use: Documentation of hardware, software and operating personnel in support of TRAC criteria

- **Archival unit level**
  - Identification: name; description; contact; access point URI
  - Attributes: update frequency, plugin required for harvesting, storage required
  - Terms of use: Required statement of content compliance with network terms.; Dissemination terms and conditions

- **TRAC Integration**
  - A number of elements comprise documentation showing how the replication system itself supports relevant TRAC criteria
  - Other elements that may be used to include text, or reference external text that documents evidence of compliance with TRAC criteria.
  - Specific TRAC criteria are identified implicitly, can be explicitly identified with attributes
  - Schema documentation describes each element's relevance to TRAC, and mapping to particular TRAC criteria
DPP LOCKSS External Cache Monitoring System

1. Determine elements to compare.

2. Query contents of elements.

3. Create difference report.

DPP Archival Scheme Instance File schema.xml

Comparable Elements

Perl Script

MySQL DB

Cache manager/monitor

Interrogate LOCKSS nodes for contents & status

lockss.xml

Harvesting plugin

lockss-2

lockss-3

Voting, Repairs, & Communication

Interrogate LOCKSS nodes for contents & status

lockss.xml

Harvesting plugin

lockss-2

Plugin conducts harvesting operations

Original Content Server(s)

4. Issue invitations to LOCKSS node lockss.xml files based on diff.xml.

Interesting

Human-readable

diff.html

XSLT Transform

Make diff xml human readable

Future automation

Parse diff.xml

diff.xml

Exception report
Issues & Future Work

- Move from prototype to production
- Look for other applications
- Examine scalability issues
- Bulk recovery to home repositories
- Work toward a fully automated update system
- Examine stability issues around Cache Manager
- Work with the community to develop standard PLN Auditing
Summary

- Replication ameliorates institutional risks to preservation
- Data PASS requires policy based, auditable, asymmetric replication commitments
- Formalize policy in schema
- (Re)Configure & audit LOCKSS using schema
- Replication uses standard LOCKSS mechanisms
Contact Information

Website:  http://www.icpsr.umich.edu/DATAPASS/
http://www.odum.unc.edu

E-mail:  Data-PASS@icpsr.umich.edu

Jonathan Crabtree
jonathan_crabtree@unc.edu