Replicated & Distributed Storage Technologies:
“Impact on Social Science Data Archive Policies”

IASSIST 2010
Ithaca, New York

Jonathan Crabtree
June 2, 2010
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 IQSS
• National Archives and Records Administration
One of the key functions of social science data archives is to *preserve* historic and import data used in social science research.
How can we promise Preservation

• There are many definitions of preservation and many key components to policies that support preservation of social science data.

• “Social science archives should consistently update and evaluate policies to ensure they meet the goals of their organizations”

Data Replication

“Storage alone will not solve the problem of digital preservation. Academic materials have many enemies beyond natural bit rot: ideologies, governments, corporations, and inadequate budgets. It is essential that sound storage and administration practices are complemented with the institution of communities acting together to thwart attacks that are too strong or too extrinsic for such practices to protect against.”

Distributed Replication and Storage Projects

- Policy-Based Replication and Auditing
  - Data-PASS project
  - LOC funded prototype
  - Currently IMLS funded project
  - LOCKSS PLN foundation
  - Schema based auditing

- Rules-Based Distributed Storage
  - NARA/Odum/UNC Chapel Hill SILS project
  - NARA funded
  - iRODS grid based foundation
  - Rules based policy enforcement
Policy-Based Replication and Auditing

- Data-PASS Syndicated Storage Technology “SSP”
Multi-Archival: Syndicated Storage Platform

Before: Institutions maintain copies of their own data

After: Institutions maintain copies of others’ data as well
Preservation Failures

• 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 SSP

- **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
    - Educopia Institute/MetaArchive perhaps the best known
- Biggest selling point was independence of each node in the PLN
PLNs

• Other differences between traditional PLN and our needs
  – Our content isn’t harvestable via HTTP
    • In our case we use OAI-PMH
  – Our PLN nodes are different sizes
  – Our trust model requirement prevents a centralized authority controlling the network
SSP 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

```xml
<SSP>
  <network>
    <networkIdentity>
      <networkCapabilities>
        <protocolVersion version="1.0"/>
        <numberReplicates min="4"/>
        <replicationFrequency maxHours="72"/>
        <versioningPolicy policyType="required"/>
        <deletionPolicy policyType="required"/>
      </networkCapabilities>
    </networkIdentity>
    <networkTerms/>
  </network>
  <host>
    <hostIdentity>
      <hostCapabilities>
        <lockssVersion protocolVersion="1.0" softwareVersion="1.0"/>
        <storageAvailable maxGB="500"/>
      </hostCapabilities>
    </hostIdentity>
    <hostTerms/>
  </host>
  <archivalUnit>
    <cduIdentity>
      <name uniqueID="au1">Gallop</name>
      <description fulltextURI="http://datapass.org">Gallop Polls</description>
      <accessBase adminEmail="support@lcdsr.org" accessURL="http://somesite.com/reference"/>
    </cduIdentity>
    <cduCapabilities>
      <updateFrequency minDays="24"/>
      <storageRequired maxMB="1000"/>
      <pluginRequired pluginURI="http://someplugin.com"/>
    </cduCapabilities>
    <cduTerms>
      <contentTermsCompliance fulltextURI="/" complianceType="inAu">This is compliant</contentTermsCompliance>
      <disseminationTerms fulltextURI="/" disseminationType="clickthrough" disseminationCondition="failure">Some terms</disseminationTerms>
    </cduTerms>
  </cdu>
</archivalUnits>
</SSP>
```
Current Efforts

Policy-Based Replication and Auditing

Preservation Network

Museums

Archives

Libraries
IMLS Project Goals

- Move from prototype to production
- Adapt to more generic uses
- Examine scalability issues
- Bulk recovery to home repositories
- Work toward a fully automated update system
- Rework the interface to LOCKSS cache
- Work with the community to develop standard PLN auditing
Rules-Based Distributed Storage

• Rules-Based policy enforcement
• iRODS grid based technology
• OAI-PMH harvesting from Odum Dataverse network
National Archives and Records Administration
Transcontinental Persistent Archive Prototype

Federation of Seven
Independent Data Grids

Extensible Environment, can federate with additional research and education sites. Each data grid uses different vendor products.
Using approach modeled on MIT Pledge project

- **Step 1** = define policy areas
- **Step 2** = create policy declaration statements for each policy area; state the requirements for operation, not technical specifics
- **Step 3** = each entity in a policy statement is defined in language descriptions: humans and machine-readable references
- **Step 4** = deontic statements: logical statements define actors, actions, and constraints that enforce a policy statement.
- **Step 5** = Write iRODS rules for each statement

Policy Areas

• Organization, Environment, and Legal Policies
• Community and Usability Policies
• Process and Procedure Policies
• Technology and Infrastructure Policies

Initial Rules Developed

• Organization, Environment, and Legal Policies
  – Defined dataset succession plan
  – Defined access policies
  – Log access for accountability
  – Reference TRAC criteria
• Community and Usability Policies
  – Require a deposit agreement
• Process and Procedure Policies
  – Defined iCAT to DDI discovery crosswalk
  – Store dataset’s DDI metadata as object
  – Defined persistent identifiers
  – Defined UNF’s and Checksums
  – Provide reporting of preservation network
• Technology and Infrastructure Policies
  – Defined number of replication copies
  – Defined geographic location for the copies
  – Provide authentication policy
  – Provide versioning
  – Provide control for deletion/replacement
  – Defined replica validation frequency via UNF’s and Checksums
Summary

- Replication ameliorates institutional risks to preservation
- Strengthen preservation through institutional diversification
- Data-PASS requires policy based, auditable, asymmetric replication commitments
- Formalize policies in schema or rules
- Build trust models
- Data-PASS approach to preservation combines Trust Models, Institutional Collaborations and Digital Replication Infrastructures
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