ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

A Community Approach to Preservation: "Experiences with Social Science Data"

DigCCurr Spring 2009

Jonathan Crabtree

April 3, 2009

Digital PRESERVATION
National Digital Information Infrastructure & Preservation Program
A Collaborative Initiative of the Library of Congress

DATA-PASS
DATA PRESERVATION ALLIANCE FOR THE SOCIAL SCIENCES

ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

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

ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

The Plan

- Identify significant data collections (classic)
- Identify important contemporary data ("at risk")
- Develop common standards and procedures across partnership

ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

Partnership Goals

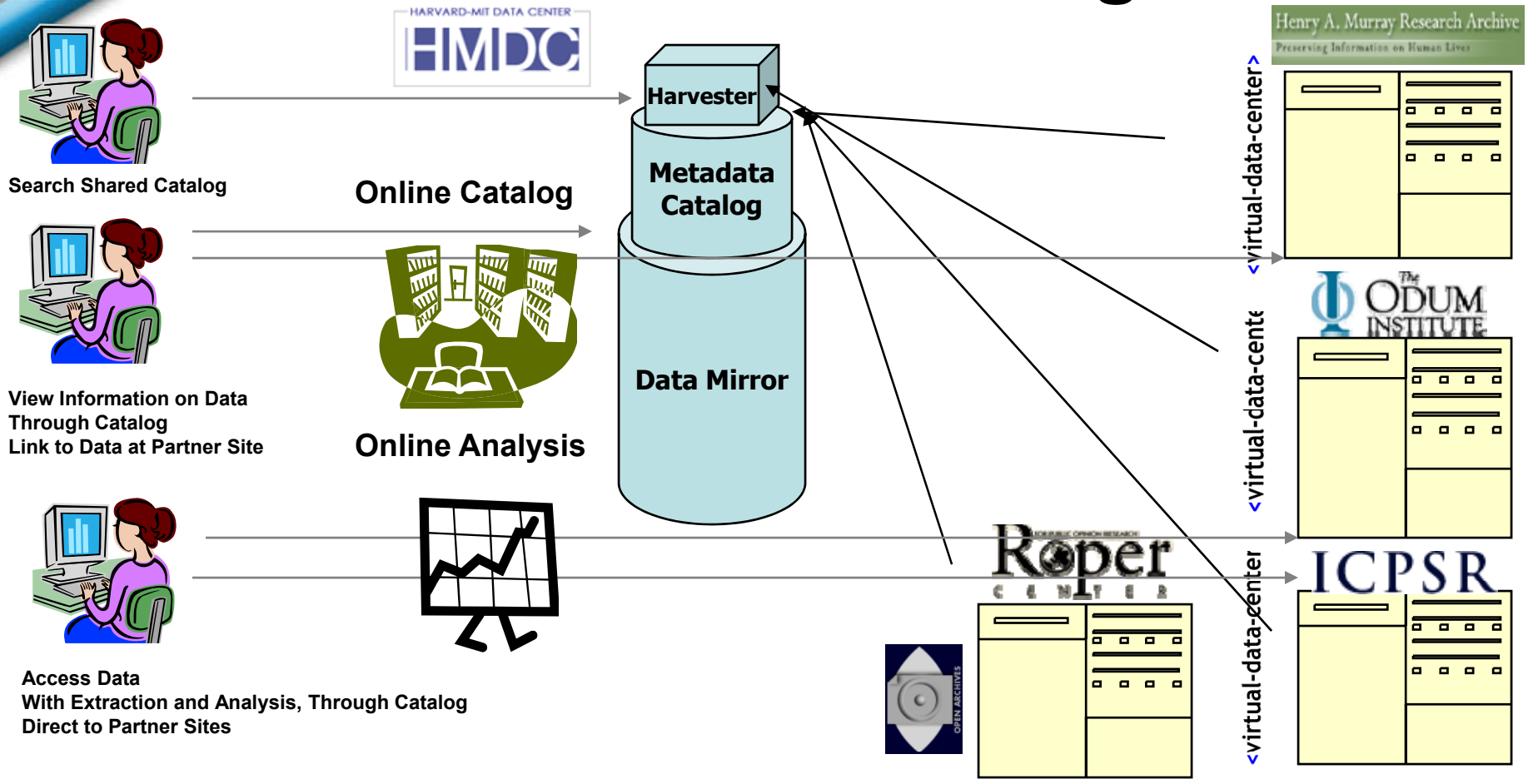
- Develop common standards and activities
- Determine how the partnership can expand
- Use technological advances to encourage metadata standards and a shared catalog

ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

Dataverse Network

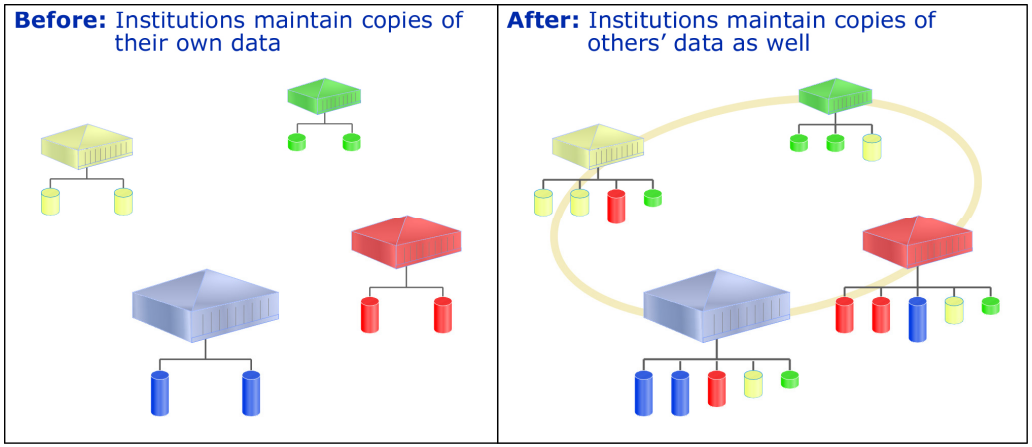
- Open source platform
- OAI server
- DDI metadata standards
- Federated Approach

Overview of How Catalog Works



ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

Multi-Archival: Syndicated Storage Platform



ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

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

ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

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

ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

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

ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

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

ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

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

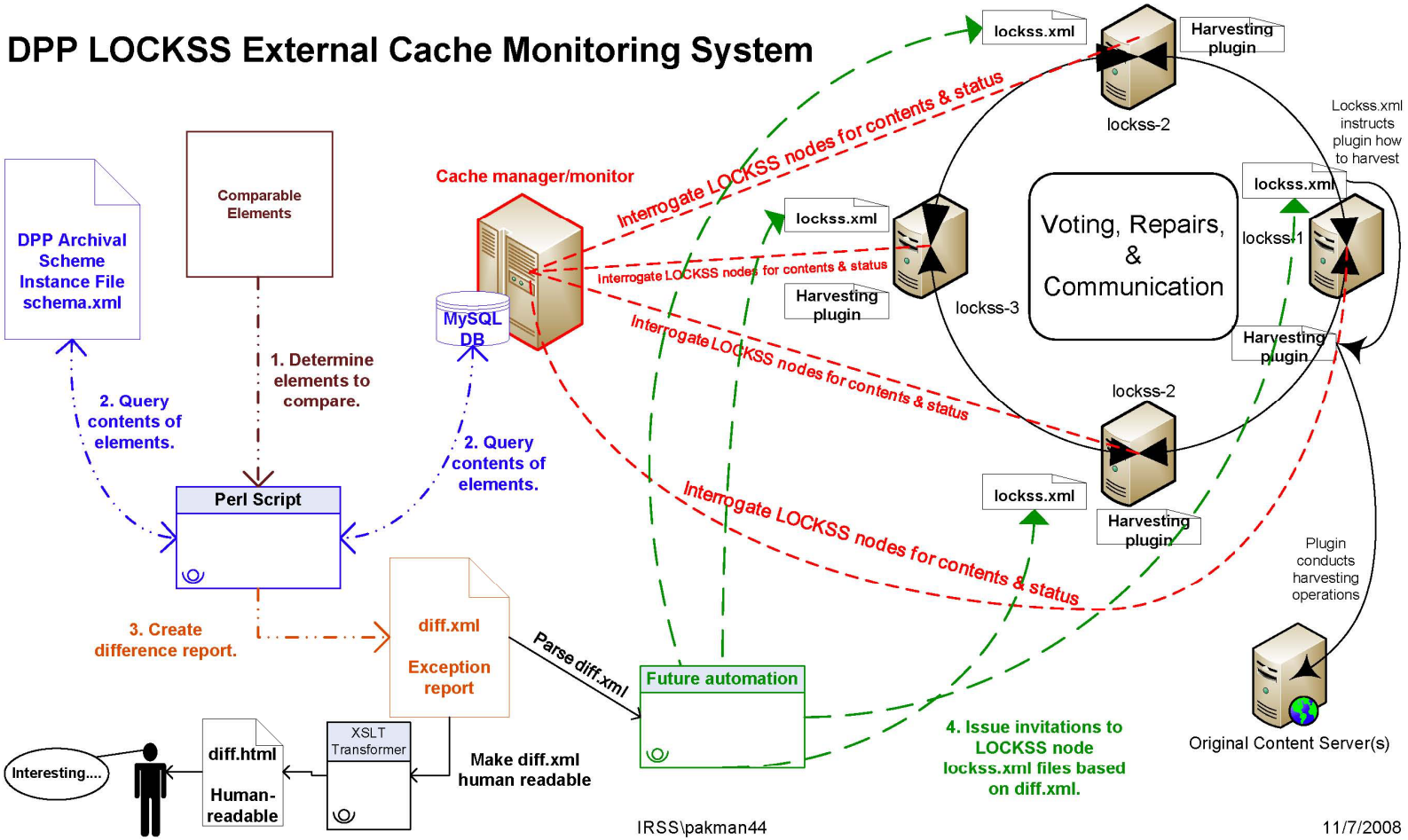
SPP Commitment Schema

- 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 use 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 elements relevance to TRAC, and mapping to particular TRAC criteria

```
- <SSP>
  - <network>
    + <networkIdentity></networkIdentity>
    - <networkCapabilities>
        <protocolVersion version="1.0"/>
        <numberReplicates min="4"/>
        <replicationFrequency maxHours="72"/>
        <verificationFrequency maxHours="72"/>
        <versioningPolicy policyType="required"/>
        <deletionPolicy policyType="required"/>
      </networkCapabilities>
    + <networkTerms></networkTerms>
    </network>
  - <hosts>
    - <host>
      + <hostIdentity></hostIdentity>
      - <hostCapabilities>
          <lockssVersion protocolVersion="1.0" softwareVersion="1.0"/>
          <storageAvailable maxGB="500"/>
        </hostCapabilities>
      + <hostTerms></hostTerms>
      </host>
    + <host></host>
    </hosts>
  - <archivalUnits>
    - <au>
      - <auIdentity>
          <name uniqueID="au1">Gallop</name>
          <description fulltextURI="http://datapass.org">Gallop Polls</description>
          <accessBase adminEmail="support@icpsr.org" accessURI="http://somesite.com/reference"/>
        </auIdentity>
      - <auCapabilities>
          <updateFrequency minDays="24"/>
          <storageRequired maxMB="1000"/>
          <pluginRequired pluginURI="http://someplugin.com"/>
        </auCapabilities>
      - <auTerms>
          <contentTermsCompliance fulltextURI="" complianceType="inAu">This is
          compliant</contentTermsCompliance>
          <disseminationTerms fulltextURI="" disseminationType="clickthrough"
          disseminationCondition="failure">Some terms</disseminationTerms>
        </auTerms>
      </au>
    + <au></au>
    </archivalUnits>
</SSP>
```

ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

DPP LOCKSS External Cache Monitoring System



ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

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

ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

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

ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

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