



How collaborative preservation works

Micah Altman, Harvard University

IASSIST 2010, Ithaca New York



What's next?

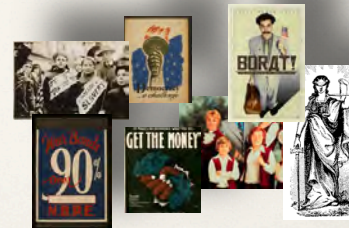
- * What is Data-PASS?
- * Challenges of preserving scientific evidence
- * Converging trends
- * Benefits of institutional collaboration
- * Evolving structure of collaboration
- * Services and infrastructure

Collaborators and Co-Conspirators

- ✦ Margaret Adams, Caroline Arms, Ed Bachman, Nitin Borwankar, Adam Buchbinder, Ken Bollen, Bryan Beecher, Steve Burling, Jonathan Crabtree, Darrell Donakowski, Myron Gutmann, Gary King, Patrick King, Jared Lyle, Marc Maynard, Amy Pienta, Lois Timms-Ferrarra, Copeland Young.
- ✦ **Research Support**
Thanks to the Library of Congress (PA#NDP03-1), the National Science Foundation (DMS-0835500, SES 0112072), IMLS (LG-05-09-0041-09), the Harvard University Library, the Institute for Quantitative Social Science, the Harvard-MIT Data Center, and the Murray Research Archive.

What is Data-PASS?

- ✦ Data-PASS is a broad-based partnership of data archives dedicated to acquiring and preserving data at-risk of being lost to the social science research community.
- ✦ Data-PASS partners have rescued thousands of data sets and created the largest catalog of social science data in existence.
- ✦ Data-PASS partners collaborate to
 - ✦ identify and promote good archival practices,
 - ✦ seek out at-risk research data,
 - ✦ build preservation infrastructure,
 - ✦ and mutually safeguard collections.
- ✦ Our current initiatives include:
 - ✦ improving data citation practices,
 - ✦ automatic policy-based archival replication



How collaborative preservation works.

Challenges of Preserving Scientific Evidence

- ♦ Scientists expectations are changing
 - ♦ Movements toward open access and open data
 - ♦ Specialized workflow systems
 - ♦ Diversity of approaches to managing replication and community data
- ♦ Scientific change creates technical challenges:
 - ♦ Forms, formats, and research workflows change
 - ♦ Data is not self-documenting
 - ♦ Intellectual property & privacy law are evolving
 - ♦ **Resources to deal with these changes are limited**
- ♦ Much of the empirical base of science becomes lost
 - ♦ Journal articles & books are only summaries
 - ♦ Full replication is expensive or impossible
 - ♦ **This slows scientific progress:** cooked results, publication bias, citation authority distortion, challenges of meta-analysis



Source: Wikimedia Commons

How collaborative preservation works.

Converging trends in preservation

- ✦ Standardized criteria for evaluating trustworthiness of archives
 - ✦ TRAC; NARA TDR; Drambora
- ✦ Collaborative stewardship by memory institutions
 - ✦ Meta-Archive, CLOCKSS, COPUL, PeDALS, ADN, Chronopolis
- ✦ Technology for replication and verification
 - ✦ Solutions developed within the library / archival community:
LOCKSS, IRODS, ACE, Duraspace
 - ✦ Commercial HPC and Cloud solutions:
Hadoop, Crashplan, Mozy, AWS, etc.
 - ✦ P2P sharing:
freenet, grunet, Tahoe-LAFS

Benefits of Collaboration

"Nothing new that is really interesting comes without collaboration" -- James Watson

- ✦ General Benefits
 - ✦ Exposure to funding opportunities; collection development leads
 - ✦ Division of labor in tracking law, technology, information science
 - ✦ Combined experience in preservation practice
- ✦ Data-PASS Focus*
 - ✦ Expanded discoverability of collections
 - ✦ Reach new audiences
 - ✦ Holdings across the joint collection are more complete
 - ✦ Virtual collections can be built from slices of the joint collection
 - ✦ Development and advocacy of archival good practices
 - ✦ *(Current initiative: outreach to professional associations in support of data citation)*
 - ✦ Insurance against institutional and technological failure

How collaborative preservation works.

** And the museum of obsolete data storage technologies*

How Collaborative Stewardship acts as Insurance Against Preservation Failure

✦ *Collaborative replication & stewardship can substantially mitigate preservation risk from:*

✦ External threats to institution failure:

- ✦ funding loss; attacks;
legal regime change;
mission drift

✦ Institutional failure:

- ✦ Unintentional curatorial modification
Loss of institutional knowledge;
Change in mission

✦ And also reduce preservation risk from:

- ✦ Media failure (from storage & media characteristics);
Software & hardware infrastructure failures



How collaborative preservation works.

Shared Infrastructure

- ✦ Shared infrastructure can
 - ✦ reduce costs
 - ✦ reduce risk
 - ✦ coordinate operations
 - ✦ validate shared standards
- ✦ Data-PASS Shared Infrastructure
 - ✦ Shared Catalog
 - ✦ Policy-Driven Distributed Replication (in development)
 - ✦ The Dataverse Network (overlapping infrastructure)



How collaborative preservation works.

Shared Catalog

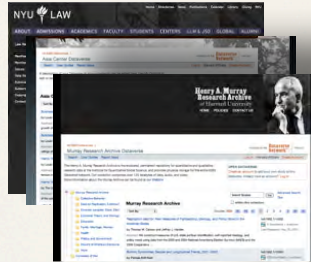
- ♦ Unified Discovery
 - ♦ Simple & fielded search
 - ♦ Virtual collection across entire catalog
 - ♦ Browse by subject, data, source
- ♦ Metadata delivery
 - ♦ Descriptive study, file, and variable information
 - ♦ Provenance & rights metadata
 - ♦ Human, OAI, Z39.50 interfaces
- ♦ Layered Services
 - ♦ Data reformatting for delivery
 - ♦ On-line analysis

How collaborative preservation works.

The screenshot shows the Data-PASS DataPASS Database website. The header includes the Data-PASS logo and the tagline 'DATA PRESERVATION ALLIANCE for the SOCIAL SCIENCES'. Below the header is a navigation menu with options like 'ABOUT THE PROGRAM', 'ABOUT THE PARTNERS', 'DATA', 'PUBLICATIONS & PRESENTATIONS', 'NEWS & EVENTS', 'MEMBERSHIP AND SUPPORT', and 'CONTACT US'. The main content area features a search bar, a 'DataPASS Database' title, and a list of 'DataPASS Collected Studies'. The list includes entries such as 'Economic State Poll, December 1997' and 'Economic State Poll, December 2007', each with a brief description and a 'Download' button. The footer of the page contains the number '10'.

The Dataverse Network[®]

For Organizations



- ✦ *Dataverses are Data-PASS ready* -- all dataverses can provide:
 - ✦ DDI (2.x) metadata export (intuitive form-based entry)
 - ✦ Catalog access through OAI-PMH (and Z39.50)
 - ✦ LOCKSS compatibility
 - ✦ Version control (**new**); Terms of use metadata; Flexible contributor-curator-editor workflows

ideal for "living collections" & (better) self-archiving

For Scholars



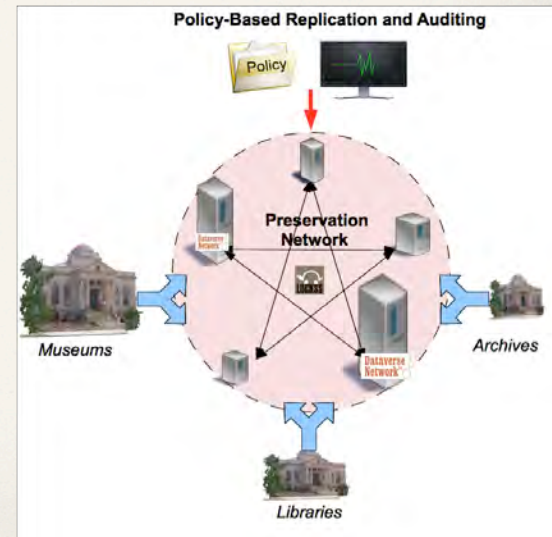
- ✦ The Dataverse Network System is Open-Source *and*
- ✦ Creating a Dataverse requires no software.
- ✦ IQSS & MRA host an open DVN and offer no-cost permanent storage:

<http://dvn.iq.harvard.edu>

Policy-Driven Distributed Replication

- ♦ Policy Based
 - ♦ Preservation requirements shape policy
 - ♦ Policy drives replication rules
 - ♦ Auditing demonstrates conformance with preservation requirements
- ♦ Copies are distributed
 - ♦ Across space
 - ♦ Among institutions
 - ♦ Across time (version history retained)
- ♦ Commitments scaled to participant resources
 - ♦ Collection size
 - ♦ Technology

How collaborative preservation works.



Structure of Collaboration

Areas of collaboration...

- ♦ Partnership agreements
 - agreement on good practice;
permission to preserve;
partners offer to accept data transfer if archive fails
- ♦ Coordinated operations
 - shared leads;
regular communication;
collegial review available
- ♦ Shared good practice
 - metadata; preservation; confidentiality
- ♦ Circle of gifts norm
 - in-kind effort & resource;
contributions are voluntary & proportional

Steps to participation

Partners agree to...

- ♦ Publishing metadata
- ♦ Use of replication system
- ♦ Good archival practice
(TRAC compliance *not* required)
- ♦ Transfer protocols

Partners use the following technologies

- ♦ Light-weight protocols:
OAI-PMH + DDI 2-lite +
HTTP harvestable data
- ♦ Software:
Could use a hosted dataverse or;
install open source OAI-PMH server, etc.
- ♦ **No fear - we can help!**

More Questions?



- * Know of research data at risk of loss?
- * Need help preserving your research data?
- * Want more visibility and protection for your collections ?

<http://data-pass.org>

data-pass@icpsr.umich.edu