Data-PASS Shared Catalog

Micah Altman
Harvard University

Archival Director, Henry A. Murray Research Archive
Associate Director, Harvard-MIT Data Center
Senior Research Scientist, Institute for Quantitative Social Sciences

E: micah_altman@harvard.edu
W: http://maltman.hmdc.harvard.edu/

Jonathan Crabtree
University of North Carolina

Assistant Director for Archives and Information Technology
HW Odum Institute for Research in Social Science

E: Jonathan_Crabtree@unc.edu
W: http://www.odum.unc.edu

Collaboration for Preservation

The ODUM INSTITUTE
Henry A. Murray Research Archive
Preserving Information on Human Lives
Roper CENTER FOR PUBLIC OPINION RESEARCH
HARVARD-MIT DATA CENTER HMDC
NARA
ICPSR

- Strategic Partnership Agreements
- Coordinated Operations
- Joint “not-bad” practices
- Shared catalog
- Shared tools & technologies

Technical Collaboration

Shared Catalog
- Unified Discovery
- Content exchange
- Layered Services

Shared Technologies & tools
- Schema's and crosswalks
- Fingerprint and persistent identifiers
- Digital libraries and ingest tools
- Storage and replication

"Not-bad" practices and Standards
- Identification & selection
- Metadata
  - Cataloging
  - Exchange
- Security
- Confidentiality
- Citation

Shared Catalog
- Unified Discovery
- Content exchange
- Layered Services

Data-PASS Shared Catalog

- A unified catalog of the partners' *entire* holdings
- Completes the unification of social science data that was the dream of the first Council of Social Science Data Archives in 1969

- Discovery Services
  - Simple & fielded search
  - Virtual collection browsing
- Metadata delivery
  - Descriptive study, file, & variable information
  - Provenance metadata
  - Human and OAI interfaces
- Enhanced Delivery
  - Proxy delivery
  - Replication
  - Layered analysis services

Finding Data

- Search Across Entire Partners' Catalogs
- Find Studies Collected for Data-PASS
- Simple and Fielded Search
- Browse by Subject, Date, Source

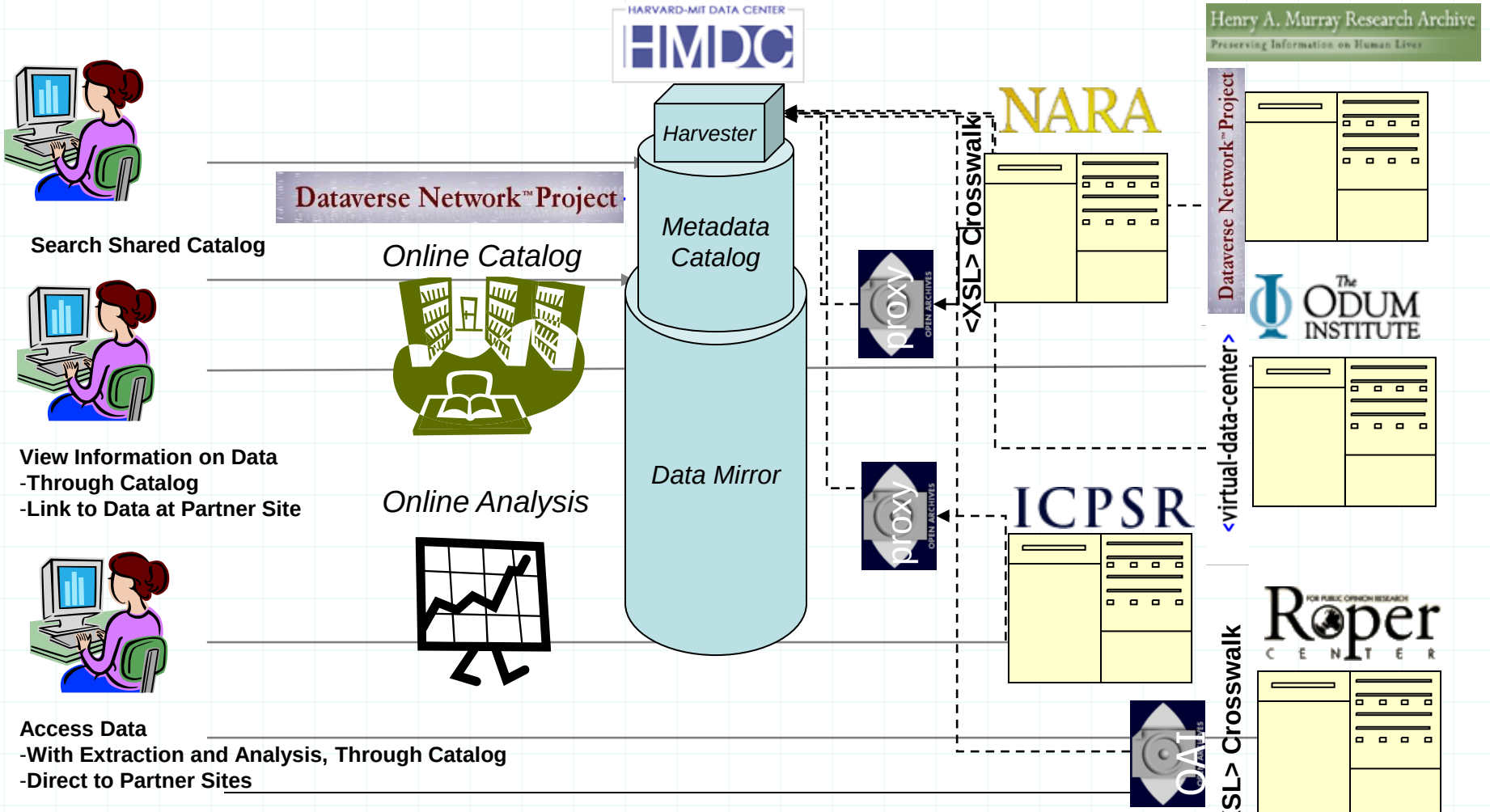
Delivering Data

- Through Partners' Sites
  - Shared catalog results always give link to data at partners site
  - If no file information supplied to catalog, this is the only option
- Through Shared Catalog
  - Catalog server may cache a copy of data for performance
  - Catalog can bundle requests for multiple files
- Through Analysis Services
  - If partner site runs DVN(or data access proxy), analysis and extraction is available
  - Download data in multiple formats
  - Extract subsets, in multiple formats, with citations and UNF's
  - Run descriptive stats, crosstabs
  - Advanced analysis -- dozens of statistical models

Enabling Technologies

- Metadata harvesting:
  - OAI-pmh
- Metadata standards and tools:
  - DDI
  - XSL
- Citation, validation:
  - Handles
  - UNF
- Federated Search, Virtual Archives:
  - Dataverse Network
  - OAI Servers

Catalog Distributed Architecture



Micah Altman & Jonathan Crabtree

Metadata Harvesting

- Each partner catalog is exposed via
  - Dataverse Network via OAI
  - Other OAI Server, running on-site
  - Proxy OAI Server, running at HMDC
    - Harvested ad-hoc
    - XSL Metadata to cross-walk applied
    - Made available through OAI
- DDI-lite schema subset used for exchange
  - Data Documentation Initiative (DDI) – international effort to establish specification schema for the content, presentation, transport, and preservation of documentation for datasets in the social and behavioral sciences
  - Provenance, and structural metadata, including: document description (meta-meta data), study description, file description, variable description

http://www.icpsr.org/DDI/

The Dataverse Network

Includes integrated developments in **web application** software, **networking**, data **citation** standards, and **statistical methods** designed to put some of the universe of data and **data sharing** practices on firmer ground. It facilitates the public **preservation** and **distribution** of persistent, authorized, and verifiable research data.

Virtually-Hosted Archiving

- The importance of being virtual …
  - Nothing to install
  - Dynamic collections: local and federated
- Institutionally supported
  - Persistent identifiers and citations
  - No worries about file formats changing, backups, etc.
  - All the initial setup work is done for depositor
- Depositor retain total control over
  - Content
  - Access
  - Presentation

Benefits to collaboration

- Combine and blend strengths
- Bring different perspectives to the table
- Coordinate on key issues, e.g., syndicated storage
- Share knowledge and experience to develop tools and future standards

Archivists & Catalogers

- Benefit from shared workflows
- Participate in software design to enhance ingest
- Potential for increased submissions

IT Administration Perspective

- Standards based collaborations are less risky
  - More recovery paths
  - More resources to solve problems
- Collaboration provides larger test audience for software development
- Lowers developmental cost

What do data consumers say?

- Enjoy the simplicity of a “common catalog”
- Variable level searches are powerful
- Browsing the data with descriptive statistics helpful
- Excited about the advance online statistics

Benefits of Virtual Archiving

- Promotes self archiving
- Potential to reach investigators early in the data lifecycle
- Allows for professional subject area based curation
- Customized branding for producers
- Lowers the barriers to submission and in turn increasing data deposit rates

Collaboration for Preservation

- Objects protected again single institutional failure
- Standards based metadata
- Collaborations offer potential for replicated and geographically diverse distributed storage
- Collaborations may offer small archives the only way to become a “trusted archive”
- Collectively dedicated to the long-term survival of the resource

Collaboration Strengths

- Over 200 years combined experience in social science data preservation
- Innovative archival software developed uniquely for the ingest, presentation, location, analysis, and preservation of social science data
- Institutional dedication to the distribution and preservation of social science data

For More Information

Data-PASS Project:

http://www.icpsr.umich.edu/DATAPASS/

Shared Catalog:

http://dvn.iq.harvard.edu/dvn/dv/datapass/

Dataverse Network Software:

http://TheData.Org