The Dataverse Network – Combining Preservation with Scholarly Recognition

Mercè Crosas, Ph.D.
Director of Product Development
Institute for Quantitative Social Science (IQSS)
Harvard University
Acknowledgements

This work is done with contribution from:

• The Dataverse Network software development team at IQSS: Gustavo Durand, Ellen Kraffmiller, Kevin Condon, Leonid Andrev, Bob Treacy, Michael Heppler, Steve Kraffmiller
• Gary King, Albert J. Weatherhead III University Professor, Harvard University
• Micah Altman, Senior Research Scientist, IQSS, Harvard University
The Problem with Data

Professional archives focus on long term access by the wider community

- Persistent identifiers
- Fixity
- Backups and recovery
- Metadata standards
- Conversion standards
- Preservation standards

Data owners focus on recognition and control

- Branding and visibility
- Data discovery
- Ease of use
- Scholarly citation
- Control over updates
- Terms of access and use

... but do not attract maximum contribution from data owners

... but do not assure long-term use as would a professional archive
Centralized Archival Infrastructure

- Persistent identifiers
- Fixity
- Backups and recovery
- Metadata standards
- Conversion standards
- Preservation standards

Distributed Recognition and Control

- Branding and visibility
- Data discovery
- Ease of use
- Scholarly citation
- Control over updates
- Terms of access and use

The Dataverse Network

Enables people to solve a political problem through technology – Combining centralized archiving with distributed ownership
Use Case 1: Dataverse for Researchers – Web Visibility and Branding

Your website

Your dataverse

Sidney Verba
Carl H. Pforzheim University Professor
Department of Government, Harvard University (email)
Use Case 1: Dataverse for Researchers – Recognition through Citation

Credit to Authors
Credit to Distributor
Use Case 1: Dataverse for Researchers – Control and Ownership

- The researcher is the administrator of his/her dataverse.
- Manages data curation workflow.
- Sets permissions and terms of use to datasets.

**A Researcher …**

- Creates Dataverse
- Adds Study
- Sets Permissions
- Releases Study
- Enters Cataloging
- Terms of Use
- Uploads Data Files
Use Case 2: Dataverse for Journals – Replication of Published Work
Use Case 2: Dataverse for Journals – From Publication to Data

Journal, Book

Datasets, text, images

Formal Data Citation

Persistent identifier and url that never changes

Universal Numerical Fingerprint (UNF) to verify dataset
Use Case 2: Dataverse for Journals – Submission and Review Process

- A Journal for replication data can be set as an open dataverse.
- Authors upload data after registering.
- Journal reviews data before releasing it.

Author Adds Study
  Enters Cataloging
  Uploads Data Files
  Sends Study for Review
  Curator reviews data
  Releases Study
Use Case 3: Dataverse for Archives – Evidence is Never Lost

The Dataverse provides:

- **Versioning** – Keeps all changes to data and metadata to allow citation of old versions.
- **Deaccession** – Doesn’t delete studies permanently so citation is valid forever.
- **Universal Numeric Fingerprints** – Fixity that survives changes of format.
Preservation and verifiable format by:

1. Extracting variable and format metadata.
2. Converting dataset to a format independent of software package.
3. Applying cryptographic algorithm to canonical format.
4. Obtaining an alphanumeric string based on semantic content:
   UNF5:EKgHvTNfkkS86dNzABlhNw==
Use Case 3: Dataverse for Archives – Standards and Backups

The Dataverse facilitates:
- Multiple copies of data
- Metadata exchange with other systems.

- Export metadata in multiple formats
- Harvest metadata with OAI-PMH
- Replicate data and metadata with LOCKSS

DDI
Dublin Core
FGDC
Marc

Dataverse Network

Dataverse Network

Archival Site
New: A ‘wiki’ dataverse is an archive to share and improve data through contributions (with a moderator to review them before their release)
Use Case 4: Wiki Dataverses – Multiple Roles and Workflows

- Multiple contributors can be updating the same study.
- Changes are only visible after being reviewed.
- A new way to gather more data for a study and improve its cataloging.
Current Usage

- The IQSS Dataverse Network at Harvard holds:
  - 257 public dataverses from around the world
  - 36,000 studies (mostly in Social Science)
  - 640,000 files
- ... and there are over 10 more installations of the Dataverse Network at other Institutions.
How to Get Started …

- Create a dataverse at: http://dvn.iq.harvard.edu
- Or install the **Dataverse Network** software for your Institution.
- For more information, go to: http://thedata.org
- Contact me at: mcrosas@hmdc.harvard.edu

THANK YOU!