Dataverse Network
A Data Sharing System

Merce Crosas (mcrosas@hmdc.harvard.edu)

Director of Product Development

Institute of Quantitative Social Science (IQSS)

Harvard University

A long history of data sharing has yielded revolutionary impacts

Galileo's Jupiter observations

Night 1

Night 2

Tycho Brahe's Mars Observations

Declination (degrees)

1582 1584 1586 1588 1590 1592 1594 1596 1598 1600

Image Copyright 2000, Wayne Pafko

1582–1600 Tycho Brahe collects data of mars position

1605 Kepler infers three laws of planetary motion based on Brahe's observations

1610 Galileo publishes observations of the moons of Jupiter

1687 Newton publishes the theory of universal gravitation based partly on above observations

Who is sharing now?

In the natural and physical sciences, many Journals require authors to share their data

Nature's Policy on availability of data and materials:

"An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. Therefore, a condition of publication in a Nature journal is that **authors are required to make materials, data and associated protocols promptly available to readers without preconditions**. …. The preferred way to share large data sets is via public repositories".

Who is sharing now?

In health research, funding agencies are requiring grantees to share their data

Since 2003…

"NIH reaffirms its support for the concept of data sharing. We believe that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health. The **NIH endorses the sharing of final research data** to serve these and other important scientific goals".

Who is sharing now?

In Social Sciences, organizations and archives are encouraging scholars and data collectors to share their data

Council of European Social Science Data Archives on sharing data benefits…

- "Sharing data reinforces open scientific inquiry, allowing effective self-correction of research; secondary analysts can verify, refute, or refine original results.
- It facilitates high-quality, policy-relevant research.
- Sharing encourages diversity of analysis and opinions, and of a multiplicity of perspectives.
- Sharing promotes new research and allows for the testing of new or alternative methods.
- Sharing data reduces costs by avoiding duplicate data collection efforts.
- …"

Sharing data increases citation of researcher's published work

Distribution of 2004–2005 citation counts of 85 clinical trials by data availability. (the box encompasses the interquartile range of the citation counts, whiskers extend to 1.5 times the interquartile range, and lines within the boxes represent medians.)

Piwowar HA, Day RS, Fridsma DB 2007 Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308 doi:10.1371/journal.pone.0000308

Increasing volumes of information and new media demands digital libraries for effective sharing

Amount of "published" Information

Historical media for sharing:
- Cuneiform
- Papyrus
- Books
- Digital Objects

~160 TB

Berkeley's study, 2003:
www information
increased 30% per year
from 2000 to 2003

~25-50 TB

Ancient Alexandria Library
(700,000 scrolls)

(large # of TB today?)

Library of Congress
(~11TB):
- 120 million items
- 18 million books
- 500 miles of shelving

Cuneiform
(thousands)

| 2500BC | 288BC | First Printed Book 1455 | ARPnet 1969 | www 1989/1991 | 2000 | 2003 |

The Dataverse Network provides a digital library for research data that addresses concerns about sharing

Requirements	Our Solutions
Author's credit and persistent reference to data	New data citation standard for study/dataset
Control by author, department or project group	Each entity can have an individual "virtual archive" (dataverse)
Author's and organization's recognition	Branding to match author's or organization's website
Privacy/Restrictions	Different levels of study and file permissions set by data owner
Cataloging (as in traditional library)	Extensive study metadata, easy to enter (self-archiving)
Support for heterogeneous data formats and sources	Store data files individually, not in centralized data table
Preservation/Interoperability	Convert to preservable/exchangeable formats automatically
Data Safety	Professional archiving safer than in researcher's computer

A data citation is created for each research study in the Dataverse

Allows to cite research data from published work

Printed Citation for published work:

Verba, Sidney; Nie, Norman H., 1984, "Political Participation in America, 1967",
The Social Science Journal, v1, p.1

On-line citation for research data:

Verba, Sidney; Nie, Norman H., 1984, "Political Participation in America, 1967",
hdl:1902.2/7015 UNF:3:+DNr7jVq/5XmsPAmls4KQg==
Inter-university Consortium for Political and Social Research [Distributor]

hdl:1902.2/7015 → Persistent Identifier
UNF:3:+DNr7jVq/5XmsPAmls4KQg== → Universal Numerical Fingerprint

1. Persistent identifiers: permanence and reliability to citation of digital objects

Examples of Persistent Identifiers:
1. Handle System
2. Digital Object Identifier (DOI) – Based on handles, but with cost
3. PURL
4. Uniform Resource Names (URN)

By default, Dataverse uses **handles** because of low cost, good implementation and services:

<Handle> ::= <Handle Naming Authority> "/" <Handle Local Name>

10.1045/january99-bearman

Naming Authority
Prefix

Unique Local Name
Suffix

2. Universal Numerical Fingerprint (UNF): used to uniquely identify and verify data

$$\begin{pmatrix} 1 & 4 & 4 & 21 & \cdots & 121 \\ 1 & 2 & 2 & 91 & \cdots & 212 \\ 1 & 9 & 2 & 72 & \cdots & 104 \\ 0 & 2 & 2 & 2 & \cdots & 321 \\ 1 & 6 & 2 & 12 & \cdots & 204 \\ 1 & 9 & 4 & 52 & \cdots & 311 \\ 0 & 3 & 2 & 23 & \cdots & 92 \\ 0 & 2 & 5 & 91 & \cdots & 212 \\ 0 & 5 & 8 & 91 & \cdots & 91 \\ 1 & 9 & 1 & 72 & \cdots & 104 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & 2 & 91 & \cdots & 212 \end{pmatrix}$$

- Apply a cryptographic algorithm
- Solely based on semantic contents of the digital object:
  - data changes result in new UNF
  - format or location changes retain original UNF
- Final alphanumeric string:
  - uniquely summarizes the contents,
  - but does not convey its information

ZNQRI14053UZq389x0Bffg?==

Dataverse Network ™

The "Network" offers an extensive and flexible functionality to organize data

Scholars get an individual "archive" with a dataverse

Scholar's website

Scholar's Dataverse

And the same for Departments, projects or centers

Department's or project website

Department's or project Dataverse

What can you do with a dataverse?

Features	Admin	Curator	Contributor	End-user
Search Studies and Browse collections	X	X	X	X
Advanced Search (by metadata field)	X	X	X	X
View metadata and download files	X	X	X	(depending on permissions)
Subset and analyze quantitative data	X	X	X	(depending on permissions)
Create Study and upload files	X	X	X	
Set Permissions to Study	X	X	X	
Release Study	X	X		
Update Study after Release	X	X		
Organize Studies by Collections	X	X		
Brand dataverse	X			
Add new admins, curators and contributors	X			
Change Settings (contact e-mail, dataverse name, etc)	X			
Release dataverse when ready	X			

Studies provide multiple permission levels controllable by the data owner (admin, curator or contributor)

Completely Public	Public but with Agreement	Files Restricted	Entire Study Restricted
All users can access entire study - both cataloging information (metadata) and files -	All users can access entire study, but need to agree to terms of use to download the files.	All users can access metadata, but files are only available to a set of user: - Authorized by dataverse login - Authorized based on IP address	Only a set of users can access metadata and files. This means that the study can not even be found by searching for non-authorized users.

Authors can self-archive their studies with extended cataloging options

Cataloging Information

* Required Fields * Recommended Fields A light blue backgro
A light orange form indicates that a Date format is required. Please
⊕ Add a new row (e.g., when add multiple authors). ⊖ Remove an e

○ Show Required and Recommended Fields ⦿ Show All Fields

Citation Information

Title *

Subtitle

Study ID * hdl:1902.1/ 11839

Other ID
Other ID
Other ID Agency

Author *
Name *
Affiliation

Producer *
Producer *
Affiliation
Abbreviation
URL (Enter full url, e.g., http://...)
Logo URL (Enter full url for image, http://...)

Production Date * (Enter date as YYYY or YYYY-MM or YYYY-MM-DD)

- Up to 100 metadata fields to describe each study, including:
  - geospatial information
  - data collection and methodology
  - terms of use and conditions
- Only study title and id are required
- All fields are searchable
- Templates with pre-defined fields

A dataverse supports ANY file type, but offers additional services to quantitative data files

- Author can upload any file type or format of audio, images, video, text, etc

- Additional features are available to subset and analyze quantitative data files (SPSS, STATA files)

- File metadata includes:
  - File name
  - description
  - category

Cataloging Information | **Documentation, Data and Analysis**

Access to some files is restricted

Download All Files Note: you will be prompted to save a single arc... format.

▼ 1. Documentation
00022Colby-Responsibility-StudyDescription.pdf   Do...
Adobe PDF - 10 KB - 0 downloads

▼ 2. Data   Files in this category are restricted
00022Colby-Responsibility-Data.por   Restricted Access   Midlife in the United States Data in SPSS Portable Format
SPSS Portable - 435 KB - 0 downloads

00022Colby-Responsibility-Data.tab   Restricted Access   Midlife in the United States Data in Tab Delimited Format
Tab Separated - 305 KB - 0 downloads + analyses

SUBSETTING | 94 Cases | 1445 Variables   Access Subset/Analysis   View Data Citation
Anne Colby, 1983, "Social Responsibility at Midlife", hdl:1902.1/00022
UNF:3:N3JFv0pJ64KJWJ5tDY8ksQ== Murray Research Archive [Distributor]
00022Colby-Responsibility-Data.tab [fileDscr/fileName (DDI)]
Collapse [-]

▼ 3. Supplementary Documentation   Download all files in this category
00022Colby-Responsibility-CodebookAddendum.pdf   Download   Clarifications of discrepancies in the data files
Adobe PDF - 9 KB - 0 downloads

▼ 4. Detailed Usage Terms   Download all files in this category
00022Colby-Responsibility-MemoOfAgreement.pdf   Download   Legal agreement between data depositor and Murray Archive
Adobe PDF - 440 KB - 0 downloads

A rich set of data analysis based on R statistical package

- Download a subset of variables

- Recode a variable

- Apply descriptive statistics or and advanced statistical models (from Zelig/R)

| Download Subset | Recode and Case-Subsetting | Descriptive Statistics | **Advanced Statistical Analysis** |

Selected Variables
```
ID
QL2
HH_CELL
```

ℹ Select variables from table below (s...

– Choose a Statistical Model–

– Choose a Statistical Model–

Categorical Data Analysis
  Cross-Tabulation

Ecological inverence models
  Hierarchical Multinomial-Dirichlet Ecological Inference Model for R x C Tables

Event Count Models
  Negative Binomial Reg for Event Count Dep Vars
  Poisson Reg for Event Count Dep Vars

Models for Continous Bounded Dependent Variables
  Exponential Reg for Duration Dep Vars
  Gamma Reg for Cont, Positive Dep Vars
  Log-Normal Reg for Duration Dep Vars
  Weibull Reg for Duration Dep Vars

Models for Continous Dependent Variables
  Fit an Analysis of Variance Model

	Variable Type
☑	Discrete
☑	Discrete
☐	Discrete
☑	Discrete
☐	Discrete
☐	Discrete

NPRB_SEX   probability of selection - gender
GENDER   gender of R

Dataverse Network

The Dataverse Network software supports data archive standards for interoperability

Open Archive Initiative: Protocol for Metadata Harvesting (OAI-PMH)

Data Documentation Initiative (DDI): standard in XML for metadata describing social science data

Dublin Core: bibliographic metadata standards for describing resources

Handles: Persistent Identifiers

Metadata and files are converted automatically for preservation and interoperability

Author enters study metadata using a web form

Dataverse exports metadata from database into an XML format (DDI and Dublin Core) for preservation and harvesting

Author uploads quantitative data file in statistical format (SPSS, STATA)

Dataverse converts into a plain tabular file + variable metadata (independent of statistical package, versions, etc)

J2EE architecture is at the core of the Dataverse Network software

Multiple layers for maintainability and scalability:

• JSF: User interface layer

• EJB: Business logic layer

• OAI client and server: For harvesting metadata

•Lucene: Index server

• PostgreSQL: Database for persitence storage of metadata

• File System: For storage of data data + complementary files

• R Serve: Analysis component for quantitative data file

IQSS Dataverse Network Case Study

- Project was initiated 3 years ago (based on the Virtual Data Center previously implemented by Harvard-MIT Data Center)

- In production for Harvard and MIT for more than 1 year, with ~ 300-500 users per day.

- Post-release adoption was rapid: ~100 dataverse owners within 6 months.

- We have now about 160 dataverses, with 30,000 studies, 500,000 files, including datasets from archives and from individual scholars from universities around the world.

- About 100 additional dataverses currently in preparation

The Dataverse Network expansion outside Harvard/MIT is underway…

• ODUM at University of North Carolina (In production)

• Woods Hole Oceanographic Institution (In Testing)

• Inter-University Consortium for Political and Social Research, University of Michigan (In testing)

• Data Archiving and Networked Services at the Netherlands (in testing)

• University of Utrecht (in testing)

• Australian Social Science Data Archive (In evaluation)

• South Africa National Archives and Records Service (NARS) (In evaluation)

And Finally, It is Free, Open Source Software

• Number of open source projects growing at an exponential rate.

• More and more organizations and companies are using open source for every day operations.

Graph of total number of open source projects

How to contribute (New):

• Advisory Committee: Open to **active users** of the software (installing a Dataverse Network for their organization, own a dataverse).

• Technical Committee: Open to **developers** who are participating in the design and implementation

What's next?

- Expand features to more file types (UNF, convert to preservable format, etc)
  - For other quantitative file formats in addition to SPPS, STATA
  - For qualitative file formats
- Data Visualization:
  - Geospatial tools for locating data
  - Graphical representations of datasets
- Expand support to health, biomedical data and other research fields (additional metadata, ontologies)
- User comments, data versioning, and other added-value features
- Remote authorization (Shibboleth?)
- And more …

References and Acknowledgements

- **http://thedata.org**

- Development team at IQSS, Harvard University:
    Ellen Kraffmiller, Gustavo Durand, Kevin Condon, Leonid Andrev, Wendy Bossons, Akio Sone, Michael Heppler, Isabelle Chopin, Elena Villalon
- Gary King, An Introduction to the Dataverse Network as an Infrastructure for Data Sharing, Sociological Methods and Research, 32, 2 (November, 2007): 173–199.
- Micah Altman and Gary King. A Proposed Standard for the Scholarly Citation of Quantitative Data, D-Lib Magazine, 13, 3/4(March/April, 2007).
- Other contributors to project: Bob Treacy and Ann Starkey