Sharing Data with Dataverse

Mercè Crosas, Ph.D.
Director of Product Development
Institute for Quantitative Social Science (IQSS)
Harvard University
1. Data Sharing and Replication

2. A Solution with the Dataverse Network
SHARE your data
it’s good for you, and for the world.

Come. Eat lunch. Accelerate the pace of science.

CfA, PHILLIPS Auditorium, 11:45 MONDAY 4/2/12

theastrodata.org

Slide acknowledgment: Alyssa Goodman
From Data Sharing to Replication

“The replication standard holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author.”

King, G. 1995 “Replication, Replication”
“Sufficient Information”? 

- codebook
- field data
- parameters
- codebooks
- external data
- workflow description
- script
- refined data
- aggregated data
- discussion with experts
- corrected data
- visualization
- presentation
- notes
- blog posting
- publication

Slide acknowledgment: Andrea Goethals
“Sufficient Information”? 

- codebook
- field data
- parameters
- codebooks
- external data
- workflow description
- script
- refined data
- aggregated data
- discussion with experts
- corrected data
- visualization
- presentation
- blog posting
- publication
- tool descriptions
- notes

= data
= context (documentation)
= research products

Slide acknowledgment: Andrea Goethals
At the very least, need sufficient documentation to judge the veracity and usefulness of the data.
In a survey of more than 2,000 American psychologists scheduled to be published this year, Leslie John of Harvard Business School and two colleagues found that 70 percent had acknowledged, anonymously, to cutting some corners in reporting data.

... an analysis of 49 studies appearing Wednesday in the journal PLoS One, by Dr. Wicherts, Dr. Bakker and Dylan Molenaar, found that the more reluctant that scientists were to share their data, the more likely that evidence contradicted their reported findings.

“We have the technology to share data and publish our initial hypotheses, and now’s the time.”
What can happen if we share...

"We found that cancer clinical trials which share their microarray data were cited about 70% more frequently than clinical trials which do not."
1. Data Sharing and Replication

2. A Solution with the Dataverse Network
The Dataverse Network is a repository for research data that takes care of long term preservation and good archival practices, while researchers keep control of and get recognition for their data.
Dataverse Network for Astronomy Data

This is the Astronomy data repository for Harvard affiliates. The project is a collaboration of the Seamless Astronomy group at the Harvard-Smithsonian Center for Astrophysics, the ADS, the Wolbach Library, and IQSS with support from the FAS Science Research Computing.

All Dataverses

Dataverses: 6 | Studies: 68 | Files: 547

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
<th>Released</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>CfA Library Datasets</td>
<td>Harvard-Smithsonian Center for Astrophysics</td>
<td>Aug 17, 2012</td>
<td></td>
</tr>
<tr>
<td>theastrodata.org</td>
<td>theastrodata.org</td>
<td>Apr 2, 2012</td>
<td></td>
</tr>
<tr>
<td>Soderberg, Alicia</td>
<td>Harvard University</td>
<td>Feb 6, 2012</td>
<td></td>
</tr>
<tr>
<td>Astrophysics of galaxies &amp; quasars</td>
<td>Harvard-Smithsonian Center for Astrophysics</td>
<td>Oct 12, 2011</td>
<td></td>
</tr>
<tr>
<td>COMPLETE</td>
<td>Harvard-Smithsonian Center for Astrophysics</td>
<td>Jun 23, 2011</td>
<td></td>
</tr>
<tr>
<td>1.2 Meter CO Survey</td>
<td>Smithsonian Astrophysical Observatory</td>
<td>May 23, 2011</td>
<td></td>
</tr>
</tbody>
</table>
In NeuroInformatics, sharing data for long term access has similar challenges:

✓ Incentives to share
✓ Unit of data citation
✓ Sufficient metadata
✓ Obsolescence of formats and software
✓ Discoverability
✓ Increasing size of data
✓ ...

...
Recognition and Credit for Author

Your own site or project site

Your Dataverse
Weisiger, Alex, "Replication data for: Logics of War: Explanations for Limited and Unlimited Conflicts", http://hdl.handle.net/1902.1/18738 UNF:5:OJCPMDOPJ96Q09V7fhXJMA== V1 [Version]

Persistent URL (Handle) for permanent reference

Universal Numerical Fingerprint (UNF) for verification

Altman, M., King, G., 2007 “A Proposed Standard for the Scholarly Citation of Quantitative Data”
A centralized software installation and data repository

Individual virtual data archive with its own branding

A study describes and holds the data (self-curated)

Metadata is searchable
Data Versioning and Data Management

Contributors, curators, admins view

New Study → In Review → Released version 1

Draft

Data File 1

Metadata

| title: |
| author: |
| abstract: |
| year: |
| methods: |
...

Data Versioning and Data Management

End user view

Christopher Casillas; Peter K. Enns; Patrick C. Wohlfarth, "How Public Opinion Constrains the Supreme Court", http://hdl.handle.net/1902.1/14568
UNF:5:YBYuXzOp6STpakyTEIoScQ== V1 [Version]

Edit Study + Add New File

Released version 2

Draft

Data File 2

Christopher Casillas; Peter K. Enns; Patrick C. Wohlfarth, "How Public Opinion Constrains the Supreme Court", http://hdl.handle.net/1902.1/14568
UNF:5:dI8qi49P0uIB9pLfXA3RCw== V2 [Version]
Connecting Publications to Data

The Milky Way in Molecular Clouds: A New Complete CO Survey
Dame, T. M.; Hartmann, Dap; Thaddeus, P.
show affiliations

1.2 m CO Survey Archive

1.2 Meter CO Survey Dataverse

REPLICATION DATA: TAUROS (DHT21)
hdl:10094/10013
Version: 5 - Released: Mon Aug 01 12:20:31 EDT 2011

Data Citation

If you use these data, please add the following citation to your scholarly references. Why cite?

Dame, T. M.; Hartmann, Dap; Thaddeus, P., "Replication data for: Taurus (DHT21)", hdl:10094/10013 W5 [Version]

Citation Format

Original Publication

The Centralized Data Repository Provides:

✓ Backups and replication of data in different locations (LOCKSS)
✓ Re-formatting for preservation (e.g. from SPSS, STATA to archival format)
✓ Extraction of Metadata from data sets
✓ Metadata standards (DDI, Dublin Core)
✓ Inter-operability (OAI-PMH, APIs)

It handles good archival practices for you
Open Source, Java EE 6 Architecture

- JSF, Javascript (jquery): user interface
- EJB: middle-tier, business logic
- OAI-PMH client and server: Harvest Metadata
- Lucene: Indexes metadata
- PostgreSQL: Persistence storage of metadata
- Files system: Store data and complementary files
- RServe: Analysis component for quantitative data files
concerns and suggestions from the neuroinformatics community?

The Dataverse Network Project: http://thedata.org

mcrosas@iq.harvard.edu
IQSS, Harvard University