Discovering + Publishing Data with Dataverse

Elizabeth Quigley
User Experience Lead
The Dataverse Project (@dataverseorg)
equigley@iq.harvard.edu
What is Dataverse?

A software framework for publishing, citing and preserving research data (open source on [github](https://github) for others to install)

Developed by the Institute for Quantitative Social Science at Harvard University.

Provides incentives for researchers to share:

- Recognition & credit via data citations
- Control over data & branding
- Fulfill Data Management Plan requirements
**Institutions** can setup/host their own Dataverse repository (UNC ODUM, Fudan Univ, Scholars Portal, DANS, etc) and within them can have dataverses for a variety of users (across all research domains): Researchers, Projects, Journals, etc.
What is a Dataverse or Dataset?

Schematic Diagram of a **Dataverse** in Dataverse 4.0

Container for your **Datasets** and/or **Dataverses***

* Dataverses can now contain other Dataverses (this replaces Collections & Subnetworks)

Schematic Diagram of a **Dataset** in Dataverse 4.0

Container for your data, documentation, and code.
Share, publish, and archive your data. Find and cite data across all research fields.

dataverse.harvard.edu
Harvard Dataverse

Open to all repository instance at Harvard currently has:

1,231 Dataverses
59,115 Datasets
281,651 Files
>1.3 Million Downloads

*number from September 15, 2015
DATAVERSE BEST PRACTICES
Dataverse Best Practices (1)

• Standard Metadata Schemas
  – DDI & OAI DC
  – New in 4.0:
    • DataCite 3.1
    • ISA-Tab (biomedical)
    • VO Resource (astronomy)
    • DC Terms
  – Metadata can be exported in JSON & XML
Dataverse Best Practices (2)

- Metadata is always public once a dataset is published

- By default, datasets receive CC0 Waiver

- Even though default is CC0 and we encourage open/public data, when needed, data files in a dataset can be made restricted, or terms of use can be added
Dataverse Best Practices (3)

• Formal Data Citation
  – Originally based off Altman + King 2007
  – Endorse + comply w/ 2014 Joint Declaration of Data Citation Principles (FORCE11)
    • Lead by Merce Crosas, Director of Data Science @ IQSS
  – Versioning and File Fixity

• Persistent IDs: DOI (DataCite/EZID)
  – Resolve to a dataset landing page, not directly to the data files
Data Citation Example

**Principle 2: Credit and Attribution** (e.g. authors, repositories or other distributors and contributors)

**Principle 4: Unique Identifier** (e.g. DOI, Handle.). **Principle 5, 6 Access, Persistence:** A persistent identifier that provides access and metadata

**Author(s), Year, Dataset Title, Data Repository or Archive, Version, Global Persistent Identifier**

**Principle 7: Specificity and verification** (e.g. the specific version used). Versioning or timeslice information should be supplied with any updated or dynamic dataset.
Dataverse Best Practices (4)

• Preservation format conversion for tabular data (extract column/variable metadata)

• File Fixity:
  – UNF (Altman, 2008) for tabular data
  – MD5 checksums for other files
Dataverse Best Practices (5)

• Data-PASS: (ICPSR, ODUM, NARA, ROPER,...)  
  – Member of Data-PASS
• OAI-PMH: Harvesting metadata (DC, DDI)  
  – From other Dataverse installations  
  – From other OAI-DC compliant repositories
• If necessary: Deaccession a Dataset
DISCOVERING DATA
Searching for Data

• Search uses **Solr**, an open source search platform

• Solr is also used by:

  ![Netflix Logo](netflix.png)  ![JSTOR Logo](jstor.png)  ![Instagram Logo](instagram.png)  ![Smithsonian Institution Logo](smithsonian.png)  ![Internet Archive Logo](archive.png)
Searching for Data
Search term metadata field match displayed for each result
Browsing for Data
Browsing for Data

• All dataverses are able to select facets by going to the General Information option under the Edit button
• Facets available for all metadata domains supported in Dataverse
PUBLISHING WITH DATAVERS
Rigorous Data Publishing Workflows

Upload

Draft Dataset

Published Dataset v1

Major + Minor Versioning

Published Dataset v1.1

Published Dataset v2

Publish Version 1

Authors, Title, Year, DOI, Repository, V1

Publish Version 1.1: small metadata change; citation doesn’t change.

Publish Version 2: File change (automatic); big metadata change (e.g., author, title).

Authors, Title, Year, DOI, Repository, UNF, V2
Publishing a Dataset

Georgia Election Years Data from 1900-2000

Quigley, Elizabeth, 2015, "Georgia Election Years Data from 1900-2000", http://dx.doi.org/10.5072/FK2/J7EHC4, Root Dataverse, DRAFT VERSION [UNF:6:xeV+tBAeRa7+wPM+G4Pbyw==]

If you use these data, please add this citation to your scholarly resources. Learn about Data Citation Standards.

Description
Data on elections in Georgia.

Subject
Social Sciences

Files

3 Files

dataveseam2015attendees.tab
Tabular Data - 27.3 KB - Sep 10, 2015 - 0 Downloads
Original File MDS: 808cb14e92c205e1d5b657da57eb5cb;
14 Variables, 170 Observations - UNF:6:xeV+tBAeRa7+wPM+G4Pbyw==

dataveseam2015attendees.xlsx
Adobe PDF - 38.4 KB - Sep 10, 2015 - 0 Downloads
MDS: e9a0470edaac4fa08bb67da72db05e7232;

Screen Shot 2015-09-09 at 3.27.07 PM.png
PNG Image - 488.0 KB - Sep 10, 2015 - 0 Downloads
MDS: 6b8f6858cd2143dfdd1c84271954aa5;
Publishing Dataverse + Dataset
Publishing a Dataset
DATAVERSSE + JOURNALS
An Integrated & Automated Journal / Data Publishing Workflow

- **Submit**
  - Features for automatic data citation and insertion into article.

- **Review**
  - Workflows + features for reviewing data before article publication.

- **Publish**
  - Long term preservation + persistent access to dataset.

- **Repository**

- **Code**

- **Data**

- **Article**

- **Journal**

- **Prepare new submission**
  - New versions of a dataset induce new research.

- **Automatic integration w/ data repositories** (common repository API).

- **Submit**

- **Review**

- **Publish**

- **Repository**

- **Journal**

- **Submit**

- **Review**

- **Publish**

- **Repository**

Features for automatic data citation and insertion into article.

Workflows + features for reviewing data before article publication.

Long term preservation + persistent access to dataset.

New versions of a dataset induce new research.

Prepare new submission.

Automatic integration w/ data repositories (common repository API).
Current Workflows in Dataverse: To Connect Data to Journals

A. Journals include Dataverse as a Recommended Repository

1. Recommend Dataverse on your site
2. Author deposits data in Dataverse
3. Author adds Data Citation to article

B. Authors Contribute Directly to a Journal’s Dataverse

1. Setup Dataverse for your Journal
2. Invite authors to deposit data in Dataverse
3. Editor(s) review data
4. If approved, add Data Citation to article

C. Automated Integration of Journal + Dataverse (e.g., OJS)

1. Setup Dataverse for your Journal
2. Connect journal with plugin to API
3. Author submits article + data to journal
4. Data also deposited to Dataverse

Slide created by: Eleni Castro
## Metadata for Journals

**American Journal of Political Science**

<table>
<thead>
<tr>
<th>Journal Volume</th>
<th>58 (52)</th>
<th>57 (38)</th>
<th>59 (38)</th>
<th>Forthcoming (30)</th>
<th>56 (2)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Journal Issue</th>
<th>3 (33)</th>
<th>2 (32)</th>
<th>1 (31)</th>
<th>4 (28)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Replication Data for: Segregation and Inequality in Public Goods</th>
</tr>
</thead>
</table>

Many United States cities function without regular problems. They have well-kept roads, sewers that never overflow, and public parks with swing sets and restrooms. Others struggle to maintain balanced budgets, fail to adequately equip or staff their police forces, and offer little...

<table>
<thead>
<tr>
<th>Replication Data for: Experimental Learning and Presidential Management of the U.S. Federal Bureaucracy: Logic and Evidence from Agency Leadership Appointments</th>
</tr>
</thead>
</table>

Presidents become increasingly effective at managing the bureaucracy because of the information and expertise that they acquire from on-the-job experience. In their appointment choices, this theory predicts that presidents become better at reducing information asymmetries incurred...

<table>
<thead>
<tr>
<th>Replication Data for: Are Voters Equal under Proportional Representation?</th>
</tr>
</thead>
</table>

We develop and apply a new conceptual framework and measure for evaluating electoral systems, focusing on (in)equality in parliamentary representation. Our main arena of interest is proportional representation with districts, an electoral system employed by more than half of demo...

<table>
<thead>
<tr>
<th>Replication Data for: Does Compulsory Voting Increase Support for Leftist Policy?</th>
</tr>
</thead>
</table>

Citizens unequally participate in referendums and this may systematically bias policy in favor of those who vote. Some view compulsory voting as an important tool to alleviate this problem while others worry about its detrimental effects on the legitimacy and quality of democratic...

<table>
<thead>
<tr>
<th>Replication Data for: When Governments Regulate Governments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Konisky, David, 2015, &quot;Replication Data for: When Governments Regulate Governments&quot;, <a href="http://dx.doi.org/10.7910/DVN/INXGMT">http://dx.doi.org/10.7910/DVN/INXGMT</a>, Harvard Dataverse, V1 [UNF:6:3QW6t6L.9boc2T5OxDR5Iq==]</td>
</tr>
</tbody>
</table>

This article advances a political theory of regulation that accounts for the behavior of regulators and regulated entities when both are governments.
ORGANIZING AN INSTITUTION’S DATA
Researcher/Faculty Member

Gary King Dataverse

Gary King Dataverse (Harvard University) http://gking.harvard.edu/

Dataverse

Gary King Dataverse

Harvard Dataverse > Gary King Dataverse

1 to 47 Results

Replication data for: Reverse Engineering Chinese Censorship: Randomized Experimentation and Participant Observation
May 22, 2015


Chinese government censorship of social media constitutes the largest coordinated selective suppression of human communication in recorded history. Although existing research on the subject has revealed a great deal, it is based on passive, observational methods, with well known...

Replication data for: Systematic Bias and Nontransparency in US Social Security Administration Forecasts
May 8, 2015

Kashin, Konstantin; King, Gary; Sonaj, Samir, 2015, "Replication data for: Systematic Bias and Nontransparency in US Social Security Administration Forecasts", http://dx.doi.org/10.7910/DVN/29112, Harvard Dataverse, V1

We offer an evaluation of the Social Security Administration demographic and financial forecasts used to assess the long-term solvency of the Social Security Trust Funds. This same forecasting methodology is also used in evaluating policy proposals put forward by Congress to modi...

Replication data for: Explaining Systematic Bias and Nontransparency in US Social Security Administration Forecasts
May 8, 2015

Kashin, Konstantin; King, Gary; Sonaj, Samir, 2015, "Replication data for: Explaining Systematic Bias and Nontransparency in US Social Security Administration Forecasts", http://dx.doi.org/10.7910/DVN/29323, Harvard Dataverse, V1

The accuracy of U.S. Social Security Administration (SSA) demographic and financial forecasts is crucial for the solvency of its Trust Funds, other government programs, industry decision making, and the evidence base of many scholarly articles. Because SSA makes public little rep...

Replication data for: A Unified Approach To Measurement Error And Missing Data: Overview
Mar 23, 2015


Although social scientists devote considerable effort to mitigating measurement error during data collection, they often ignore the issue during data analysis. And although many statistical methods have been proposed for reducing measurement error-induced biases, few have been e...

Replication data for: A Unified Approach To Measurement Error And Missing Data: Details And Extensions.
Mar 23, 2015

Blackwell, Matthew; Honaker, James; King, Gary, 2015, "Replication data for: A Unified Approach To Measurement Error And Missing Data: Details And Extensions", http://dx.doi.org/10.7910/DVN/29961, Harvard Dataverse, V1

Gary King Dataverse
Research Project/Archive

The Henry A. Murray Research Archive is the endowed, permanent repository for quantitative and qualitative research data at the Institute for Quantitative Social Science. Our collection comprises over 125 terabytes of data, audio, and video.

More information about the Murray Archive can be found at our website.

Oral History of the Tenured Women in the Faculty of Arts and Sciences at Harvard University, 1981
Jul 20, 2015

The purpose of this 1981 study was to document the history of women faculty at Harvard University. Of the then 13 women tenured in the Faculty of Arts and Sciences at Harvard in 1981, 11 professors and the husband of one deceased professor participated in interviews. (list of these...)

A Comparison of Cross-Generational Attitudes About Filial Obligations, 1982
Jul 17, 2015

The purpose of this study was to assess the attitudes of students and their parents about their obligations toward each other and toward the elderly. A total of 463 high school and college students and one parent or older relative (mostly mothers) completed a six-page, closed-end...

Success and Failure Incidents From Self-Employed Women, 1979
May 15, 2015

...
Organization or Institution

In 2003, the Bill & Melinda Gates Foundation launched Avahan, the India AIDS Initiative, to reduce the spread of HIV in India. Avahan’s ten-year charter had three distinct parts. The first was to build and operate a scaled HIV prevention program, with saturated coverage for key population most at risk, in the six states which account for the bulk of HIV infections in India. The second was to transfer the program to the Government of India and other implementers in the country; and the third encouraged the replication of best practices by fostering and disseminating learnings from the program.

Avahan provided funding and support to targeted HIV prevention programs in the six Indian states with the highest HIV prevalence, and along the nation’s major trucking routes. Gathering and using data was critical for all of Avahan’s goals to continuously refine the program and its many moving parts, to inform other HIV prevention efforts including the national prevention program and its direction, to measure impact and to capture best practices. Data in this Dataverse represent the full range of data collected and used by the lead implementing partners, and some evaluation, knowledge building and capacity building partners in Avahan. The range of data encompasses routine program monitoring data, survey data used for monitoring and for evaluation, and special studies to better understand the HIV epidemic in the program areas.

This dataverse has been divided into different sub-domains such as ‘Migration Research’, "Media and Advocacy Studies", etc. Each sub-domain contains all the studies under that thematic area, which is further grouped by the institution that has conducted the study and by specific topics/key populations.

Avahan Dataverse
This is the Astronomy data repository at Harvard. It is currently open to all scientific data from astronomical institutions worldwide. Administration and support is provided by the Harvard-Smithsonian Center for Astrophysics (CfA) in collaboration with Harvard Library (HL) and the Institute for Quantitative Social Science (IQSS). Infrastructure is provided by Harvard University Information Technology Services.

The Astronomy Dataverse Network plays an important role in fulfilling your Data Management Plan requirements (e.g., as mandated by NSF), and for providing data re-use and citation opportunities. Find out more about our team by exploring the Seamless Astronomy and Wolbach Library teams at the CfA. We are also connecting the Astronomy Dataverse to the indexing services provided by the SAD/NASA Astrophysical Data Service (ADS).

<table>
<thead>
<tr>
<th>Dataset Title</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>The International Political Economy Data Resource</td>
<td>Graham, Benjamin A.T., 2015, &quot;The International Political Economy Data Resource&quot;, <a href="http://dx.doi.org/10.7910/DVN/28003">http://dx.doi.org/10.7910/DVN/28003</a>, Harvard Dataverse, V5 [UNP:p:wr4g44AUMaaHw30f69Gw=] &lt;br&gt;Quantitative scholars in international relations often draw repeatedly on the same sources of country-year data across a diverse range of projects. The IPE Data Resource seeks to provide a public good to the field by standardizing and merging together 56 of core IPE data sources.</td>
</tr>
<tr>
<td>APEX CMZ SHFI-1 survey</td>
<td>Ginsburg, Adam, 2015, &quot;APEX CMZ SHFI-1 survey&quot;, <a href="http://dx.doi.org/10.7910/DVN/27801">http://dx.doi.org/10.7910/DVN/27801</a>, Harvard Dataverse, V5 &lt;br&gt;APEX SHFI-1 survey of the CMZ</td>
</tr>
<tr>
<td>Replication data for: Deep 3.8 Micron Observations of the Trapezium Cluster</td>
<td>Maenoh, August; Aivio, Jaqo; Lada, Charles; Lada, Elizabeth, 2015, &quot;Replication data for: Deep 3.8 Micron Observations of the Trapezium Cluster&quot;, <a href="http://dx.doi.org/10.7910/DVN/28877">http://dx.doi.org/10.7910/DVN/28877</a>, Harvard Dataverse, V2 &lt;br&gt;This is the data behind the paper, &quot;Deep 3.8 Micron Observations of the Trapezium Cluster,&quot; by Lada et al. (2004). It includes FITS image files and reference comparison files that would prove useful for interpreting the FITS image files. A note on the images: two images are given...</td>
</tr>
<tr>
<td>Replication Data for: YSOVAR: Mid-Infrared Variability in the Star-forming Region Lynds 1688</td>
<td>Guenther, Hans Moritz Gönter Dataverse &lt;br&gt;Guenther, Hans Moritz, 2015, &quot;Replication Data for: YSOVAR: Mid-Infrared Variability in the Star-forming Region Lynds 1688&quot;, <a href="http://dx.doi.org/10.7910/DVN/N96C6A">http://dx.doi.org/10.7910/DVN/N96C6A</a>, Harvard Dataverse, V1 &lt;br&gt;This dataset contains all YSOVAR lightcurves of L1688 as published in YSOVAR: Mid-infrared Variability in the Star-forming Region Lynds 1688. The data...</td>
</tr>
</tbody>
</table>

Harvard-Smithsonian Center for Astrophysics
Courses

Project TIER: Teaching Integrity in Empirical Research Dataverse

Harvard Dataverse > Project TIER: Teaching Integrity in Empirical Research Dataverse

This dataverse supports a protocol for teaching undergraduates to document the statistical analyses they do for empirical research projects in such a way that their results are completely reproducible and verifiable. The protocol is guided by the principle that the documentation prepared to accompany an empirical research project should be sufficient to allow an independent researcher to replicate easily and exactly every step of the data management and analysis that generated the results reported in the study. You will find in this dataverse examples of the protocol as applied in senior thesis and introductory statistics projects. We hope that requiring students to follow this protocol will not only teach them how to document their research appropriately, but also instill in them the belief that it is an important professional responsibility to do so. For more information, visit the Project TIER website.

Dataverse Category
- Teaching Courses (3)
- Organization or Institution (1)

Publication Date
- 2016 (14)
- 2014 (7)
- 2013 (2)

Author Name
- Adams, Gaines (1)
- Auer, Alexis (1)
- Bosssu, Nicole (1)
- Bracker, Mason (1)
- Brennan, Claire (1)

1 to 10 of 26 Results

Colgate Dataverse: Colgate University
Jul 8, 2015

Haverford Dataverse: Haverford College
Jun 17, 2015

Simplicity Versus WAR: Examining Salary Determinations in Major League Baseball’s Arbitration and Free Agent Markets
May 29, 2015

This paper examines salaries given to arbitration eligible players in Major League Baseball from 2008-2013 and compares them to free agent contracts from the same period. Anecdotal evidence suggests that simpler statistics are more successful in Major League Baseball’s final offer.

More...
Thank you!

Any questions?

Contact: equigley@iq.harvard.edu

Learn more: dataverse.org
Demo Dataverse: dataverse-demo.iq.harvard.edu

@dataverseorg

References