Sharing Your 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 Datavverse 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.
Harvard Dataverse

Open to all repository instance at Harvard currently has:

1,298 Dataverses

59,370 Datasets

284,473 Files

> 1.4 Million Downloads

*number from October 25, 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 + PUBLISHING DATA WITH DATAVERSE
Benefits for Researchers Using Dataverse

• Add your own custom branding and increase your datasets’ visibility by embedding them on your website with our widgets.

• Safe and long-term data storage in preservation format.

• Share your data with everyone, or only specific individuals you approve.

• No need to translate data when statistical software formats change.

• Domain specific metadata: making it easy for others to find your data and associated scholarship.

• Allow users to download your data in any format and run many advanced statistical methods online.
Welcome to Harvard University’s Data Sharing Portal

The institutional data available on this site are being shared to encourage the creativity and innovation of the Harvard University community. In this collection you will find both datasets that are open to the world, and some that are only accessible to members of the University. Please check in often to see what new sources of data we add to the collection.

<table>
<thead>
<tr>
<th>Harvard Faculty Finder</th>
<th>Sep 11, 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waido, Jim, 2015, &quot;Harvard Faculty Finder&quot;, <a href="http://dx.doi.org/10.7910/DVN/P7MLN8">http://dx.doi.org/10.7910/DVN/P7MLN8</a>, Harvard Dataverse, V1</td>
<td></td>
</tr>
<tr>
<td>The Harvard Faculty Finder creates an institution-wide view of the breadth and depth of Harvard faculty and scholarship, and it helps students, faculty, administrators, and the general public locate Harvard faculty according to research and teaching expertise. More info...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Library Cloud</th>
<th>Sep 11, 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weinberger, David, 2015, &quot;Library Cloud&quot;, <a href="http://dx.doi.org/10.7910/DVN/WX7KE6">http://dx.doi.org/10.7910/DVN/WX7KE6</a>, Harvard Dataverse, V1</td>
<td></td>
</tr>
<tr>
<td>Catalog metadata++ about Harvard Library’s collection + VIA images + Archival items. Total of about 17M items. Provides an API, which is documented at <a href="https://wiki.harvard.edu/confluence/display/LibraryStaffDoc/Library+Cloud">https://wiki.harvard.edu/confluence/display/LibraryStaffDoc/Library+Cloud</a>. A overview is attached as the data file. Entry point...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Harvard Library Bibliographic Dataset</th>
<th>Sep 11, 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waido, Jim, 2015, &quot;Harvard Library Bibliographic Dataset&quot;, <a href="http://dx.doi.org/10.7910/DVN/LZDQYN">http://dx.doi.org/10.7910/DVN/LZDQYN</a>, Harvard Dataverse, V1</td>
<td></td>
</tr>
<tr>
<td>This dataset contains over 12 million bibliographic records for materials held by the Harvard Library, including books, journals, electronic resources, manuscripts, archival materials, scores, audio, video and other materials. The metadata has been created, acquired and modified...</td>
<td></td>
</tr>
</tbody>
</table>
Rigorous Data Publishing Workflows

Upload → Draft Dataset → Published Dataset v1 → 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
LIVE DEMO TIME!

demo.dataverse.org
ORGANIZING AN INSTITUTION’S DATA
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.
Organization or Institution

Avahan Dataverse (Bill & Melinda Gates Foundation)

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
Department

Harvard-Smithsonian Center for Astrophysics
Courses
Thank you!

Any questions?

Contact: equigley@iq.harvard.edu

Learn more: dataverse.org
Try out Dataverse: demo.dataverse.org

@dataverseorg

References