Update on Dataverse

Image credit: David Bygott (CC-BY-NC-SA)

2014 Dryad-Dataverse Community Meeting
Mercè Crosas, Elizabeth Quigley & Eleni Castro
Data Science > IQSS > Harvard University
Introduction to Dataverse

Software framework for publishing, citing and preserving research data (open source on github for others to install)

Provides incentives for researchers to share:
- Recognition & credit via data citations
- Control over data & branding
- Fulfill Data Management Plan requirements

Harvard Dataverse (open to all repository instance at Harvard) currently has:

- 700 Dataverses
- 53,857 Datasets
- 739,326 Files
- > 1 Million Downloads
Who’s Using Dataverse?

Worldwide Dataverse Installations

Types of Dataverses (across all research domains)

Institutions (ODUM, MIT, OCUL,...)

Journals (AJPS, Open Health Data,...)

Projects (IFPRI, PSI, COMPLETE,...)

Researchers (Jonathan McDowell, Eric Dunipace,...)
Journals Working With Dataverse

Option 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

Option B. Authors Contribute Directly to a Journal 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

Option C. Seamless Integration btw 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
OJS-Dataverse Integration

Details/Updates:
- Integrating w/ PKP’s Open Journal Systems (Data Deposit API).
- Pilot with ~ 50 journals + expanding outreach.
- OJS Dataverse plugin now available with latest OJS release.
- Future: Embed Dataverse widget into journal article.

http://projects.iq.harvard.edu/ojs-dvn
Dataverse Milestones

1999-2006: Virtual Data Center (VDC)

2006: Coding of Dataverse begins (initial focus on Social Sciences data)

November 2009: Review of Economics and Statistics (RESTAT) Integration with Dataverse

2011: Expanded to include Astronomy & Astrophysics data

Fall 2012: OJS & Dataverse API Integration

March 2012: Dataverse 3.0 Released

April-May 2013: First Usability Testing of Dataverse

2013: Expanded to include Biomedical data

October 2013: Dataverse 4.0 development begins

October 2013: User centered design process integrated

October 2013-Present Dataverse 4.0 Development
How our users have influenced 4.0:

- Over 50 usability testing sessions
- Users from various disciplines participated
- Task completion rates
- System Usability Scale
- User feature requests and reported issues via support tickets
- Metadata reviews with discipline/domain experts
- Prototype review sessions with partners
Overview of Dataverse 4.0

Try our Beta site: http://dataverse-demo.iq.harvard.edu/
Rigorous Data Publishing Workflows

Upload → Draft Dataset

Note: A Published Dataset cannot be deleted (only deaccessioned, if legally needed).

Published Dataset v1

Publish Version 1
Authors, Title, Year, DOI, Repository, V1

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

Published Dataset v1.1

Publish Version 2
Authors, Title, Year, DOI, Repository, UNF, V2

Publish Version 2: File change (automatic); big metadata change; or citation changes.

Published Dataset v2

Expanding Metadata Support

<table>
<thead>
<tr>
<th>Metadata Schema</th>
<th>Version 3.6</th>
<th>Version 4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDI (General &amp; Social Science)*</td>
<td>X (v2.1)</td>
<td>X (v.2.5)</td>
</tr>
<tr>
<td>Simple Dublin Core</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Dublin Core Terms</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DataCite 3.0</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Virtual Observatory (Astrophysics)**</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>ISA-Tab (Biomedical)***</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

* Including variable level metadata found in **tabular data files**.

** Automatically extracts relevant metadata from the header **FITS files**.

*** Controlled vocabulary maps to ontologies/taxonomies (OBI, NCBI,...).
<table>
<thead>
<tr>
<th>Design Type</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Case Control</td>
<td></td>
</tr>
<tr>
<td>Cross Sectional</td>
<td></td>
</tr>
<tr>
<td>Not Specified</td>
<td></td>
</tr>
<tr>
<td>Parallel Group Design</td>
<td></td>
</tr>
<tr>
<td>Perturbation Design</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factor Type</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>Biomarkers</td>
<td></td>
</tr>
<tr>
<td>Developmental Stage</td>
<td></td>
</tr>
<tr>
<td>Cell Surface Markers</td>
<td></td>
</tr>
<tr>
<td>Cell Type/Cell Line</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measurement Type</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA Methylation Profiling (Bisulfite-Seq)</td>
<td></td>
</tr>
<tr>
<td>DNA Methylation Profiling (MeDIP-Seq)</td>
<td></td>
</tr>
<tr>
<td>Histone Modification (ChIP-Seq)</td>
<td></td>
</tr>
<tr>
<td>Protein-RNA Binding (RIP-Seq)</td>
<td></td>
</tr>
<tr>
<td>Transcription Factor Binding (ChIP-Seq)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Organism</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Danio rerio</td>
<td></td>
</tr>
<tr>
<td>Homo sapiens</td>
<td></td>
</tr>
<tr>
<td>Mus musculus</td>
<td></td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cell Type</th>
<th></th>
</tr>
</thead>
</table>
Enhanced Faceted Search
Enhanced Faceted Search

The Harvard Dataverse for Dataverse 4.0 Beta. Beta is only a testing environment so any data stored on Beta is temporary and will eventually be removed. Only datasets that have no restrictions and are non-identifiable data can be uploaded to Beta.

Search for:
election

1 to 4 of 4 results

**Election Data Dataverse (IQSS)**
May 16, 2014
Contains various election datasets.
Preview Recently Released Datasets [+]

**Elections to the United States House of Representatives, 1899-1992**
May 8, 2014 Gary King Dataverse
... and the effect of party incumbancy on election outcomes, contains election returns for elections to the United ...

**Election Data from 1850-1923 in Georgia**
May 16, 2014 Election Data Dataverse
Smith, Jon. 2014, "Election Data from 1850-1923 in Georgia", http://dx.doi.org/10.5072/FK2/139. Harvard Dataverse, V1
... Data spanning from 1850-1912 for the state of Georgia. Includes state and local election data. ...
Keyword: election

**Record of American Democracy, All Key Data Files**
May 8, 2014 Gary King Dataverse
Enhanced Faceted Search

The Harvard Dataverse for Dataverse 4.0 Beta. Beta is only a testing environment so any data stored on Beta is temporary and will eventually be removed. Only datasets that have no restrictions and are non-identifiable data can be uploaded to Beta.

Keyword: election

Record of American Democracy, All Key Data Files
May 8, 2014  Gary King Dataverse

King, Gary; Palmquist, Bradley; Adams, Greg; Altman, Micah; Benoit, Kenneth; Gay, Claudine; Lewis, Jeffrey B.; Mayer, Russ; Reinhardt, Eric, 2014, "Record of American Democracy, All Key Data Files", http://dx.doi.org/10.5072/FK2/71, Harvard Dataverse, V1

... The Record of American Democracy (ROAD) data provide election returns, socioeconomic summaries...
Expanded Advanced Search

Ability to search on specific dataset metadata fields across various domains
• Integrated with Dataverse & Zelig
• For users at all statistical levels
• Explore data, view descriptive statistics, and estimate statistical models for files in datasets
- Integrated with Dataverse & Zelig
- For users at all statistical levels
- Explore data, view descriptive statistics, and estimate statistical models for files in datasets
• Integrated with Dataverse & Zelig
• For users at all statistical levels
• Explore data, view descriptive statistics, and estimate statistical models for files in datasets
What to Expect After 4.0
WorldMap Integration

1. Upload a file containing geographic data into Dataverse
2. Easily visualize the data on the WorldMap system
3. WorldMap layer embedded into dataset in Dataverse
WorldMap Integration

1. Upload a file containing geographic data into Dataverse
2. Easily visualize the data on the WorldMap system
3. WorldMap layer embedded into dataset in Dataverse
# Data Tags
Sharing data with confidence

Start Tagging

<table>
<thead>
<tr>
<th>Harm Levels, and Their Appropriate Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level</strong></td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>NoRisk</td>
</tr>
<tr>
<td>Minimal</td>
</tr>
<tr>
<td>Shame</td>
</tr>
<tr>
<td>CivilPenalties</td>
</tr>
<tr>
<td>CriminalPenalties</td>
</tr>
<tr>
<td>MaxControl</td>
</tr>
</tbody>
</table>

Final tags may not match the tags of a specific harm level. Hover over the terms to view an explanation.
Data Tags

Sharing data with confidence

Person-specific

Does your data include personal information?

☑ YES  ☒ NO

Data Tags

<table>
<thead>
<tr>
<th>Method</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUA Agreement Method</td>
<td>n/a</td>
</tr>
<tr>
<td>Authentication Type</td>
<td>n/a</td>
</tr>
<tr>
<td>Transit Encryption Type</td>
<td>n/a</td>
</tr>
<tr>
<td>Storage Encryption Type</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Full Interview

Tagging Complete!

Direct Data Access

<table>
<thead>
<tr>
<th>Method</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criminal Penalties</td>
<td></td>
</tr>
<tr>
<td>DUA Agreement Method</td>
<td>Sign</td>
</tr>
<tr>
<td>Authentication Type</td>
<td>TwoFactor</td>
</tr>
<tr>
<td>Transit Encryption Type</td>
<td>Encrypted</td>
</tr>
<tr>
<td>Storage Encryption Type</td>
<td>Encrypted</td>
</tr>
</tbody>
</table>
Collaborations & More

Biomedical Metadata + Tools

• HSPH: TB Genomics & Molecular Epidemiology Data
• Harvard Stem Cell Institute
• FAIRport Data

Astronomy Metadata + Tools

• Seamless Astronomy Group
• Harvard-Smithsonian Center for Astrophysics

Also working on:
• Integration w/: ORCID, Open Science Framework, DataUp, DataBridge, ...
• Support for large-scale datasets with efficient data storage.
Data Science Team

Researchers

Data Science Applications and Tools

Information Scientists

Software Engineers

Data Curation & Stewardship

Tool Building & Computer Science

Statistical Innovation

Thank you!

Contact: ecastro@fas.harvard.edu; equigley@iq.harvard.edu

http://datascience.iq.harvard.edu