Dataverse 4.0 & Beyond

Eleni Castro > Institute for Quantitative Social Science (IQSS), Harvard University
Data Science Team

Researchers
Data Science Applications and Tools
Data Curation & Stewardship
Information Scientists
Software Engineers

Statistical Innovation
Tool Building & Computer Science

Find out more: http://datascience.iq.harvard.edu
What is 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 is using Dataverse?

**Worldwide Dataverse Installations**

*Institutions* can setup/host their own Dataverse installation (OCUL, UoA, etc) and within them can have dataverses for a variety of users (across all research domains): Researchers, Projects, Journals (OJS – Dataverse integration), etc.
Streamlined Workflows

Based on extensive continuous usability testing: improved account creation process, dataverse setup (incl. customizations), and dataset (prev. study) creation.
Featured Dataverses

* Beta is only a testing environment so any data stored on Beta is temporary and will eventually be removed. To upload real data and receive a formal data citation, please use the data.harvard.edu. Only datasets that have no restrictions and are non-identifiable data can be uploaded to Beta.
Improved File Upload & Handling

Select multiple files, Drag-n-Drop, Dropbox, File Previews, and extra handling for csv, tsv and excel files (no control card needed).
Rigorous Data Publishing Workflows

**Note:** A Published Dataset **cannot** be deleted (only deaccessioned, with reason included (i.e., legal)).

- **Publish Version 1:**
  - Authors, Title, Year, DOI, Repository, V1

- **Publish Version 1.1:** small metadata change (not citation); files not changed.

- **Publish Version 2:** File change (automatic); big metadata change (citation metadata).

Authors, Title, Year, DOI, Repository, UNF, 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 (Astronomy)**</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,...).
Astronomy and Astrophysics Metadata

Type
- Image
- Mosaic
- EventList
- Spectrum
- Cube

Facility

Instrument

Spatial Resolution

Spectral Resolution

Time Resolution

Bandpass

Central Wavelength (m)

Wavelength Range
- Minimum (m)
- Maximum (m)

Dataset Date Range
- Start: YYYY-MM-DD
- End: YYYY-MM-DD

Astronomy Metadata:
Certain values (e.g., Type, Facility, Instrument, etc) automatically extracted from FITS file header.
Enhanced Faceted Search

Harvard Dataverse

Search this Dataverse...

Advanced Search

Affiliation: Harvard University
Author Name: King, Gary
Author Name: Tomz, Michael

1 to 1 of 1 result

Replication Data for: Making the Most of Statistical Analyses: Improving Interpretation and Presentation
Aug 13, 2014  Gary King Dataverse

King, Gary; Tomz, Michael; Wittenberg, Jason, 2014, "Replication Data for: Making the Most of Statistical Analyses: Improving Interpretation and Presentation", http://dx.doi.org/10.5072/FK2/15, Harvard Dataverse, V1

Social Scientists rarely take full advantage of the information available in their statistical results. As a consequence, they miss opportunities to present quantities that are of greatest substantive...
Expanded Advanced Search

Ability to search on specific dataverses, dataset metadata fields across various domains, and files (variables).
Visualize & Analyze Data: TwoRavens

- Integrated with Dataverse & Zelig (statistical software)
- From beginners up to advanced stats users
- Explore data, view descriptive statistics, and estimate statistical models for files in datasets
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

Read more on: Data Science Blog.
After 4.0

- Sharing Privacy Sensitive Data
- Secure Dataverse
- **DataTags** (questionnaires based on privacy laws)
- ORCID Integration (API)

**Longer-Term**
- Large-scale datasets (efficient storage)
- Ensuring long-term preservation for more file formats (e.g., Archivematica)
Get Involved: Dataverse Community

- Let us know your thoughts on Dataverse 4.0 Beta in the Dataverse Google Group.
- Sign up to participate in usability testing of Dataverse 4.0 Beta by filling out this form.
- Contribute to our code or scripts: GitHub Pull Requests.
- Read our Data Science Blog for any upcoming updates and notifications.
Thank You!

Eleni Castro, Research Coordinator
IQSS, Harvard University
ecastro@fas.harvard.edu

Dataverse 4.0 Demo:
http://dataverse-demo.iq.harvard.edu/
Dataverse Twitter: @thedataorg

Presentation Finished

Any Questions...