Introducing the Expanding Dataverse

Elizabeth Quigley
Usability Specialist
IQSS @ Harvard University
Introduction to Dataverse

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
- Default CC0 Waiver for all uploaded datasets
1999-2006: Virtual Data Center (VDC)

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

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-April 2015 Dataverse 4.0 Development

April 2015: Dataverse 4.0 Launched
dataverse.harvard.edu
Who uses Dataverse?

- Researchers
- Librarians
- Data Archivists
- Journals
- Courses
- Institutions and Organizations
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

- Dataverse
- Dataset #1
- Dataset #2

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

- Dataset
- Descriptive Metadata
- Data files
- Documentation
- Code

Container for your data, documentation, and code.
Harvard Dataverse

- Dataverse repository run at Harvard University

dataverse.harvard.edu
Harvard Dataverse

Open to all repository instance at Harvard currently has:

1,006 Dataverses
57,892 Datasets
269,477 Files
> 1 Million Downloads

*number from April 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 CCO 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
PUBLISHING WITH DATAVEVERSE
Rigorous Data Publishing Workflows

Upload Dataset

Draft Dataset

Published Dataset v1

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

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

Published Dataset v1.1

Published Version 2: File change (automatic); big metadata change (e.g., author, title).
Authors, Title, Year, DOI, Repository, UNF, V2

Major + Minor Versioning

Published Dataset v2
Publishing a Dataset

TCDL Demo Dataverse (Harvard University) Unpublished

Demo Dataverse > TCDL Demo Dataverse > TCDL Demo Dataset

Success! - The files for this dataset have been updated.

TCDL Demo Dataset Unpublished

Quigley, Elizabeth, 2015, "TCDL Demo Dataset", http://dx.doi.org/10.5072/FKz/C79KKG, Demo Dataverse, DRAFT VERSION

If you use these data, please add this citation to your scholarly resources. Why Cite?

Description
Dataset created for demo purposes for Texas Conference on Digital Libraries.

Subject
Arts and Humanities

Files Metadata Terms Versions

AK_2002.tab
Tekvad Data - 35.4 KB - April 21, 2015 - 0 Downloads
Original File MDS: 4b2c6e3c7b0e6007ca3a092e18;
18 variables, 517 observations - UNF:K2WzBi8pAY2TgB1w

HowToDeposit4-01.png
PNG image - 60.1 KB - April 21, 2015 - 0 Downloads
MDS: 6b10415ad03c4296c2580e32b1;

IQSSLogo-2014-Large (2).png
PNG image - 168.9 KB - April 21, 2015 - 0 Downloads
MDS: 3dd67461aa101e1e000aaff0b1d108;

I1_aveg.fits
FITS - 7.7 MB - April 21, 2015 - 0 Downloads
MDS: 817a473f0c9a0d9c2456b552bdc6464A;
This is a FITS file with 1 primary HDU. The following recognized metadata keys have been found in the FITS file:
INSTRUMEN, NAXIS0, NAXIS1, TELESCOP, DATE-OBS, CRVAL1, NAXIS, OBJECT, CD1_1, CD1_2, CD2_1, CD2_2, EXP-Time.
Publishing Dataverse + Dataset

Success! The files for this dataset have been updated.

Publish Dataset

⚠️ This dataset cannot be published until TCDL Demo Dataverse is published. Would you like to publish both right now?

☑️ Once you publish this dataset it must remain published.

Yes, Publish Both  Cancel
Publishing a Dataset
DISCOVERABILITY OF DATA
Searching for Data

• Search uses Solr, an open source search platform

• Solr is also used by:

[Logos of Netflix, JSTOR, Instagram, Smithsonian Institution, and Internet Archive]
Searching for Data

Dataverse

Dataverse Category
Organization or Institution (29)
Researcher (25)
Journal (14)
Research Project (11)
Teaching Course (1)

Affiliation
Unknown (1,235)
Statistics Canada (1,172)
Statistique Canada (1,030)
United States Department of Commerce, Bureau of the Census (852)
United States Department of Justice, Office of Justice Programs, Bureau of Justice Statistics (576)

Publication Date
2015 (4,207)
2009 (4,029)
2014 (2,695)
2007 (1,303)

Socioeconomic Status Indicators of HarvardX and MITx Participants 2012-2014
Apr 22, 2015 - MITx and HarvardX Dataverse
Hanson, John; Reich, Justin, 2015, "Socioeconomic Status Indicators of HarvardX and MITx Participants 2012-2014", http://dx.doi.org/10.7910/DVN/29779, Harvard Dataverse, V3 [UNF:6.x550gICP2OF7g79bWmWvKQ==]

Underinvestment in a Profitable Technology: the Case of Seasonal Migration in Bangladesh
Apr 22, 2015 - Randomized Controlled Trials in the Social Sciences Dataverse
Bryan, Ghrard; Chowdhury, Shyamal; Mobarak, Ahmed Mushfiq, 2014, "Underinvestment in a Profitable Technology: the Case of Seasonal Migration in Bangladesh", http://dx.doi.org/10.7910/DVN/28277, Harvard Dataverse, V2

This paper presents the results of a financial intervention in northwestern Bangladesh intended to promote out-migration to nearby urban areas during the lean season before harvest in order to mitigate

Replication Data for: Careful Commitments: Democratic States and Alliance Design
Apr 22, 2015 - The Journal of Politics Dataverse

Evidence suggests that leaders of democratic states experience high costs from violating past commitments. We argue that because democratic leaders foresee the costs of violation, they are careful to...

Characterization and Detection of epsilon-Berge-Zhukovskii Equilibria
Apr 22, 2015
Gaskó, Noémí, 2015, "Characterization and Detection of epsilon-Berge-Zhukovskii Equilibria", http://dx.doi.org/10.7910/DVN/USiP5E,
## Searching for Data

### Dataverse

- **Datasets** (451)
- **Files** (184)

### Dataverse Category

- Researcher (1)

### Affiliation

- Harvard University (60)
- Unknown (10)
- Puget Sound Transportation Panel (8)
- UC San Diego (8)
- CBS News (6)

### Publication Date

- 2007 (94)
- 2008 (70)
- 2015 (37)
- 2010 (19)
- 2014 (17)

### Author Name

- King, Gary (103)
- King County (Wash.) (103)
- Adams, Greg (55)

<table>
<thead>
<tr>
<th>Dataset Title</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gary King Dataverse</strong> (Harvard University)</td>
<td>Jan 12, 2007</td>
</tr>
<tr>
<td><strong>King County, Washington : trails</strong></td>
<td>Dec 12, 2011 - Harvard Geospatial Library Dataverse</td>
</tr>
<tr>
<td>King County (Wash.), 2011, &quot;King County, Washington : trails&quot;</td>
<td>Location of trails in King County, Washington. City trails and private trails may not be complete ...</td>
</tr>
<tr>
<td><strong>King County, Washington : streams</strong></td>
<td>Dec 12, 2011 - Harvard Geospatial Library Dataverse</td>
</tr>
<tr>
<td>King County (Wash); King County (Wash). Water and Land Resources Division; Washington (State) Dept of Ecology, 2011, &quot;King County, Washington : streams&quot;</td>
<td>... King County, Washington streams. Covers also portions of Snohomish and Pierce Counties. ...</td>
</tr>
<tr>
<td><strong>King County, Washington : landmarks</strong></td>
<td>Dec 12, 2011 - King County Washington Dataverse</td>
</tr>
<tr>
<td>King County (Wash)</td>
<td>Producer Name: King Co. Water and Land Resources</td>
</tr>
<tr>
<td>Analysis Procedure: King County, Washington data was obtained via a CD-ROM. The data were in the format of ArcINFO.</td>
<td>Author Name: King County (Wash)</td>
</tr>
<tr>
<td>This dataset is harvested from our partners at Harvard Geospatial Library. Clicking the link will take you directly to the archival source of the data.</td>
<td></td>
</tr>
</tbody>
</table>
Browsing for Data

Dataverse

Search this dataverse

1 to 10 of 58,813 results

Socioeconomic Status Indicators of HarvardX and MITx Participants 2012-2014
Hansen, John; Reich, Justin, 2015, "Socioeconomic Status Indicators of HarvardX and MITx Participants 2012-2014", http://dx.doi.org/10.7910/DVN/29779, Harvard Dataverse, V3 [UNF:6.x50gCF20F7g79b9WwM+hQ==]
[[NOTE. Placeholder for DOI. Data still in preparation. Data are currently only accessible to qualified reviewers.]] This dataset includes the home mailing addresses of all participants (registrants w...]

Underinvestment in a Profitable Technology, the Case of Seasonal Migration in Bangladesh
Bryan, Gharad; Chowdhury, Shyamal; Mobarak, Ahmed Mushfiq, 2014, "Underinvestment in a Profitable Technology: the Case of Seasonal Migration in Bangladesh", http://dx.doi.org/10.7910/DVN/28277, Harvard Dataverse, V2
This paper presents the results of a financial intervention in northwestern Bangladesh intended to promote out-migration to nearby urban areas during the lean season before harvest in order to mitig...]

Replication Data for: Careful Commitments: Democratic States and Alliance Design
Chiba, Daina; Johnson, Jesse C.; Leeds, Brett Ashley, 2015, "Replication Data for: Careful Commitments: Democratic States and Alliance Design", http://dx.doi.org/10.7910/DVN/DGHX1E, Harvard Dataverse, V1 [UNF:6.8cwy7ANI/KsXGydFl/aQ==]
Evidence suggests that leaders of democratic states experience high costs from violating past commitments. We argue that because democratic leaders foresee the costs of violation, they are careful to...

Characterization and Detection of epsilon-Berge-Zhukovskii Equilibria
Gaskó, Noemi, 2015, "Characterization and Detection of epsilon-Berge-Zhukovskii Equilibria", http://dx.doi.org/10.7910/DVN/USIP5E,
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

![Browse/Search Facets](image-url)
Thank you!

Any questions?

Contact: equigley@iq.harvard.edu

Learn more: dataverse.org

@dataverseorg

References