Update on Data Publishing With Dataverse

Eleni Castro, Research Coordinator
Institute for Quantitative Social Science (IQSS)
Harvard University

DataCite Annual 2014
Nancy, France
August 25, 2014
Introduction to Dataverse

Software framework for publishing, citing and preserving research data (open source on [github](https://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:

- **761** Dataverses
- **54,828** Datasets
- **748,554** Files
- **> 1 Million Downloads**

[Harvard Dataverse](https://dataverse.harvard.edu)
Who’s Using Dataverse?

Worldwide Dataverse Installations

Institutions can setup & host their own Dataverse installation (e.g., Odum, OCUL, DANS, Fudan, etc) and within them can support datasets from a variety of users (across all research domains): Researchers, Projects, Departments, Journals, etc.
Journals Publishing Data w/ 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
Details/Updates: 2 Year Project 2012-2014

- Integrating w/ PKP’s Open Journal Systems (Data Deposit API).
- Pilot with ~ 50 journals + expanding outreach (100s).
- OJS’ Dataverse plugin now available with latest OJS release.
- Future: Embed Dataverse widgets into journal article.

http://projects.iq.harvard.edu/ojs-dvn
OJS Plugin: Journal Data Policies Boilerplate Templates

Dataverse Plugin

Data Policies
Configure data policies.

Data Availability Policy
General data availability policy for the journal. This will appear in About the Journal.

Super Plugin Testing Happy Hour requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive. Super Plugin Testing Happy Hour recommends the Harvard Dataverse, which is free and open to all researchers worldwide to share, cite, reuse and archive research data. Data are important products of the scientific enterprise, and they should be preserved and usable for decades in the future. Authors may elect to have the data publicly available at time of publication. Exceptions may be granted at the discretion of the editor, especially for sensitive information such as human subject data or the location of endangered species. Any exceptions should be documented in a statement in the public article. (Adapted from Joint Data Archiving Policy (JDAP)).

Option to: (A) deposit into Dataverse AND/OR; (B) if data is already in a repository can include the data citation (w/ persistent URL/identifier).
OJS Plugin: Editor Reviews Article + Data
Data Published in Dataverse w/ OJS Plugin

2 Options in OJS:
1) Dataset Published (with DOI) at Article Approval.
2) Dataset Published when Journal Issue is Released.
Data citation

Towards An Integrated Publishing Lifecycle

Submit

Integration with data repositories (common repository API)

Review

Guidelines for reviewing data before article publication

Prepare new submission

New versions of a dataset induce new research

Reuse & Extend

Data citation to reuse data and validate published results

See: Data Citation Principle #1 Importance

Image Credit: Mercè Crosas
Publishing in 4.0 (Late Fall 2014)
Rigorous Data Publishing Workflows

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

Dataset Versioning (1)
10 Million International Dyadic Events


When the Palestinians launch a mortar attack into Israel, the Israeli army does not wait until the end of the calendar year to react. Yet, most modern data collections are aggregated to the month or year. The data available here include almost 10 million individual events, each coded to the exact day they occur or become known. Each event is summarized in the data as "Actor A does something to Actor B", with Actors A and B recording about 450 countries and other within-country actors and "does something to" coded in an ontology of about 200 types of actions. The data are coded by computer from millions of Reuters news reports.

Keyword: events; palestine
Subject: Social Sciences
Related Publication:

<table>
<thead>
<tr>
<th>Versions</th>
<th>Show Differences</th>
<th>Show Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Additional Citation Metadata: (2 Added, 1 Changed);</td>
<td>admin Privileged August 22, 2014</td>
</tr>
<tr>
<td>2.0</td>
<td>Files (Added: 3) Show Details</td>
<td>admin Privileged August 15, 2014</td>
</tr>
<tr>
<td>1.0</td>
<td>This is the first published version.</td>
<td>admin Privileged August 13, 2014</td>
</tr>
</tbody>
</table>

© Copyright 1997-2014, President & Fellows Harvard University.
Dataset Versioning (3)

Ex. Added files to a Dataset so it bumped up to a major version change.
Ex. Added small metadata change to a Dataset so it bumped up to a minor version change.
Deaccession Data in 4.0

Before a Dataset is published the DOI is private (reserved). Only when published is it made public & searchable.

In accordance w/ Data Citation Principle #6 Persistence: A Published Dataset cannot be deleted; only deaccessioned, with a reason.

You can Deaccession (in 4.0):
1. a **version(s)** of a Dataset, or
2. an **entire** Dataset.
Deaccession Workflow (Step 1)

Ex. This file was added in v2 and has identifiable information.
Deaccession Workflow (Step 2)
Deaccession Workflow (Step 3)
Deaccession Workflow (Step 4)

Data Citation Principle #6
Persistence

Sample Dataset For Demo Purposes Only


Deaccession Reason: There is identifiable data in one or more files. Gary King can be identified by the photo, and file name, added in this version.

<table>
<thead>
<tr>
<th>Version</th>
<th>Deaccession Reason</th>
<th>Admin Privileged</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>There is identifiable data in one or more files. Gary King can be identified by the photo, and file name, added in this version.</td>
<td>admin Privileged</td>
<td>July 17, 2014</td>
</tr>
<tr>
<td>1.0</td>
<td>This is the first published version.</td>
<td>admin Privileged</td>
<td>July 17, 2014</td>
</tr>
</tbody>
</table>
Data Publishing After 4.0 (2015)

Publishing Privacy Sensitive Data
- Secure Dataverse
- DataTags (demo) (based on Privacy Laws and DUAs)

Integration with ORCID (API): create ORCID account, connect all Dataverse datasets to ORCID account. (Note: 4.0 will already allow for authors to enter ID.)
Thank you!
Contact: ecastro@fas.harvard.edu
More information: http://datascience.iq.harvard.edu
Twitter: @thedataorg