Data Publishing with Dataverse

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

NDSA NE Regional Meeting,
October 30, 2014 > Amherst, MA

https://flic.kr/p/7QeoWp
Data Science Team

Data Curation & Stewardship

Information Scientists

Researchers

Statistical Innovation

Data Science Applications and Tools

Tool Building & Computer Science

Software Engineers

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

Software framework for Research Data (since 2006):
• Publishing
• Citing
• Analyzing
• Preserving

Provides incentives for researchers to share:
• Recognition & credit via data citations
• Control over data & branding
• Fulfill journal data availability and funder requirements.
Institutions can setup/host their own Dataverse installation (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.
Harvard Dataverse

Open to all researchers; general community repository instance:

810 Dataverses

55,069 Datasets

>1.1 Million Downloads

750,903 Files

http://thedata.harvard.edu
# National Digital Stewardship Alliance Dataverse

Data archive for data collected by the national digital stewardship alliance.

## National Digital Stewardship Alliance

<table>
<thead>
<tr>
<th>Sort By:</th>
<th>Global ID</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Study</th>
<th>Title</th>
<th>Description</th>
<th>DOI/URL</th>
<th>Downloads</th>
<th>Last Released</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>NDSA Web Archiving Surveys</strong></td>
<td>The NDSA conducted two surveys of organizations in the United States that are actively involved in, or planning to start, programs to archive content from the web. The goal of the survey was to better understand the landscape of web archiv...</td>
<td>doi:10.7910/DVN/27593</td>
<td></td>
<td>Oct 23, 2014</td>
</tr>
<tr>
<td>NDSA Web Archiving Surveys</td>
<td>by Content Working Group</td>
<td></td>
<td>8 downloads</td>
<td></td>
<td>Oct 14, 2014</td>
</tr>
<tr>
<td></td>
<td><strong>Replication data for: Staffing for Effective Digital Preservation: An NDSA Report</strong></td>
<td>Businesses, cultural memory institutions, repositories, and government bodies seeking to preserve digital assets responsibly face significant staffing challenges. How many staff and what types of positions are required? What skills, educat...</td>
<td>doi:10.7910/DVN/27562</td>
<td></td>
<td>Oct 9, 2014</td>
</tr>
<tr>
<td></td>
<td><strong>Data for NDSA Storage Report 2011</strong></td>
<td>The NDSA storage survey aims to gather information on preservation storage systems. The respondents represent a diverse cross section of organizations working with preservation storage systems including university libraries, consortia, ins...</td>
<td>hdl:1902.1/19768</td>
<td></td>
<td>Feb 10, 2013</td>
</tr>
<tr>
<td></td>
<td>by Altman, Micah; Bailey, Jefferson; Cariani, Karen; Gallinger, Michelle; Owens, Trevor</td>
<td></td>
<td>12 downloads</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[NDSA.org](http://thedata.harvard.edu/dvn/dv/ndsa)
Dataverse 4.0 (December 2014)
Dataverse Best Practices (1)

• Standard Metadata Schemas
  – DDI, FGDC & OAI DC
  – Coming in 4.0: DataCite 3.1, ISA-Tab (biomedical), and VO Resource (astronomy) → Export to JSON & XML

• Formal Data Citation (Altman & King, 2007)
  – Endorse + comply w/ Joint Declaration of Data Citation Principles (FORCE11)
  – Versioning and File Fixity

• Persistent IDs: Handles & DOI (DataCite/EZID)
Dataverse Best Practices (2)

• 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 (3)

• LOCKSS (replication of files)
  → Data-PASS: (ICPSR, ODUM, NARA, ROPER,...)

• OAI-PMH: Harvesting metadata (DC, DDI)
  – From other Dataverse installations
  – From other OAI-DC compliant repositories

• If necessary: Deaccession a Dataset
Rigorous Data Publishing Workflows

Upload → Draft Dataset

Published Dataset v1

In 4.0: Major + Minor Versioning

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

Files Metadata Versions

Show Differences
- 2.1 Additional Citation Metadata: (2 Added, 1 Changed); Show Details
- 2.0 Files (Added: 3); Show Details
- 1.0 This is the first published version.

Show Details

admin Privileged  August 22, 2014
admin Privileged  August 15, 2014
admin Privileged  August 13, 2014
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. version(s) of a Dataset, or
2. an entire Dataset.
Deaccession A Version (1)

Ex. This file was added in v2 and has identifiable information.
Deaccession A Version (2)
Deaccession A Version (3)

Deaccession Landing Page

Data Citation Principle #6
Persistence

Sample Dataset For Demo Purposes Only

Deaccessioned


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>Versions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
</tr>
<tr>
<td>Deaccessioned 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.</td>
</tr>
<tr>
<td>1.0</td>
</tr>
<tr>
<td>This is the first published version.</td>
</tr>
</tbody>
</table>

© Copyright 1997-2014, President & Fellows Harvard University.

Powered by The Dataverse Project v. 4.0
Data Publishing After 4.0 (2015)

Publishing Privacy Sensitive Data

- Secure Dataverse
- **DataTags (demo)** (based on Privacy Laws and DUAs)

- Data Citation Provenance Registry (w/ SEAS funded by NSF)
- Author Disambiguation: ORCID Integration (API)
- Long-term preservation for more file formats (beyond tabular)
  - UDFR, Archivematica,...
Thank you!
Contact: ecastro@fas.harvard.edu

References