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
Data Science Team

Researchers

Statistical Innovation

Data Science Applications and Tools

Information Scientists

Software Engineers

Tool Building & Computer Science

Data Curation & Stewardship

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

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

Open source on [github](https://github).

Provides incentives for researchers to share:
- Recognition & credit via **data citations**
- Control over data & branding
- Fulfill journal data availability and funder requirements.
Who Uses Dataverse?

Worldwide Dataverse Installations

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

750,903 Files

>1.1 Million Downloads

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

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

**National Digital Stewardship Alliance**

**Sort By:** Global ID

- **NDSA Web Archiving Surveys**
  by Content Working Group
  **Description:** 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...Continue [+]
  **Distribution Date:** October 14, 2014
  **doi:** 10.7910/DVN/27593
  **Downloads:** 8
  **Last Released:** Oct 23, 2014

- **Replication data for: Staffing for Effective Digital Preservation: An NDSA Report**
  **Description:** 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...Continue [+]
  **Distribution Date:** October 01, 2013
  **doi:** 10.7910/DVN/27562
  **Last Released:** Oct 9, 2014

- **Data for NDSA Storage Report 2011**
  by Altman, Micah; Bailey, Jefferson; Cariani, Karen; Gallinger, Michelle; Owens, Trevor
  **Description:** 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...Continue [+]
  **Distribution Date:** 2012
  **hdl:** 1902.1/19768
  **Downloads:** 12
  **Last Released:** Feb 10, 2013


**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

1. **Upload**
   - Dataset

2. **Draft Dataset**
   - Published Dataset v1
     - Authors, Title, Year, DOI, Repository, V1

3. **Publish Version 1.1**: small metadata change; citation doesn’t change.
   - Published Dataset v1.1

4. **Publish Version 2**: File change (automatic); big metadata change (e.g., author, title).
   - Published Dataset v2
     - Authors, Title, Year, DOI, Repository, UNF, V2

5. **In 4.0: Major + Minor Versioning**
Dataset Versioning (1)
Dataset Versioning (2)
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)

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.

Versions

2.0 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. admin Privileged July 17, 2014

1.0 This is the first published version. admin Privileged July 17, 2014

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