Sharing With Dataverse

Eleni Castro
IQSS Harvard
NDS2 October 24, 2014
Data Science Team

Find out more: http://datascience.iq.harvard.edu
Introduction to 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 journal data availability and funder requirements.

Harvard Dataverse (open to all; general community repository instance):

808 Dataverses
55,058 Datasets
750,850 Files

> 1 Million Downloads
Who Uses Dataverse?

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.
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)

• Formal Data Citation (Altman & King, 2007)
  – Endorsed + comply w/ Joint Declaration of Data Citation Principles (FORCE11)

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

• File Fixity:
  – UNF for tabular data
  – MD5 checksums for other files

• Open Data (+ metadata) Licenses (CC0)

• OAI-PMH: harvesting metadata (DC, ...)

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

SafeArchive  Data-PASS  LOCKSS
Towards An Integrated Publishing Workflow

- **Submit**
  - Guidelines for authors to include data citations in Reference list
  - Integration with data repositories (common repository API)

- **Review**
  - Guidelines for reviewing data before article publication

- **Publish**
  - Data citation to reuse data and validate published results

- **Prepare new submission**
  - New versions of a dataset induce new research

- **Reuse & Extend**
API Integration with Dataverse

Data Deposit API (metadata + data w/ SWORDv2)

For depositing datasets into Dataverse via API
See: OJS-Dataverse Journal Integration Project
http://projects.iq.harvard.edu/ojs-dvn/home

Also: dvn R Package, OSF Dataverse Add-on, etc

Data Sharing API

For searching/downloading Dataverse datasets (metadata + data) via API.
See: Thomas Leeper’s dvn R package
Visualize & Analyze Data: TwoRavens

- With Dataverse 4.0: for Tabular Data (i.e., R, csv, SPSS, ...)
- From beginners up to advanced stats users
- Explore data, view descriptive statistics, and estimate statistical models for files in datasets
1. Upload a shape file containing geographic data into Dataverse
2. WorldMap layer embedded into dataset in Dataverse
3. Easily visualize the data on the WorldMap system.

Read more on: Data Science Blog.
Future Dataverse Collaborations

Sharing Privacy Sensitive Data

- Secure Dataverse
- **DataTags** (questionnaires based on privacy laws)

Longer-Term

- Data Citation Provenance Registry (NSF funder w/ SEAS)
- Large-scale datasets (efficient storage) ⟷ iRods
- Integrate w/ more Publishing Systems (new Sloan grant!)
- Author Disambiguation: ORCID Integration (API)
- Long-term preservation for more file formats (Archivematica)
Thank you!

Contact: ecastro@fas.harvard.edu
More information: http://datascience.iq.harvard.edu/
Twitter: @thedataorg