Data Publishing workflows and APIs with Dataverse

Mercè Crosas, Ph.D. @mercecrosas

Director of Data Science

Institute for Quantitative Social Science, Harvard University

http://datascience.iq.harvard.edu

The continuum from publishers to data repositories: models to support seamless scholarship

SSP, May 29, 2014

Introduction to Dataverse

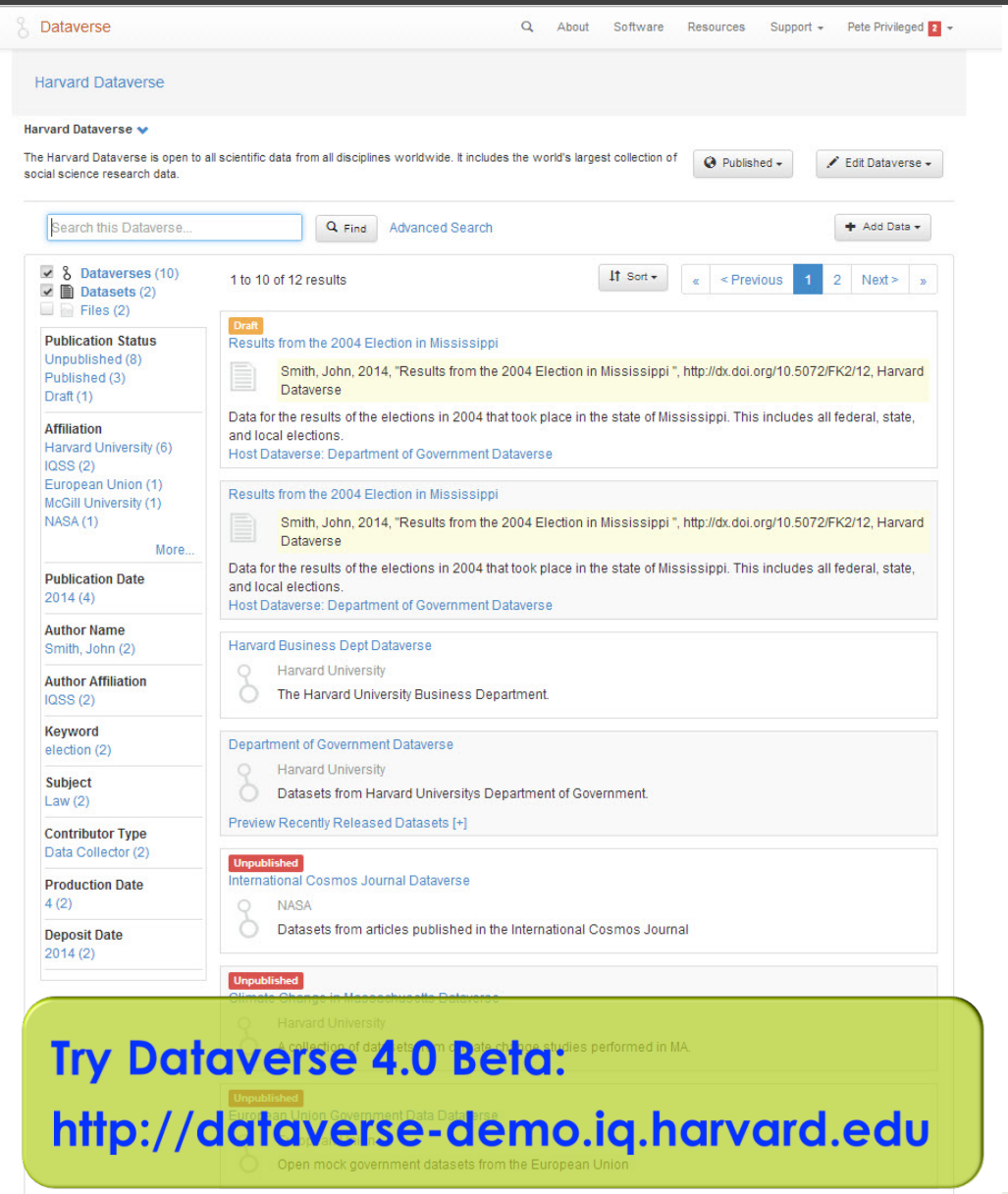
Dataverse Software

- A framework for publishing, citing and preserving research data
- Open-source, available at GitHub
- Started in 2006
- Several installations around the world, supporting all data types across multiple disciplines.

Dataverse Repository

- Harvard hosts a Dataverse instance **free and open** to all researcher data
- It currently holds > 53,000 datasets, with 735,000 files.
- Contains dataverses for researchers, journals, organizations.
- Find or deposit data at: http://thedata.harvard.edu

Dataverse 4.0



Try Dataverse 4.0 Beta:
http://dataverse-demo.iq.harvard.edu

This summer:

- New UI
- New rich, faceted search
- Reformatting and metadata extraction for more data types (excel, CSV, R data, Stata, SPSS, FITS)
- Metadata standards for social sciences, astronomy, biomedical sciences.
- Integration with a new data exploration and analysis tool (TwoRavens)

Citation Metadata

Title *
Replication Data for: Building a Bridge Betwe
Add 'Replication Data for' to Title

Author
Name* Castro, Eleni
Affiliation IQSS

Contact E-mail * ecastro@fas.harvard.edu

Description *
Research dataset for my publication on connecting journal articles and their underlying research data. Includes analysis of current data publication practices.

Keyword data publication

Subject *
- [ ] Mathematical Sciences
- [ ] Physics
- [x] Social Sciences
- [ ] Other

Citation Metadata:
Compliant with DataCite, Dublin Core, DDI study description. Applies to all datasets.

Social Science and Humanities Metadata

Topic Classification
Term | Vocabulary | URL

Software
Name | Version

Series
Name | Information

Time Period Covered
Start (YYYY-MM-DD) | End (YYYY-MM-DD)

Date of Collection
Start (YYYY-MM-DD) | End (YYYY-MM-DD)

Country/Nation

Geographic Coverage

Geographic Unit

Geographic Bounding Box
West Longitude | East Longitude
North Latitude | South Latitude

Social Sciences and Humanities Metadata:
Compliant with DDI

Astronomy and Astrophysics Metadata

Type

- Image
- Mosaic
- EventList
- Spectrum
- Cube

Facility

Instrument

Spatial Resolution

Spectral Resolution

Time Resolution

Bandpass

Central Wavelength (m)

Wavelength Range — **Minimum (m)** — **Maximum (m)**

Dataset Date Range — **Start** (YYYY-MM-DD) — **End** (YYYY-MM-DD)

Astronomy Metadata:
Compliant Virtual Observatory (VO) schema; extract metadata from FITS files

Design Type

- [ ] Case Control
- [ ] Cross Sectional
- [ ] Not Specified
- [ ] Parallel Group Design
- [ ] Perturbation Design

Factor Type

- [ ] Age
- [ ] Biomarkers
- [ ] Developmental Stage
- [ ] Cell Surface Markers
- [ ] Cell Type/Cell Line

Measurement Type

- [ ] DNA Methylation Profiling (Bisulfite-Seq)
- [ ] DNA Methylation Profiling (MeDIP-Seq)
- [ ] Histone Modification (ChIP-Seq)
- [ ] Protein-RNA Binding (RIP-Seq)
- [ ] Transcription Factor Binding

Organism

- [ ] Danio rerio
- [ ] Homo sapiens
- [ ] Mus musculus
- [ ] Rattus norvegicus

Cell Type

+

Bio Metadata:
Compliant with ISA-Tab schema, plus biomedical ontologies

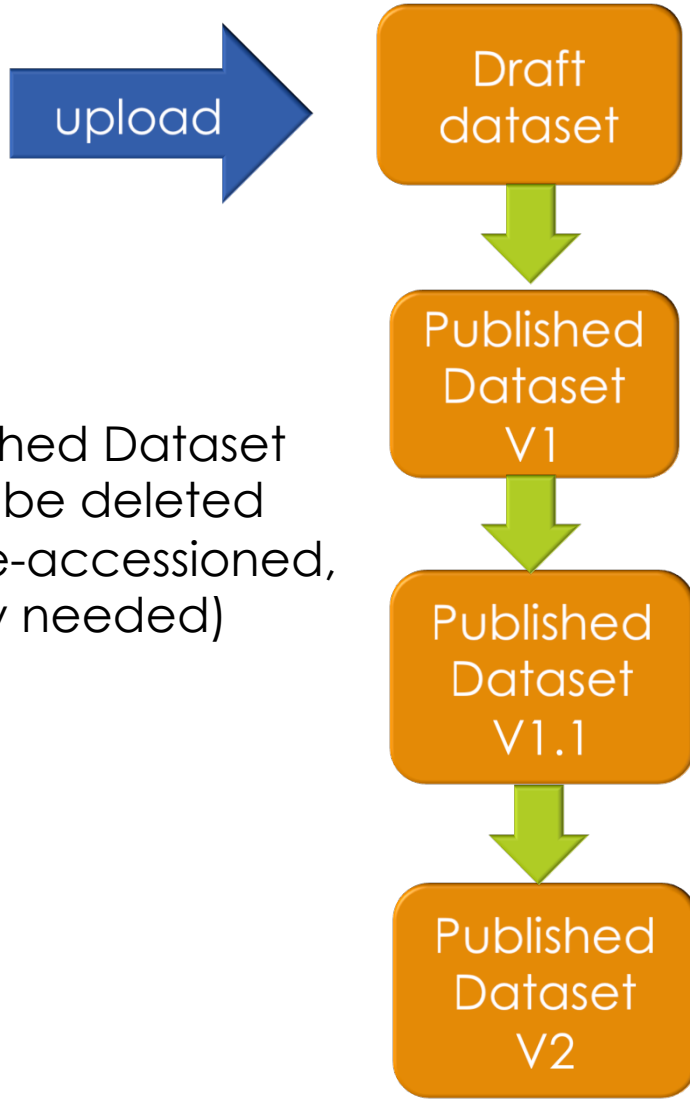
Data Publishing Guidelines

Three pillars to Data Publishing:

- A trusted data repository to guarantee long-term access
- A formal data citation*
- Sufficient information to understand and reuse the data (metadata, documentation, code)

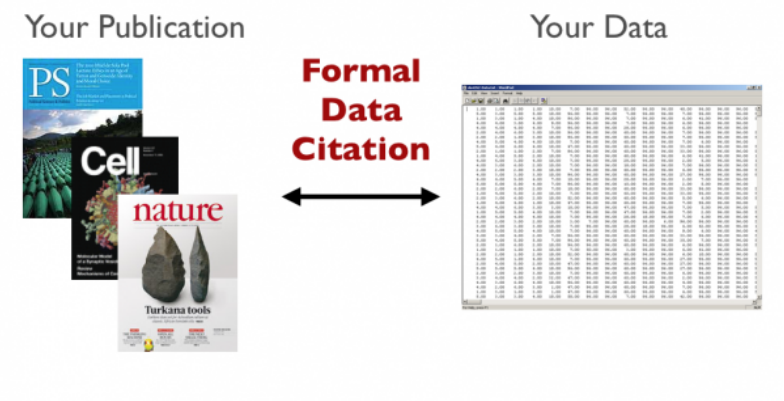
* Data Citation Principles: https://www.force11.org/datacitation

A Rigorous Data Publishing Workflow



A Published Dataset cannot be deleted (only de-accessioned, if legally needed)

Workflows that Integrate with Journals



Option A. Author publishes a dataset to his/her Dataverse, then provides the Data Citation to the journal.

Option B. Author contributes to a journal Dataverse:

1. Add dataset to Journal Dataverse as a draft.
2. Journal Editor reviews it, and approves it for release.
3. Dataset is published with Data Citation and link from journal article to the data.

Option C. Seamless Integration between journal system and Dataverse.

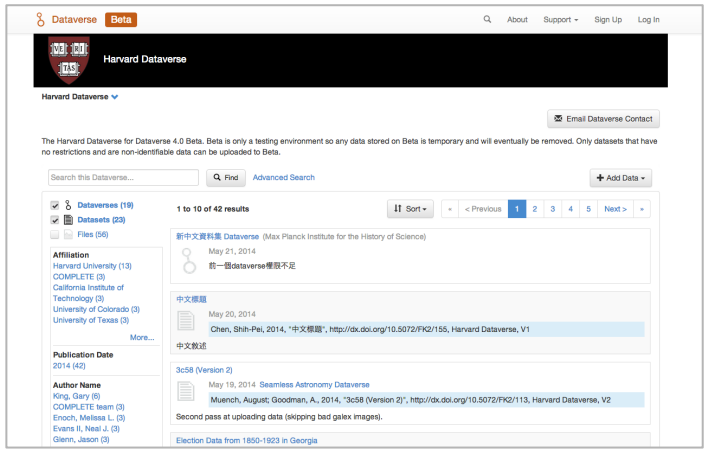
Example of Option C: OJS and Dataverse Integration

OJS Journal

Citation to Data

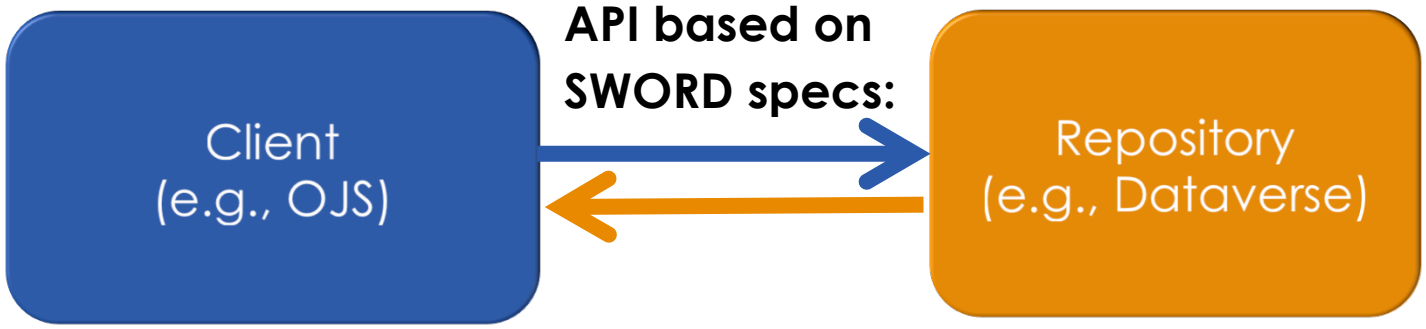
Citation to Article

Journal Dataverse



- Sloan funded project to integrate PKP's Open Journal System (OJS) with the Dataverse software.
- Pilot with ~ 50 journals
- OJS Dataverse plugin now available with latest OJS release
- http://projects.iq.harvard.edu/ojs-dvn

Toward a common API between journal systems and data repositories



- ✓ XML file: AtomPub "entry" with Dublin Core Terms (e.g., title, creator)
- ✓ Zip file: All data files associated with that dataset.

- ✓ XML file: "Deposit Receipt" → send **data citation** from repository to client

Send updates from client to server during lifecycle:
In review, publish first version, new versions

Toward an integrated, living publishing workflow