Data Publishing Workflows with Dataverse

Mercè Crosas, Ph.D.
Twitter: @mercecrosas
Director of Data Science
Institute for Quantitative Social Science, Harvard University

MIT, May 6, 2014
Intro to our Data Science Team and Projects
Data Science at the Institute for Quantitative Social Science
http://datascience.iq.harvard.edu

Data Science
Research Infrastructures for Data-Intensive Science, Analytical Tools and Data Stewardship

From Information to Knowledge
Data Science at IQSS combines expertise in software engineering, statistical innovation and data curation to build software applications and tools for sharing, exploring and analyzing data in today’s data-intensive research environment. In the last decade, our team has developed and supported the Dataverse and Zelig applications, now widely-used by researchers throughout the world for data publishing and statistical analysis respectively. More recently, a new generation of tools - Consilience for text clustering analysis, DataTags for sharing sensitive data and SolaFide for data exploration and automated analysis -, as well as collaborations across disciplines, have broadened the research and development carried out by the team.
Combines Expertise

Researchers

Data Science Applications and Tools

- Statistical Innovation
- Data Curation & Stewardship

Information Scientists

- Tool Building & Computer Science

Software Engineers
With a Team of 20

Mercè Crosas,
Director of Data Science

Gary King,
Director of IQSS

Cris Rothfuss,
Executive Director

Statistics and Analytics
James Honaker
Christine Choirat
Vito d'Orazio

Software Development
Gustavo Durand
Robert Treacy
Ellen Kraffmiller
Michael Bar-Sinai
Leonid Andreev
Phil Durbin
Steve Kraffmiller
Xiangqing Yang
Raman Prasad (BARI)

Data Curation and Archiving
Sonia Barbosa
Eleni Castro
Dwayne Liburd

Usability and UI
Elizabeth Quigley
Michael Heppler
Two widely-Used Frameworks Developed in the last Decade

A framework that allows analysts to use and interpret a large body of R statistical models from heterogeneous contributors through a common interface.

A data publishing framework that allows researchers to share, preserve, cite and analyze data, while keeping control and gaining credit for their data.
New Tools that Integrate with our Initial Work

**SolaFide**

An interactive web interface that allows users at all levels of statistical expertise to explore their data and appropriately construct statistical models.

Integrates with Zelig and Dataverse.

**DataTags**

A framework that allows data contributors to set a level of sensitivity for their dataset based on legal regulations, which defines how the data can be stored and shared.

Integrates with Dataverse.

In collaboration with NSF Privacy Tools project
Expanding in other Areas

Consilience

A web application that assists researchers to discover new clusters to categorize large document sets, leveraging all the clustering methods in the literature.

RBuild

An application that provides a continuous integration build solution for R packages shared in Git to archived published code in CRAN.
Support Throughout the Research Cycle

Develop Quantitative Methods

Analyze Quantitative Datasets

Publish Data

Share Sensitive Data

Explore, reanalyze and reuse data

Cite Data from Published Results

Consilience

Zelig

Rb RBuild

DataTags

SolaFide

Develop > Analyze > Share > Explore > Validate & Reuse
Reproducible and Reusable Science: “encourage open data and methodological transparency, and promote and enable data citation” (with Dataverse, Zelig and SolaFide)

Computationally Assisted Exploration: “with Consilience and SolaFide, assist researchers to understand and discover new insights from their data”

Interdisciplinary Quantitative Scientific Scope: “our tools and research frameworks address broad methodological issues in quantitative science and are often employed in other domains”

When Data are Not Open: “solutions to preserve privacy, while still providing science the fundamental ability to learn, access and replicate findings, with DataTags and PrivateZelig”

Large-Scale Data Sets: “will handle large-scale data sets, as Big Data science reaches all disciplines: Consilience for millions of text documents, and Zelig and Dataverse to handle TB-PB-scale data sets.”
Harvard Dataverse
The Harvard Dataverse Repository

- In collaboration with the Harvard Library, Harvard hosts a Dataverse instance free and open to all researchers.

- It currently holds > 53,000 datasets, with 735,000 files.

- Find or deposit data at: http://thedata.harvard.edu
Collaborations with MIT

- Membership through the Harvard-MIT Data Center (e.g., statistics training, access to ICPSR collection)

- The MIT Libraries Dataverse disseminates data purchased by the MIT Libraries (with Kate McNeill):
  - [http://thedata.harvard.edu/dvn/dv/mit](http://thedata.harvard.edu/dvn/dv/mit)

- MIT faculty and research groups are already disseminating their data through the Harvard Dataverse

- Research collaborations (with Micah Altman):
  - Integration of Publications with Data (Funded by Sloan): [http://projects.iq.harvard.edu/ojs-dvn](http://projects.iq.harvard.edu/ojs-dvn)
Dataverse 4.0

Target release date: June 23

- New UI
- New rich, faceted search
- New data file ingest (excel, CSV, R, Stata, SPSS)
- New metadata for social sciences, astronomy, biomedical sciences.
- Integration with SolaFide.
Data Publishing Workflows
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 Publishing Workflow

A Published Dataset cannot be deleted (only deaccession, if legally needed)

1. **Draft dataset**
2. **Published Dataset v1**
   - Release Version 1
   - Authors, Title, Year, DOI Repository, UNF, V1
3. **Published Dataset V1.1**
   - Push Version 1.1: small metadata change; citation doesn’t change
4. **Published Dataset V2**
   - Push Version 2: big metadata change, or file change; citation changes
   - Authors, Title, Year, DOI Repository, UNF, V2
Workflows that Integrate with Journals

1. Publish a dataset to your Dataverse, then provide the Data Citation to the journal.

2. Contribute 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.

3. Seamless Integration between journal system and Dataverse.
Sloan funded project to integrate PKP’s Open Journal System with the Dataverse software.

Pilot with ~ 50 journals

OJS Dataverse plugin now available with latest OJS release

http://projects.iq.harvard.edu/ojs-dvn
Detailed System Integration

- XML file: AtomPub "entry" with Dublin Core Terms (e.g., title, creator)
- Zip file: All data files associated with that dataset.
- HTTP header "In-Progress: false" to publish datasets.
- Support HTTP verbs: GET, PUT, POST, and DELETE.

- XML file: “Deposit Receipt”
- HTTP status code: 200, 201, 204, 404, 405, 406, 412, 415

Client can query repository (server) any time to get status
Follows SWORD2 specifications

SWORD is known and supported within academic publishing; a “profile” of the AtomPub standard.

The SWORD project provides client libraries for Python, Java, Ruby, and PHP:
- OJS uses the PHP client library
- OSF uses the Python client library
- DataUp and DVN-R built a custom Dataverse client
How it differs from SWORD

- Dataverse does not use SWORD download API:
  - Use own Data API
  - Plan to add this support in the future

- Add XML attribute to pass article citation from client:
  - Allow DCterms:isReferencedby to contain attributes such as HoldingsURI to link back to article from Dataverse
  - This is now part of the SWORD PHP client library

- Use “In-Progress: false” to indicate that dataset is ready to be published (In SWORD spec means deposit complete)
Support for Metadata Standards

- A core or **citation metadata** that applies to all datasets –
  Supported currently by Data Deposit API

- Extensible metadata blocks for specific domains:
  - **Social sciences:**
    - Maps to DDI schema;
    - file metadata extracted from tabular data file
  - **Astronomy:**
    - Maps to VO schema;
    - partially extracted from FITS file
  - **Biomedical sciences:**
    - Maps to ISA-tab schema
    - Controlled vocabularies maps to EFO, OBI, and Ontology of Clinical Research
    - Extended and managed using SKOS (support taxonomies within the framework of the semantic web)
Title: Replication Data for: Building a Bridge Between Connecting Journal Articles and their Underlying Research Data. Includes an analysis of current data publication practices.

Author: Castro, Eleni
Affiliation: IQSS

Contact Email: ecastro@fas.harvard.edu

Description:

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

Keyword: data publication

Subject:
- Mathematical Sciences
- Physics
- Social Sciences
- Other
<table>
<thead>
<tr>
<th><strong>Social Science and Humanities Metadata</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topic Classification</strong></td>
</tr>
<tr>
<td><strong>Term</strong></td>
</tr>
<tr>
<td><strong>Vocabulary</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>Software</strong></td>
</tr>
<tr>
<td><strong>Name</strong></td>
</tr>
<tr>
<td><strong>Version</strong></td>
</tr>
<tr>
<td><strong>Series</strong></td>
</tr>
<tr>
<td><strong>Name</strong></td>
</tr>
<tr>
<td><strong>Information</strong></td>
</tr>
<tr>
<td><strong>Time Period Covered</strong></td>
</tr>
<tr>
<td><strong>Start</strong></td>
</tr>
<tr>
<td><strong>End</strong></td>
</tr>
<tr>
<td><strong>Date of Collection</strong></td>
</tr>
<tr>
<td><strong>Start</strong></td>
</tr>
<tr>
<td><strong>End</strong></td>
</tr>
<tr>
<td><strong>Country/Nation</strong></td>
</tr>
<tr>
<td><strong>Geographic Coverage</strong></td>
</tr>
<tr>
<td><strong>Geographic Unit</strong></td>
</tr>
<tr>
<td><strong>Geographic Bounding Box</strong></td>
</tr>
<tr>
<td><strong>West Longitude</strong></td>
</tr>
<tr>
<td><strong>East Longitude</strong></td>
</tr>
<tr>
<td><strong>North Latitude</strong></td>
</tr>
<tr>
<td><strong>South Latitude</strong></td>
</tr>
</tbody>
</table>
Ontology Editor:
(e.g., Publicly available SKOS Concept Schemes for Biomedical Metadata)

Controlled Vocabulary updates sent to Dataverse (via RDF/XML export or API (JSON-LD))

Dataverse Metadata

Export in FAIRPORT DCAT RDF/XML
Upcoming
Expanding to Larger and More Types of Data

- Sharing sensitive data with DataTags and Secure Dataverse

- Integration with other systems:
  - OSF
  - DataUp
  - WorldMap
  - DataBridge
  - ORCID
  - DASH (at Harvard)

- Expand to Larger data sets
## Data Tags

Sharing data with confidence

### Data Tags: For Sharing Sensitive Data

#### Harm Levels, and Their Appropriate Tags

<table>
<thead>
<tr>
<th>Level</th>
<th>D.U.A. Agreement Method</th>
<th>Authentication</th>
<th>Transit Encryption</th>
<th>Storage Encryption</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoRisk</td>
<td>None</td>
<td>None</td>
<td>Clear</td>
<td>Clear</td>
</tr>
<tr>
<td>Minimal</td>
<td>None</td>
<td>Email_or_OAuth</td>
<td>Clear</td>
<td>Clear</td>
</tr>
<tr>
<td>Shame</td>
<td>ClickThrough</td>
<td>Password</td>
<td>Encrypted</td>
<td>Encrypted</td>
</tr>
<tr>
<td>CivilPenalties</td>
<td>Sign</td>
<td>Password</td>
<td>Encrypted</td>
<td>Encrypted</td>
</tr>
<tr>
<td>CriminalPenalties</td>
<td>Sign</td>
<td>TwoFactor</td>
<td>Encrypted</td>
<td>Encrypted</td>
</tr>
<tr>
<td>MaxControl</td>
<td>Sign</td>
<td>TwoFactor</td>
<td>DoubleEncryption</td>
<td>DoubleEncryption</td>
</tr>
</tbody>
</table>

Final tags may not match the tags of a specific harm level. Hover over the terms to view an explanation.
Data Tags

Sharing data with confidence

Person-specific

Does your data include personal information?

- YES
- NO

Data Tags

- DUA Agreement Method: n/a
- Authentication Type: n/a
- Transit Encryption Type: n/a
- Storage Encryption Type: n/a

Tagging Complete!

Direct Data Access

- CriminalPenalties

- DUA Agreement Method: Sign
- Authentication Type: TwoFactor
- Transit Encryption Type: Encrypted
- Storage Encryption Type: Encrypted
THANKS
mcrosas@iq.harvard.edu Twitter: merceccrosas
http://datascience.iq.harvard.edu (Beta)