Data Publishing with Dataverse

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

ACRL Webinar, May 22, 2014
Introduction to our Team and Projects
Data Science at IQSS combines expertise in software engineering, statistical innovation and data curation. Meet our team.

CURRENT EFFORTS

Reproducible and Reusable Science
Connecting research results to the underlying data and analysis is central to the validation and extensibility of scientific discoveries. Our tools encourage open data and methodological transparency, when possible, and promote and enable data citation.

Computationally Assisted Exploration
We build analytical tools, such as Consilience and TwoRavens, that assist a wide range of users in their work.

SOFTWARE PROJECTS

Zelig
Zelig is an interface, that allows a large body of different statistical models in the R statistical language to be implemented and interpreted in a common framework and interface.

Dataverse 4.0 Beta
May 8, 2014
The Dataverse team has been hard at work on an extensive rewrite of the Dataverse application. Thanks to helpful feedback...
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

**QA and Tech Support**
- Kevin Condon
- Elda Sotiri

**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

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.

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

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

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

Zelig

Analyze Unstructured Text

Consilience

Analyze Quantitative Datasets

Publish Data

Share Sensitive Data

Cite Data from Published Results

Explore, reanalyze and reuse data

Develop > Analyze > Share > Explore > Validate & Reuse
In collaboration with the Harvard Library, Harvard hosts a Dataverse instance free and open to all researchers across all disciplines.

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

Find or deposit data at: http://thedata.harvard.edu
Dataverse 4.0

This summer:
- 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 TwoRavens.
Integration with TwoRavens

- Users can explore, get summary statistics, and analyze tabular data
- It has access to statistical models in Zelig
User Feedback in Every Step

Dataverse 4.0 Beta available now for Testing:
http://dataverse-demo.iq.harvard.edu
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/dacitation
A Published Dataset cannot be deleted (only de-accessioned, if legally needed)

- **Draft dataset**
- **Published Dataset v1**
- **Published Dataset V1.1**
- **Published Dataset V2**

**Release Version 1**

- Authors, Title, Year, DOI, Repository, UNF, V1
- Push Version 1.1: small metadata change; citation doesn’t change

**Push Version 2**

- Authors, Title, Year, DOI, Repository, UNF, V2
Option A. Publish a dataset to your Dataverse, then provide the Data Citation to the journal.

Option B. 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.

Option C. 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
Seamless System Integration

For Option C:

- 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.

Client (e.g., OJS, DVN-R) → Repository (Dataverse)

- XML file: “Deposit Receipt”

Client can query repository (server) any time to get status
Deposit API based on SWORD

- Follows SWORD2 specifications
- SWORD is supported within academic publishing; based on the web standard Atom Publishing Protocol.

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:
  - Instead, Dataverse uses own Data API
  - Plan to support SWORD download in the future

- Added XML attribute to pass article citation from client:
  - Allow DCterms:isReferencedby to contain attributes (HoldingsURI) to link back to article from Dataverse
  - This is now part of the SWORD PHP client library
Support for Metadata Standards

- **Citation metadata:** Applies to all datasets – Supported currently by Data Deposit API

- Extensible metadata blocks for specific domains (in 4.0):
  - **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

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

Compliant with DataCite, Dublin Core, DDI study description
<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Social Science and Humanities Metadata</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Topic Classification</strong></td>
<td>Term</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Software</strong></td>
<td>Name</td>
</tr>
<tr>
<td><strong>Series</strong></td>
<td>Name</td>
</tr>
<tr>
<td><strong>Time Period Covered</strong></td>
<td>Start: YYYY-MM-DD</td>
</tr>
<tr>
<td><strong>Date of Collection</strong></td>
<td>End: YYYY-MM-DD</td>
</tr>
<tr>
<td><strong>Geographic Coverage</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Geographic Unit</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Geographic Bounding Box</strong></td>
<td>West Longitude</td>
</tr>
<tr>
<td></td>
<td>East Longitude</td>
</tr>
<tr>
<td></td>
<td>North Latitude</td>
</tr>
<tr>
<td></td>
<td>South Latitude</td>
</tr>
</tbody>
</table>

**Compliant with DDI for Social Sciences**
Compliant Virtual Observatory (VO) schema
<table>
<thead>
<tr>
<th>Design Type</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Case Control</td>
<td></td>
</tr>
<tr>
<td>Cross Sectional</td>
<td></td>
</tr>
<tr>
<td>Not Specified</td>
<td></td>
</tr>
<tr>
<td>Parallel Group Design</td>
<td></td>
</tr>
<tr>
<td>Perturbation Design</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factor Type</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>Biomarkers</td>
<td></td>
</tr>
<tr>
<td>Developmental Stage</td>
<td></td>
</tr>
<tr>
<td>Cell Surface Markers</td>
<td></td>
</tr>
<tr>
<td>Cell Type/Cell Line</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measurement Type</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA Methylation Profiling (Bisulfite-Seq)</td>
<td></td>
</tr>
<tr>
<td>DNA Methylation Profiling (MeDIP-Seq)</td>
<td></td>
</tr>
<tr>
<td>Histone Modification (ChIP-Seq)</td>
<td></td>
</tr>
<tr>
<td>Protein-RNA Binding (RIP-Seq)</td>
<td></td>
</tr>
<tr>
<td>Transcription Factor Binding (ChIP-Seq)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Organism</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Danio rerio</td>
<td></td>
</tr>
<tr>
<td>Homo sapiens</td>
<td></td>
</tr>
<tr>
<td>Mus musculus</td>
<td></td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cell Type</th>
<th></th>
</tr>
</thead>
</table>
A workflow to access and update Controlled Vocabularies

Ontology Editor:
(e.g., Publicly available SKOS Concept Schemes for Biomedical Metadata)

Dataverse Metadata

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

Export in FAIRPORT DCAT RDF/XML

HSCI Stem Cell Taxonomy DEMO
- HSCI Stem Cell Concept Scheme (8)
  - Cell Type (0)
  - Cellular Material (0)
  - Development Stage (0)
  - Disease State (0)
  - Measurement Type (0)
  - Organism (4)
    - Danio rerio (0)
    - Homo sapiens (0)
    - Mus musculus (0)
    - Rattus norvegicus (0)
  - Tissue Type (0)
  - Treatment (0)
Future Projects
Expanding to support more Data

- Sharing sensitive data with DataTags and Secure Dataverse

- Integration with other systems:
  - Open Science Framework
  - DataUp
  - WorldMap
  - DataBridge
  - ORCID
  - ...

- Expand to Large-scale datasets with efficient data storages
## DataTags: For Sharing Sensitive Data

Data Tags: Sharing data with confidence

### 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>Civil Penalties</td>
<td>Sign</td>
<td>Password</td>
<td>Encrypted</td>
<td>Encrypted</td>
</tr>
<tr>
<td>Criminal Penalties</td>
<td>Sign</td>
<td>TwoFactor</td>
<td>Encrypted</td>
<td>Encrypted</td>
</tr>
<tr>
<td>Max Control</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

Full Interview

Tagging Complete!

Direct Data Access

<table>
<thead>
<tr>
<th>Data Tag</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUA Agreement Method</td>
<td>n/a</td>
</tr>
<tr>
<td>Authentication Type</td>
<td>n/a</td>
</tr>
<tr>
<td>Transit Encryption Type</td>
<td>Encrypted</td>
</tr>
<tr>
<td>Storage Encryption Type</td>
<td>Encrypted</td>
</tr>
</tbody>
</table>

Sign
TwoFactor
Try Dataverse 4.0 Beta and give us feedback:
http://dataverse-demo.iq.harvard.edu
Learn more about our projects at:
http://datascience.iq.harvard.edu