Dataverse with DataTags: Sharing Data you can’t share

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

Michael Bar-Sinai  @michbarsinai
Architect, Senior Software Engineer,
Institute for Quantitative Social Science, Harvard University

http://datascience.iq.harvard.edu
Introduction to Dataverse

Dataverse Software

- A framework for publishing, citing and preserving research data: http://thedata.org
- Open-source, available at GitHub
- Started in 2006 at IQSS
- Can support all data types across multiple disciplines
- APIs to integrate with journal systems and other repositories

Dataverse Repository

- Harvard hosts a Dataverse instance free and open to all research data: http://thedata.harvard.edu
- More than 53,000 datasets, with 735,000 files
- Dataverses can be created for researchers, journals, organizations, educators, …
- It federates with > 10 Dataverse installations around the world.
Find and publish data at: http://thedata.harvard.edu
Dataverse allows you to:

- Get a formal citation for your data
- Link your data set to the original publication(s)
- Publish multiple versions of your datasets
- Set terms of use for your data
- Restrict data files, while metadata and documentation can be kept public (but we encourage open data, when possible)
- Brand your dataverse banner with your logo, image or colors
- Track downloads for your data, and enable a guestbook
- List data sets from other dataverses in your dataverse
Dataverse 4.0 (Fall 2014)

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

Try Dataverse 4.0 Beta: http://dataverse-demo.iq.harvard.edu
Dataverse 4.0 will include a new interactive data exploration and analysis tool, TwoRavens, which integrates with Zelig statistical framework.
Title: Replication Data for: Building a Bridge Between

Author: 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 data analysis of current data publication practices.

Keyword: data publication

Subject: Social Sciences
<table>
<thead>
<tr>
<th>Topic Classification</th>
<th>Term</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>URL</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Software</th>
<th>Name</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Series</th>
<th>Name</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time Period Covered</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>YYYY-MM-DD</td>
<td>YYYY-MM-DD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date of Collection</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>YYYY-MM-DD</td>
<td>YYYY-MM-DD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country/Nation</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Geographic Coverage</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Geographic Unit</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Geographic Bounding Box</th>
<th>West Longitude</th>
<th>East Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>North Latitude</th>
<th>South Latitude</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Astronomy Metadata: Compliant Virtual Observatory (VO) schema; extract metadata from FITS files
Bio Metadata:
Compliant with ISA-Tab schema, plus biomedical ontologies
Dataverse is part of a 4 years NSF funded project on Privacy Tools for Sharing Sensitive Data [http://privacytools.seas.harvard.edu/](http://privacytools.seas.harvard.edu/) (with Harvard SEAS, Berkman Center, Data Privacy Lab, and IQSS).

This project includes:

- **DataTags**: A framework that provides data handling prescriptions to comply with numerous privacy regulations and data user agreements
- **Private Zelig**: A differential privacy version of the Zelig statistical framework
Try our new Beta version: http://datatags.org

Currently supporting HIPAA and FERPA (and DUAs)

<table>
<thead>
<tr>
<th>Harm Levels and Their Appropriate Tags</th>
<th>DUA Agreement Method</th>
<th>Authentication</th>
<th>Transit</th>
<th>Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Risk</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>Click Through</td>
<td>Password</td>
<td>Encrypted</td>
<td>Clear</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>Two Factor</td>
<td>Encrypted</td>
<td>Encrypted</td>
</tr>
<tr>
<td>Max Control</td>
<td>Sign</td>
<td>Two Factor</td>
<td>Double Encryption</td>
<td>Double Encryption</td>
</tr>
</tbody>
</table>
DataTags behind the scenes:
A complex interview generation framework, which is automatically converted to a user-friendly interface
Interview Example: First question ...

Person-specific. Does your data include personal information?

Terms
personal information
as defined in HIPAA

data
0s and 1s in some structured way

yes  no
Interview Example: After several questions ...
Interview Example: ... and a Final Tag

---

Your dataset is tagged as

Very sensitive identifiable personal information, shared with strong verification of approved recipients under signed agreement.

### Full Tags

<table>
<thead>
<tr>
<th>DataTags</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>code</td>
<td>red</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DataType</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>harm</td>
<td>criminal</td>
</tr>
<tr>
<td>effort</td>
<td>identifiable</td>
</tr>
<tr>
<td>standards</td>
<td>HIPAA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Handling</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>storage</td>
<td>encrypt</td>
</tr>
<tr>
<td>auth</td>
<td>approval</td>
</tr>
<tr>
<td>transit</td>
<td>encrypt</td>
</tr>
<tr>
<td>basis</td>
<td>HIPAABusinessAssociate</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DUA</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>timeLimit</td>
<td>_5yr</td>
</tr>
</tbody>
</table>

---
The DataTags project consists of several distinct components.
Algorithm

- "Harmonizes law and technology"
- Consists of a tag ontology and an interview process
- Created by legal and technological experts
- Currently Supports HIPAA, FERPA, CIPSEA and Privacy Act
- Developed by Berkman, DPL and IQSS
Ontology definition language
• Define an interview and coding process: ask Questions, Set values to the tags
• Allows localization and extension
• Supports any closed-ended questionnaire. DataTags is a private case of this.

Interview and coding language
• Defines tagging ontologies
• Allows atomic (simple), aggregate and compound values
Tag Definition

DataTags: code, basis, Handling, DataType, DUA, IP, identity, FERPA, CIPSEA.

TODO: IP.

code: one of
- blue     (Non-confidential information),
- green    (Potentially identifiable but not...),
- yellow   (Potentially harmful personal information...),
- orange   (May include sensitive, identifiable information...),
- red      (Very sensitive identifiable personal information...),
- crimson  (Requires explicit permission for each transaction...).

Handling: storage, transit, authentication, auth.

storage: one of clear, encrypt, doubleEncrypt.

standards: some of HIPAA, FERPA, ElectronicWiretapping, CommonRule, CIPSEA.
Questionnaire Definition

(>medical-start< ask:
  (text: Person-specific. Does your data include personal information?)
  (terms:
    (data: 0s and 1s in some structured way)
    (personal information: as defined in HIPAA))
  (no:
    (set: code=green, storage=clear, transit=clear, auth=none,
      basis=notApplicable, identity=notPersonSpecific,
      harm=negligible)
    (end)
  )
)
(>ec< ask:
  (text: Explicit Consent. Did each person whose information appears in the data give explicit permission to share the data?)
  (yes:
    (set: basis=consent)
    (ask:
      (text: Did the consent have any restrictions on data sharing?)
      (no: (set: code=green, storage=clear, transit=clear, auth=none))
      (yes: (call: dua))
      (end)
    )
  ))
Tools

- Editing: Any text editor
- Compiler
- Visualizers
- Runtime Engine
- Java library
- Command-line Runner
Tools: Visualizations
Tagging Server

- Web-based GUI for the runtime engine
- Focus on usability
- Integration with other systems, most notably data repositories such as Dataverse, via API
- Will allow other teams to develop tagging interviews
http://www.datatags.org
Standard Tag Set

- Allows the tagging process to be machine-actionable
- Data repositories will recognize the set, and will know how to operate according to its possible tagging values
Secure Dataverse

- A data repository that can interpret a standard set of data tag, and handle datasets accordingly
- Tagging the data is part of the data ingest process
Learn more at: http://datascience.iq.harvard.edu