Mercè Crosas, Ph.D.  
Chief Data Science and Technology Officer  
Institute for Quantitative Social Science  
Harvard University  
@mercecrosas

THE DATAVERSE COMMONS

The Future of the Commons, November 18, 2015,
Most scientific studies now involve large amounts of digital data & software for analysis.
Data Repositories vs Repository Software

**Domain-specific repositories**
- GenBank
- WW Protein Data Bank
- SBGrid Data Bank
- ...

**General-purpose repositories**
- Harvard Dataverse
- DataDryad
- Figshare
- ...

**Repository Software**
- Dataverse Software
- Dspace
- Fedora
- ...

...
Dataverse follows best practices to support Data Publishing

A formal data citation
- Reference with attribution
- Access with a persistent identifier

Information about the data (metadata)
- Discovery
- Data reuse

A trusted data repository
- Access to data and metadata (long-term archival)

Data Publishing
The Dataverse Project

dataverse.org

Open-source software developed at Harvard’s IQSS since 2006
Installed in 12 sites world wide
Serving 100s of universities and organizations
Harvard Dataverse: dataverse.harvard.edu

Started as a community data repository for Social Science
Now open to all research fields and all researchers
More than 1300 dataverses
More than 59,000 datasets
More than 1,400,000 downloads
Dataverses are containers for Datasets

Each Dataverse can be for a researcher, a research project, a department, a journal, or a larger organization.
Dataverse offers a rich feature set to publish research data

**Credit and Visibility**
- Standard, persistent data citation
- Branding for each dataverse
- Widgets to embed in your own website

**Discovery**
- Faceted search for all metadata
- Standard metadata:
  - citation
  - scientific domain
  - file-level

**Access Control & Roles**
- CCO waiver for public datasets
- Tiered access:
  - terms of use
  - guestbook
  - restricted data
- Publishing workflow
- Multiple roles:
  - contribute
  - curate, review
  - administrate

**Data Features**
- Versioning
- Conversion of tabular data files to standard format
- Automatic extraction of file metadata (R, STATA, SPSS, XSLX, FITS)

**Interoperability through APIs**
- Journal Systems (Open Journal System, ScholarOne); Open Science Framework
- Data Analysis (TwoRavens); Spatial Viz (WorldMap); Preservation systems (Archivematica)
Impact on the Social Science research community and on the World

Antislavery petitions data

Election Data Archive

Project TIER

Boston Area Research Initiative
3,500 antislavery and antisegregation petitions sent to Massachusetts from 1600s to 1870
A collaborative archive to share election results, voting behavior, and electoral politics.

Alaska electoral data: 1,500 data downloads
Project TIER
Teaching Integrity in Empirical Research

Provides a protocol to document all steps in data management and analysis:
• Data Files
• Metadata Files
• Computing Command Files
• Readme File
Boston Area Research Initiative
Daniel O’Brien (Northeastern), Robert Sampson, Christopher Winship (Harvard)

Scholars, policymakers, practitioners and civic leaders collaborating on social science research and public policy

• Dataset of Bicycle Collisions in Boston (in collaboration with Boston Police, Harvard School of Public Health, and Cyclists Union)
• Data visualization with WorldMap
Future impact on other research communities: Biomedical and Astronomy

- Data archival
- Conversion to standard formats
- Extraction of file-level metadata
Current Collaborations: Addressing the Next Challenges in Data Sharing

Structural Biology Grid Data Repository (Sliz, HMS, Crosas, IQSS)

Social Science Big Data (King, Crosas, IQSS, CGA)

Data Provenance (Seltzer, SEAS, Crosas, King, IQSS)

Privacy Tools to share sensitive data (SEAS, Berkman Center, Privacy Lab, IQSS, MIT)
# Sharing Sensitive Data with Confidence: DataTags System

<table>
<thead>
<tr>
<th>Tag Type</th>
<th>Description</th>
<th>Security Features</th>
<th>Access Credentials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>Public</td>
<td>Clear storage, Clear transmit</td>
<td>Open</td>
</tr>
<tr>
<td>Green</td>
<td>Controlled public</td>
<td>Clear storage, Clear transmit</td>
<td>Email- or OAuth Verified Registration</td>
</tr>
<tr>
<td>Yellow</td>
<td>Accountable</td>
<td>Clear storage, Encrypted transmit</td>
<td>Password, Registered, Approval, Click-through DUA</td>
</tr>
<tr>
<td>Orange</td>
<td>More accountable</td>
<td>Encrypted storage, Encrypted transmit</td>
<td>Password, Registered, Approval, Signed DUA</td>
</tr>
<tr>
<td>Red</td>
<td>Fully accountable</td>
<td>Encrypted storage, Encrypted transmit</td>
<td>Two-factor authentication, Approval, Signed DUA</td>
</tr>
<tr>
<td>Crimson</td>
<td>Maximally restricted</td>
<td>Multi-encrypted storage, Encrypted transmit</td>
<td>Two-factor authentication, Approval, Signed DUA</td>
</tr>
</tbody>
</table>

*DataTag: A set of security features and access requirements for file handling*

Sweeney, Crosas, Bar-Sinai, 2015, Technology Science
Data Sharing Workflow for Sensitive Data

Sensitive Dataset → The Dataverse Project → Sensitive Dataset

Direct Access
Authorized Signed DUA
Privacy Preserving Access

http://datatags.org
http://privacytools.seas.harvard.edu