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

DATaverse ON THE MOC

MOC Workshop, Boston University, November 19, 2015
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
- ...

...
The Dataverse Project
dataverse.org

Open-source software developed at Harvard’s IQSS since 2006
Used to share, publish, cite and archive research data
Installed in 12 sites worldwide
Serving 100s of universities and organizations
Harvard Dataverse: dataverse.harvard.edu

Started as a community 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

Dataverses are containers for Datasets

Schematic Diagram of a Dataverse in Dataverse 4.0

Schematic Diagram of a Dataset in Dataverse 4.0

Container for your Datasets and/or Dataverses*

* Dataverses can now contain other Dataverses (this replaces Collections & Subnetworks)

Container for your data, documentation, and code.

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 (Archivemática)
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 (Red)
- Authorized Signed DUA (Red)
- Privacy Preserving Access (Green)

Sensi=ve Dataset

http://datatags.org
http://privacytools.seas.harvard.edu
Dataverse on the MOC

Current Architecture

- UI Layer (PrimeFaces, js)
- Application Logic (Java EE)
- PostgreSQL (user data, metadata)
- Solr (Index)

On the MOC

- UI Layer (PrimeFaces, js)
- Application Logic (Java EE)
- PostgreSQL (user data, metadata)
- Solr (Index)

- SWIFT object storage
- CINDER block storage

- COMPUTE SERVICES
  (R, Python, Spark, Hadoop, …)

Network File System (data files)
RServe (R ingest, analysis)
Dataverse-MOC Projects

We propose to pilot a framework for sharing and publishing large and streaming datasets, and enabling collaborative computing on them.

**Boston Data from City Hall and Boston Area Research Initiative (BARI):**
- Dan O’Brien (Northeastern University)
- Storage: 911 and 311 calls streaming data:
  - 311 calls: 800,000 reports, at 500 reports/day
  - 911 calls: 2,500,000 reports, at 1,500 reports/day
- Computing: merge data + geospatial, temporal exploration

**Partners Healthcare Clinical Trial Data:**
- Shawn Murphy (Partners Healthcare)
- Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS)
  - Data from 14 health systems sites
- Sensitive data sets, categorized using the DataTags levels
THANKS

@mercecrosas
mcrosas@iq.harvard.edu
http://scholar.harvard.edu/mercecrosas