ADDRESSING THE NEXT CHALLENGES IN DATA SHARING: LARGE-SCALE DATA AND SENSITIVE DATA

Mercè Crosas, Ph.D.
Chief Data Science and Technology Officer
Institute for Quantitative Social Science
Harvard University
@mercecrosas
## Data sharing:
**good for you and good for the world**

<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researchers</td>
<td>Get credit for their data</td>
</tr>
<tr>
<td>Publishers and Journals</td>
<td>Verify published work</td>
</tr>
<tr>
<td>Federal funding agencies</td>
<td>Make public assets accessible</td>
</tr>
<tr>
<td>Science</td>
<td>Validate, reuse and extend previous work</td>
</tr>
</tbody>
</table>
Data Sharing needs to support data discovery, referencing, access, and reuse

- A formal data citation
  - Reference
  - Access (persistent identifier)

- Information about the data (metadata)
  - Discovery
  - Use

- A trusted data repository
  - Access (long-term archival)
The Dataverse Project

Open-source software developed at Harvard’s IQSS since 2006
Used to share, publish, cite and archive research data
Installed in 12 sites world wide
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
# Data Sharing with Dataverse

<table>
<thead>
<tr>
<th>Now</th>
<th>The Next 5 Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>• No sensitive data</td>
<td>• Highly-sensitive data</td>
</tr>
<tr>
<td>• Seldom versioning</td>
<td>• Streaming or frequently updated data</td>
</tr>
<tr>
<td>• Datasets up to ~GB</td>
<td>• Datasets &gt; GBs, TBs, PBs</td>
</tr>
<tr>
<td></td>
<td>– Thousands of files per dataset</td>
</tr>
<tr>
<td></td>
<td>– Large dataset in a Big Data, NoSQL storage (MongoDB, Cassandra, Lucene)</td>
</tr>
</tbody>
</table>

Large-scale data sharing needs to continue supporting discovery, referencing, access and reuse.
Adhering to the same high standards for large-scale data

• Metadata for **discovery**:
  – citation metadata
  – domain-specific descriptive metadata
  – file-level or variable metadata

• Data citation for **reference and access**:
  – for entire dataset and for subsets of the dataset
    (based on time of retrieval or variables selected)

• Fast queries, data exploration and visualizations for **reuse**:
  – might not be able to download entire dataset
Data retrieval, explorations and visualizations of large-scale datasets require **data repositories** be closer to computing resources.
Current collaborations to address the next challenges in data sharing

- SB Grid Data Repository (HMS, IQSS)
- Social Science Big Data (IQSS)
- Data Provenance (SEAS, IQSS)
- Privacy Tools to share sensitive data (SEAS, Berkman, Privacy Lab, IQSS, MIT)
Sharing and Preserving Large Structural Biology Data

https://data.sbgrid.org/
Structural Biology
Primary Data

Integration with Dataverse:

- Long-term access
- Formal Data Citation
- Standard Metadata
- Data Exploration (OME)
- Preservation, with copies in multiple sites (following dataPASS approach)

1 Dataset is 180-360 images of X-ray diffraction data, 3.5-7 GB; ~1TB per dataset, with a total up to 100 PBs
Dataverse on the Massachusetts Open Cloud (MOC): Computing closer to data storage

Current Architecture

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

Network File System (data files)
RServe (R ingest, analysis)

On the MOC

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

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

- SWIFT object storage
- CINDER block storage
# 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, “Sharing Sensitive Data with Confidence: The DataTags System” Technology Science
Data Sharing Workflow for Sensitive Data

Sensitive Dataset

Direct Access

Authorized Signed DUA

Privacy Preserving Access

The Dataverse Project

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

Piotrek Sliz (SBGrid, HMS), Latanya Sweeney (Data Privacy Lab, Harvard), Dataverse team (IQSS, Harvard)
@mercecrosas