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

DATA PUBLISHING
350 years ago, the first issue of Philosophical Transactions was published by the Royal Society, under the motto “Nullius in verba” (or “Take nobody’s word for it”)
Most scientific studies now involve large amounts of digital data & software for analysis.
Data publishing: It’s good for you and good for the world

- **You**: Get credit for your data
- **Publishers and Journals**: Verify published work
- **Federal funding agencies**: Make public assets accessible
- **Science**: Validate, reuse and extend previous work
Sharing Data Increases Citations

*From 10,555 studies with gene expression microarray data:*

- Studies that shared data received 9% **more citations**

- **Data reuse** by third-party investigators continued for 6 years

Long-Term Accessibility must be Considered

Analysis of 7,641 Publications from 4 major journals in Astronomy and Astrophysics, between 1997 and 2008

Pepe, Goodman, Muench, Crosas, Erdmann, 2014 “Sharing, Archiving and Citing Data in Astronomy” PLOSOne
Data Publishing needs to support data discovery, reference, access, and use.

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

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

- A trusted data repository
  - Access (long-term archival)
Data Citation Principles

1. Data should be citable products of research
2. Credit and Attribution
3. Evidence
4. Unique Identification
5. Access
6. Persistence
7. Specificity and Verifiability
8. Interoperability and flexibility

Full Principles: https://www.force11.org/datacitation
Data Repositories vs Repository Software

Domain-specific repositories
- Gen Bank
- Protein Data Bank
- SBGrid Data
- ...

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

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

...
The Dataverse Project

dataverse.org

Open-source software developed at Harvard’ IQSS since 2006
Installed in 12 sites world wide
Serving 100s of universities and organizations
Harvard Dataverse: dataverse.harvard.edu
Open to all research fields and all researchers
More than 1200 dataverses
More than 59,000 datasets
More than 1,400,000 downloads

The Institute for Quantitative Social Science
HARVARD UNIVERSITY
Information Technology
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

**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, XSD, FITS)

**Interoperability through APIs**
Journal Systems (Open Journal System, ScholarOne); Open Science Framework
Data Analysis (TwoRavens); Spatial Viz (WorldMap); Preservation systems (Archivematica)
What you can do with file-level metadata and APIs Now

Anti-slavery petitions data

Statistical analysis with TwoRavens

Tuberculosis Genomics data

Boston Area Research Initiative
data visualization in WorldMap
What you will be able to do with Image Data in Dataverse

- OME-TIFF Files
  - Conversion to standard formats
  - Extraction of file-level metadata

- FITS Files

OMERO

WORLD WIDE TELESCOPE
Current Collaborations

SB Grid Data Repository (HMS, IQSS)

Data Provenance (SEAS, IQSS)

Social Science Big Data (IQSS)

Privacy Tools to share sensitive data (SEAS, Berkman, 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
A DataTags Repository is a repository of files held for Data Sharing that:

1. Supports more than one datatag
2. Each file in the repository must have one datatag
3. A recipient of a file from the repository must:
   a. satisfy file’s access requirements,
   b. produce sufficient credentials as requested,
   c. and agree to any terms of use required to acquire the file.
4. Provides technological guarantees for requirements 1, 2 and 3.

Data Publishing Workflow for Sensitive Data

Sensitive Dataset → Blue, Green, Yellow, Orange, Red, Crimson → The Dataverse Project → Sensitive Dataset → Direct Access, Curator Model
A Curator Model for Privacy-Preserving Analysis

Differentially Private statistics (summaries, causal inference, regression, interactive queries)

Acknowledgement: Honaker, J. and Nissim, K., Data Privacy Tools Project
Acknowledgement: Latanya Sweeney, James, Honaker, Eleni Castro, Margo Seltzer, Piotrek Sliz, Christine Choirat, Garth Griffin, and the Dataverse team for graphics and slides