CLOUD DATAVERSE

Mercè Crosas, Institute for Quantitative Social Science, Harvard University

@mercecrosas

MOC WORKSHOP, OCTOBER 3, 2017, BOSTON UNIVERSITY
OUR INSTITUTE PROVIDES A TECHNOLOGY SOLUTION TO DATA SHARING

Institute for Quantitative Social Science, Harvard University

@IQSS
An open-source software to share, cite, and find data. Developed at Harvard's Institute for Quantitative Social Science with the contribution of an active and growing community.
2006 (we started)

26 Dataverse installations serving hundreds of institutions

dataverse.org
**DataVerse**

A public repository for research data

- > 70,000 datasets total
- > 49,000 datasets uploaded to
Harvard Dataverse repository
- 200 datasets/month

- > 340,000 files
- 4,000 files/month

- > 2.5 M downloads
- 60,000 downloads/month

[dataverse.harvard.edu](http://dataverse.harvard.edu)
<table>
<thead>
<tr>
<th>OUR CONTRIBUTIONS TO ENHANCE DATA SHARING</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>King, 1995, Replication, Replication</strong></td>
</tr>
<tr>
<td>Altman et al, 2001, A Digital Library for the Dissemination and Replication of Quantitative Social Science</td>
</tr>
<tr>
<td>Altman and King, 2007, A Proposed Standard for the Scholarly Citation of Quantitative Data</td>
</tr>
<tr>
<td>King, 2007, An Introduction to the Dataverse Network as an Infrastructure for Data Sharing</td>
</tr>
<tr>
<td>Crosas, 2012, The Dataverse Network: an open source application for sharing, discovering, and preserving research data</td>
</tr>
<tr>
<td>Crosas, 2013, A Data Sharing Story</td>
</tr>
<tr>
<td>Altman and Crosas, 2013, The Evolution to Data Citation: from principles to implementation</td>
</tr>
<tr>
<td><strong>2014, Joint Declaration of Data Citation Principles</strong></td>
</tr>
<tr>
<td>Pepe et al, 2014, How Do Astronomers Share Data?</td>
</tr>
<tr>
<td>Goodman et al, 2014, Ten Simple Rules for the Care and Feeding of Scientific Data</td>
</tr>
<tr>
<td>Bierer, Crosas, Pierce, 2017, Data Authorship as an Incentive to Data Sharing</td>
</tr>
<tr>
<td><strong>Castro et al, 2015, Achieving Human and Machine Accessibility of Cited Data</strong></td>
</tr>
<tr>
<td>Sweeney, Crosas, Bar-Sinai, 2015, Sharing Sensitive Data with Confidence: The DataTags System</td>
</tr>
<tr>
<td>Meyer et al. 2016, Data Publication with the Structural Biology Data Grid Supports Live Analysis</td>
</tr>
</tbody>
</table>
Data should be ...

FINDABLE
ACCESSIBLE
INTEROPERABLE
REUSABLE

Wilkinson et al., 2016, "The FAIR Guiding Principles for Scientific Data Management and Stewardship"
Nature Scientific Data
FAIR DATA IN DATAVERSE

Data Citation with Persistent Identifier

Data Files

Metadata

Data Licenses, User Agreements, Restrictions

Versions

APIs
Cloud Dataverse combines the power of cloud computing and storage with access to thousands of datasets from a feature-rich data repository platform
WHY CLOUD DATAVERESE?

- Big Data should also be **FAIR Data**
- Datasets are replicated to the Cloud for efficient access and reuse
- Computing on a dataset is enabled directly from any repository
WHAT WE HAVE BUILT

• Dataverse integration with Swift storage
• Compute access to MOC from a dataset page in Dataverse
• Temporary url to access restricted files in MOC

IN PROGRESS

• Implement Swift Access Control List (ACL) for file restriction
• Support InCommon for MOC to use same credentials as in Dataverse

NEXT

• Replicate data from any Dataverse to Cloud Dataverse
• Upload data directly in Swift; publish dataset from Swift to Dataverse
INTEGRATION WITH OTHER PROJECTS
Draft Architecture for Geospatial Analytics with GPUs and Spark on the MOC
Big geodata exploration and analytics
DATA PROVENANCE
TRACK THE ORIGINAL SOURCE OF A DATASET
Comment: If these data could talk

Thomas Pasquier, Matthew K. Lau, Ana Trisovic, Emery R. Bose, Ben Coutier, Mercè Crosas, Aaron M. Ellison, Valerie Gibson, Chris R. Jones & Margo Seltzer

In the last few decades, data-driven methods and open-source software have enabled the measurement and analysis of the growing flood of data. However, data fields exhibit distressingly low rates of reproducibility. Here, we believe that there is a lack of formal records that connect the data source to the analysis to the results, which contributes to issues of reproducibility through systematic and formal records of the publications and researchers.

Reproducibility

The success and power of science depends on the ability to reproduce experiments and test hypotheses by other scientists. The lack of reproducibility is a growing concern in fields ranging from biology to medicine. Although the lack of reproducibility remains a worrisome issue, this comes at a time when the amount of data and the computational demands are growing exponentially. At the same time, the data and the computing infrastructure are becoming more accessible and affordable.

Alice

wasAssociatedWith

Process 1

used

File 1

wasInformedBy

File 2

used

Process 2

wasGeneratedBy

File 3

wasAssociatedWith

Bob

Pasquier, Lau, Trisovic, Bose, Coutier, Crosas, Ellison, Gibson, Jones, Seltzer, 2017, If These Data Could Talk, Nature Scientific Data
(Data Provenance examples from CERN and Harvard Forest)
DATA PRIVACY
CLASSIFY AND HANDLE DATASETS BASED ON THEIR PRIVACY LEVEL
Dataverse® as a DataTags repository

Data file deposit
Assistance to assign DataTag from:
- DataTags automated interview
- RobotLawyer autogenerated data user agreements (DUA)
- Review Board

Direct Access
Requires:
- User registration
- Approval needed for access
- Signed DUA

Privacy Preserving Access
- Requires user registration
- Provides access to differentially private statistics using Private data Sharing Interface (PSI)

Harvard Data Privacy Tools Project: privacytools.seas.harvard.edu
DataTags Project: datatags.org
THANKS

@mercecrosas
@iqss

scholar.harvard.edu/mercecrosas
dataverse.org