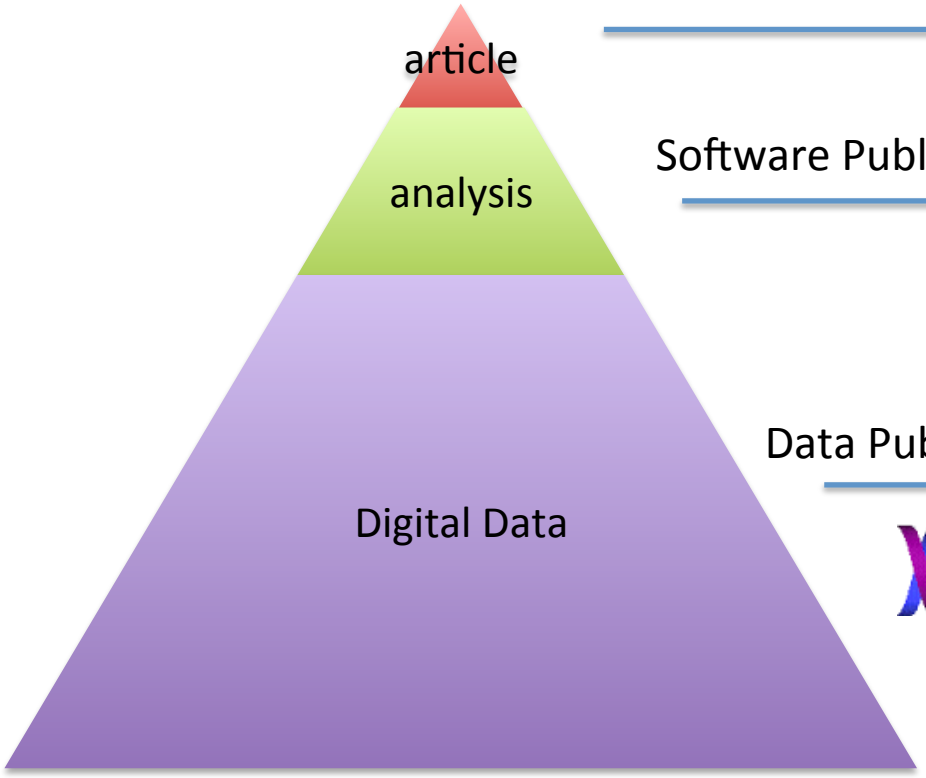
Mercè Crosas, Ph.D.
Chief Data Science and Technology Officer
Institute for Quantitative Social Science
Harvard University
@mercecrosas

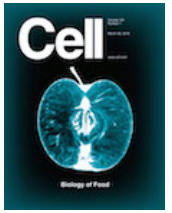
THE DATAVERSE COMMONS

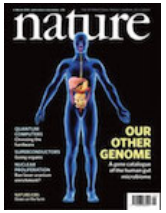
The Future of the Commons, November 18, 2015,

Today's Scholarly Publication Landscape



Article Publishing





PLOS

Software Publishing

GitHub

zenodo

Data Publishing

WORLDWIDE PDB
PROTEIN DATA BANK

NCBI GenBank

The Dataverse Project

OpenfMRI

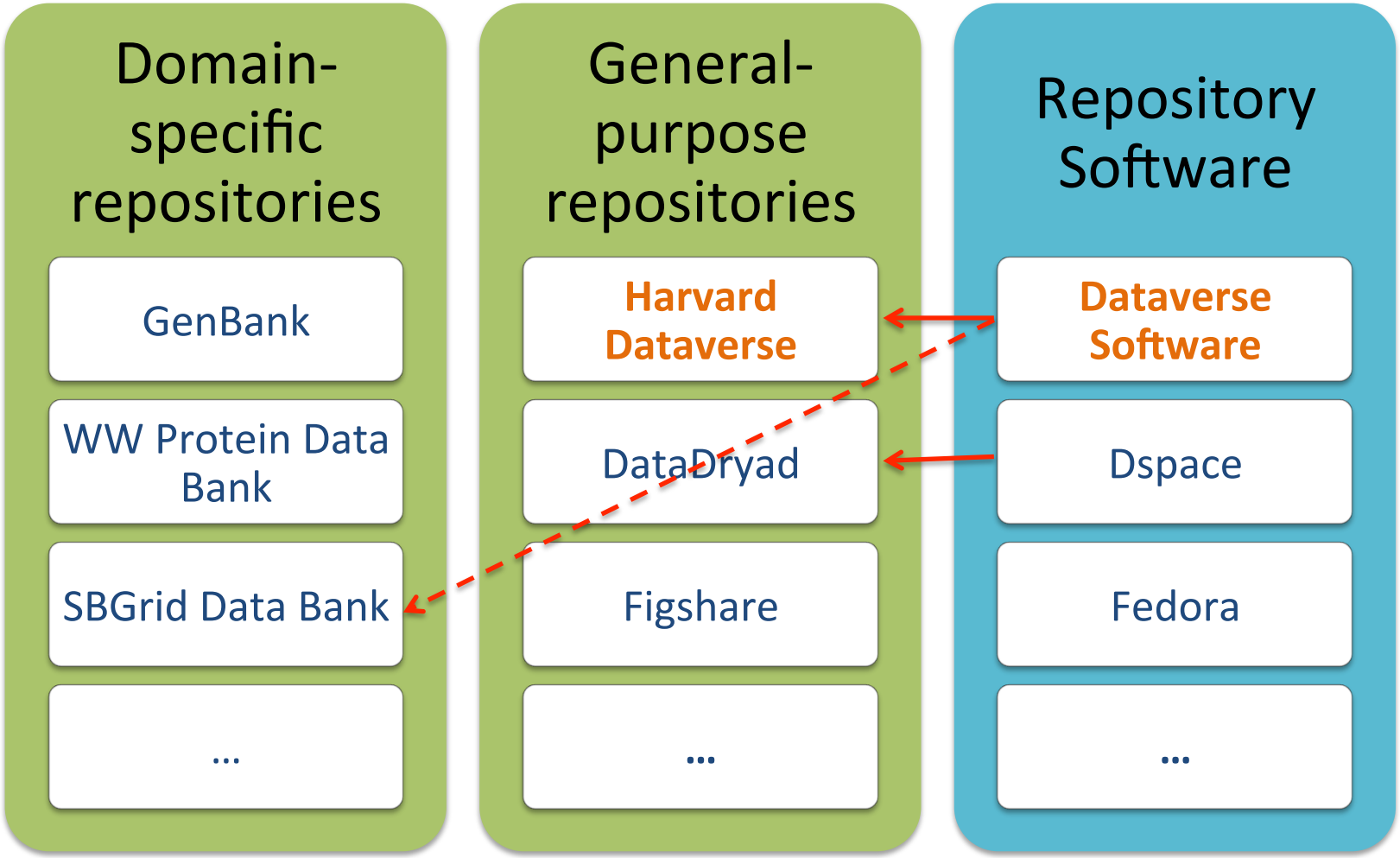
DRYAD

SBGrid CONSORTIUM

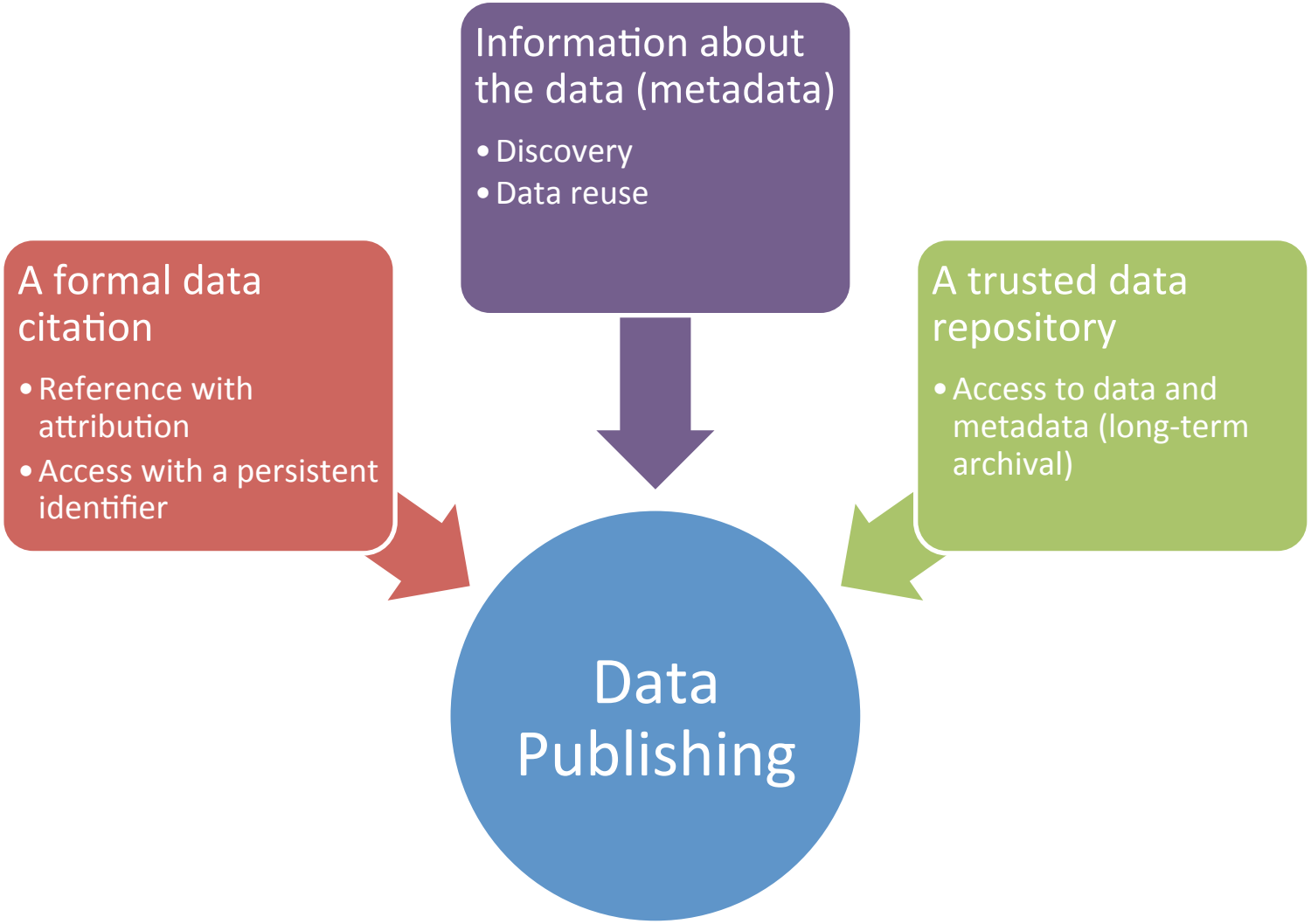
figshare

Most scientific studies now involve large amounts of digital data & software for analysis.

Data Repositories vs Repository Software



Dataverse follows best practices to support Data Publishing



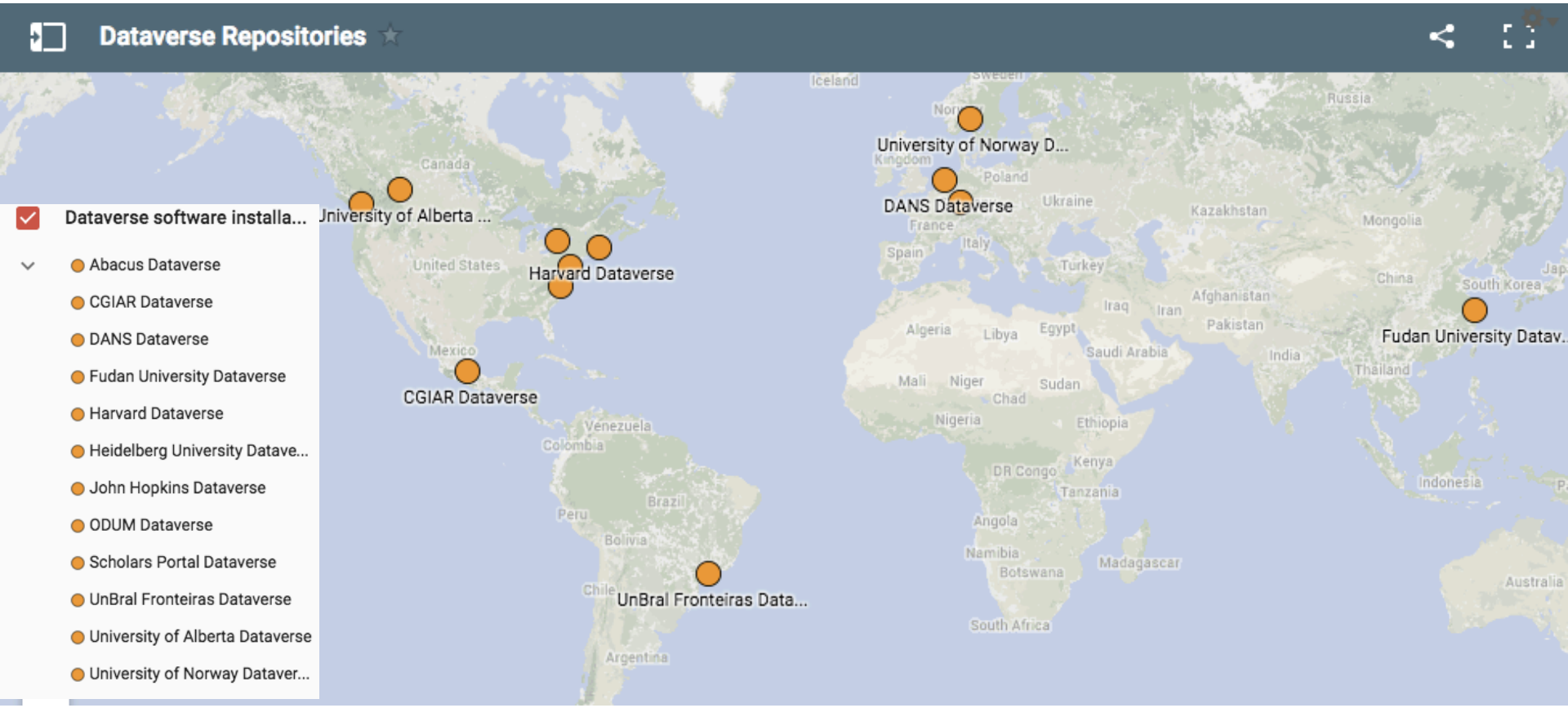
The Dataverse Project

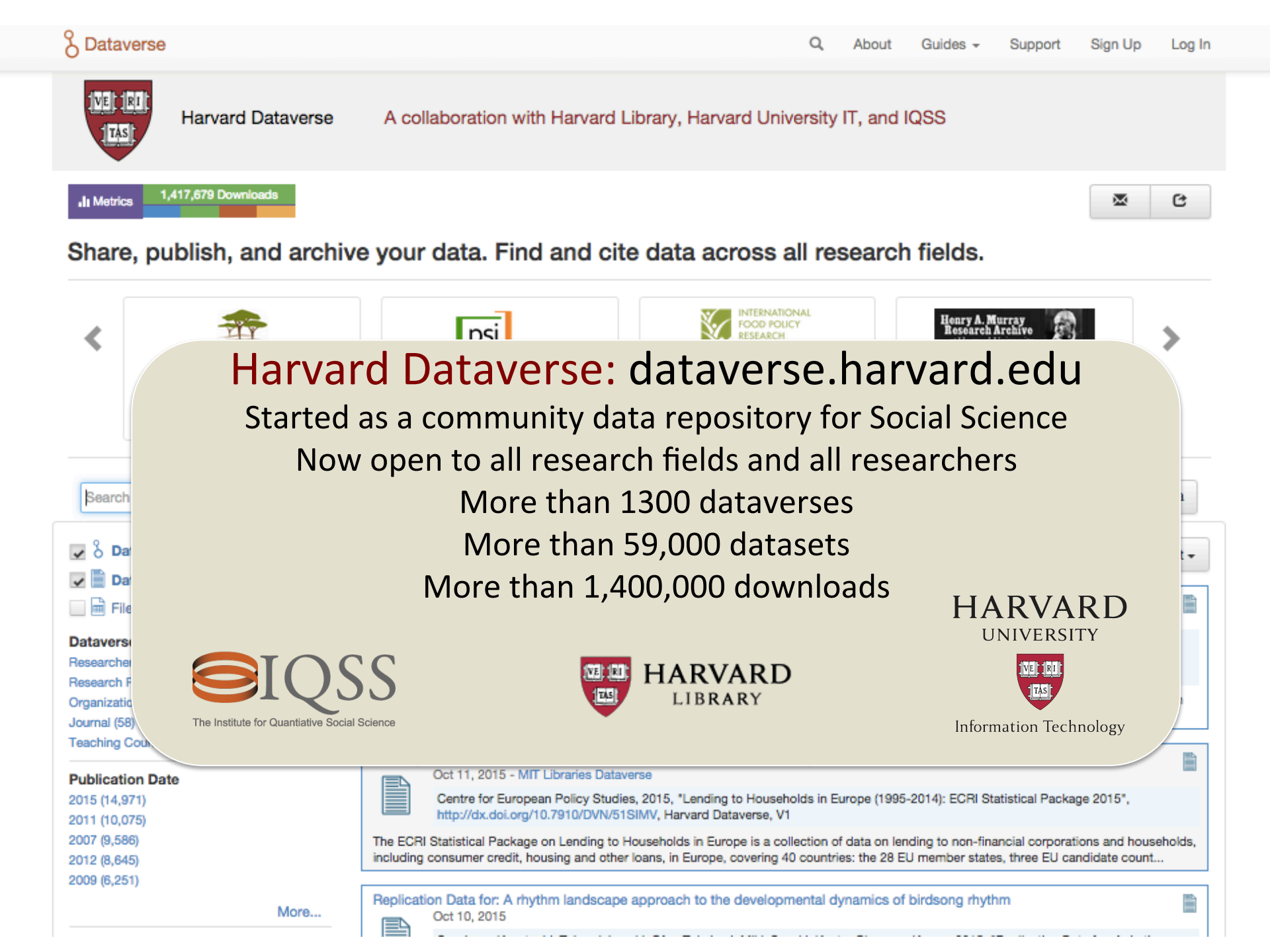
dataverse.org

Open-source software developed at Harvard's IQSS since 2006

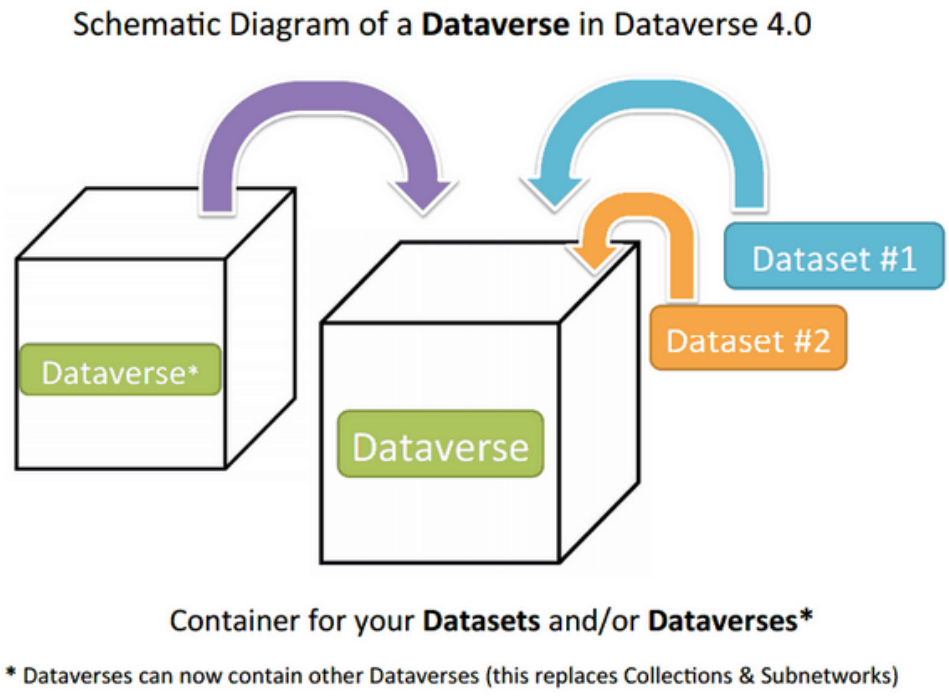
Installed in 12 sites world wide

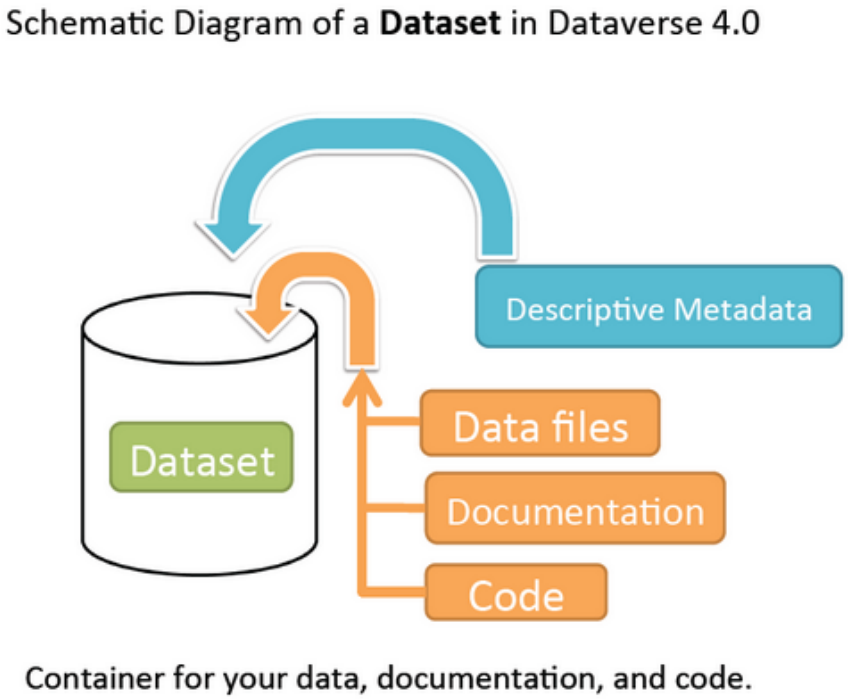
Serving 100s of universities and organizations




Dataverse
About
Guides
Support
Sign Up
Log In
Harvard Dataverse
A collaboration with Harvard Library, Harvard University IT, and IQSS
Metrics
1,417,679 Downloads
Share, publish, and archive your data. Find and cite data across all research fields.
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
IQSS
The Institute for Quantiative Social Science
HARVARD
LIBRARY
HARVARD
UNIVERSITY
Information Technology
Publication Date
2015 (14,971)
2011 (10,075)
2007 (9,586)
2012 (8,645)
2009 (6,251)
More...
Oct 11, 2015 - MIT Libraries Dataverse
Centre for European Policy Studies, 2015, "Lending to Households in Europe (1995-2014): ECRI Statistical Package 2015", http://dx.doi.org/10.7910/DVN/51SIMV, Harvard Dataverse, V1
The ECRI Statistical Package on Lending to Households in Europe is a collection of data on lending to non-financial corporations and households, including consumer credit, housing and other loans, in Europe, covering 40 countries: the 28 EU member states, three EU candidate count...
Replication Data for: A rhythm landscape approach to the developmental dynamics of birdsong rhythm
Oct 10, 2015

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

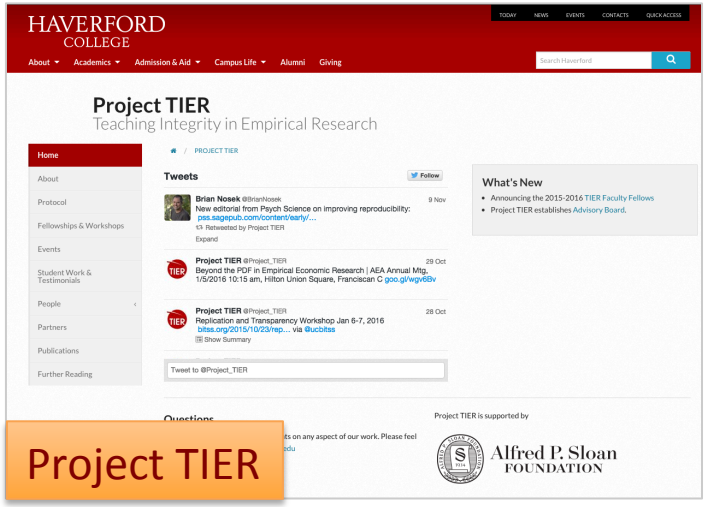
Journal Systems (Open Journal System, ScholarOne); Open Science Framework
Data Analysis (TwoRavens); Spatial Viz (WorldMap); Preservation systems (Archivematica)

Impact on the Social Science research community and on the World

Antislavery petitions data

Election Data Archive

The Dataverse Project



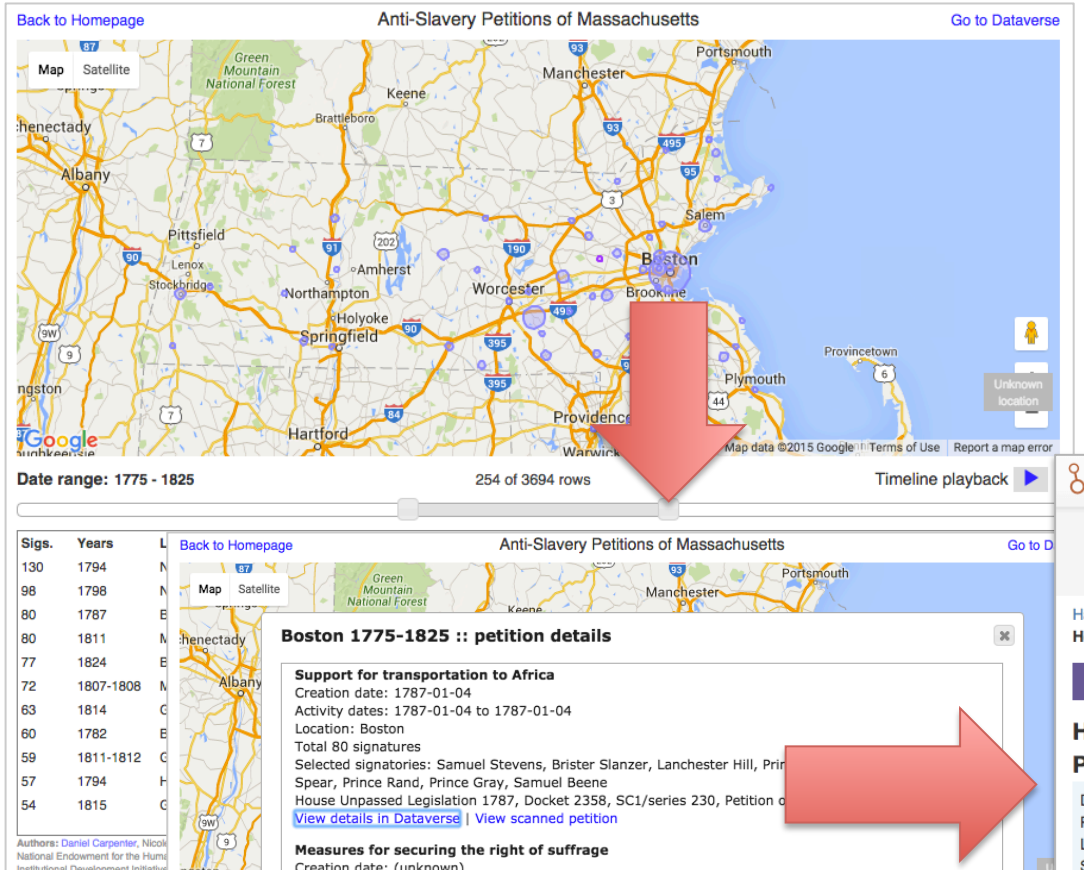
Project TIER

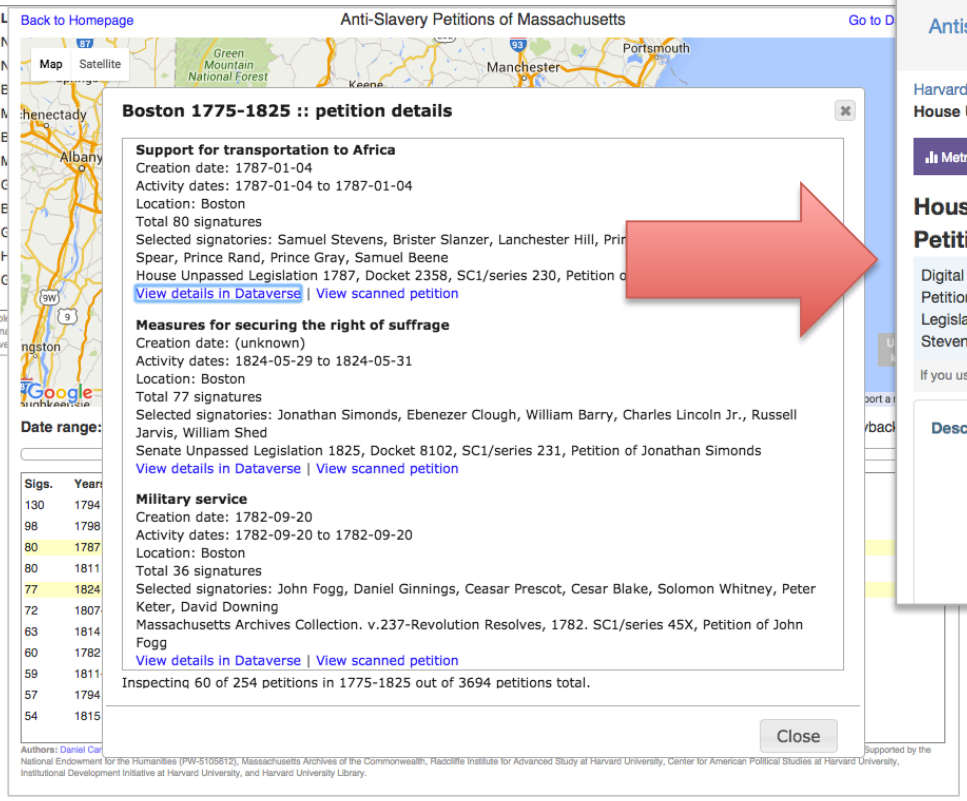
Boston Area Research Initiative

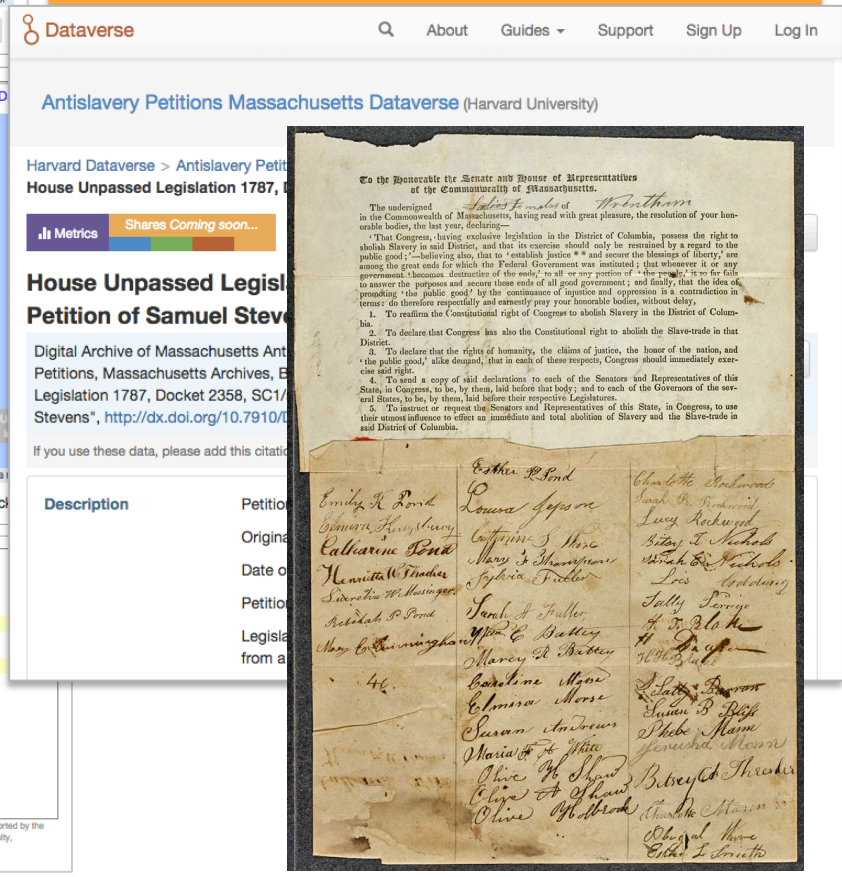
Antislavery Petition Data

Daniel Carpenter, Garth Griffin (Harvard University)

3,500 antislavery and antisegregation petitions sent to Massachusetts from 1600s to 1870





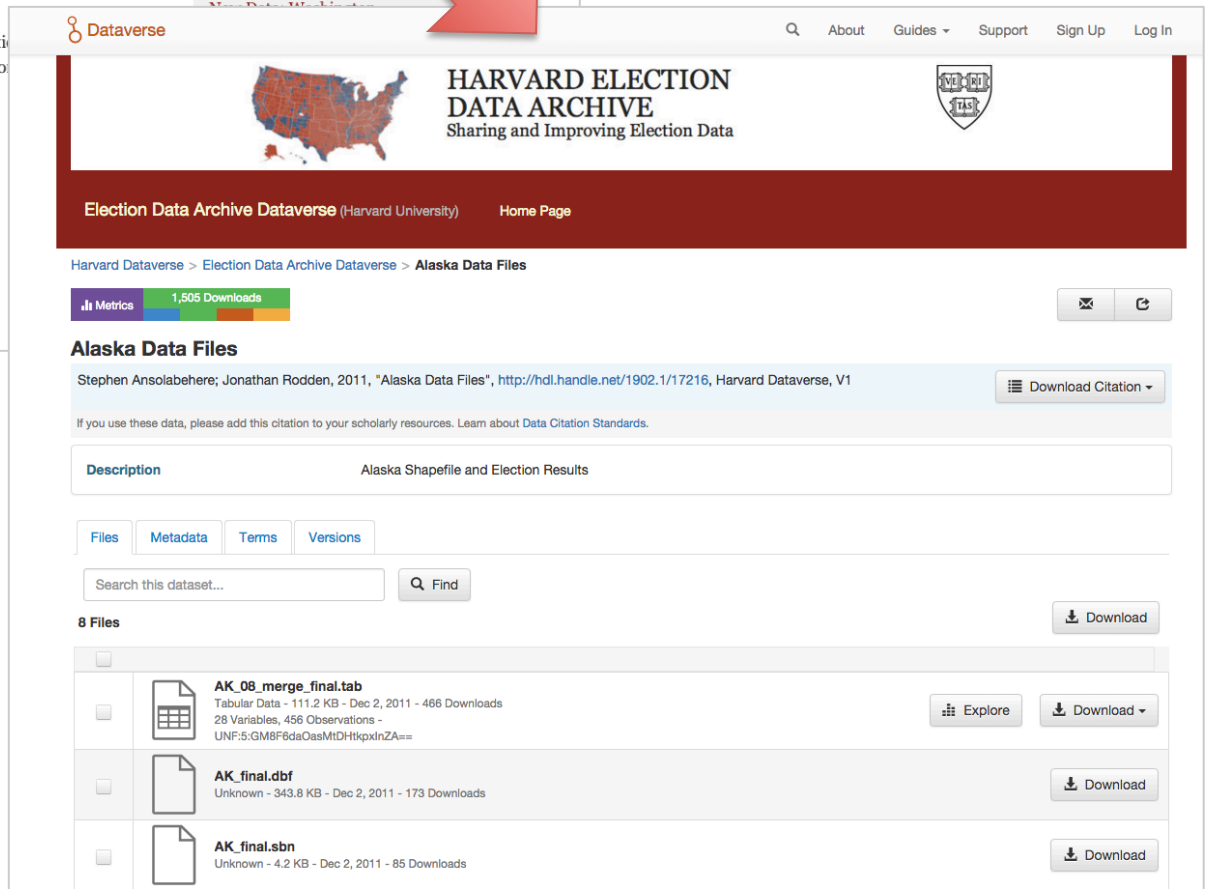


Election Data Archive

Steve Ansholabehere (Harvard), Jonathan Rodden (Stanford)

A collaborative archive to share election results, voting behavior, and electoral politics.

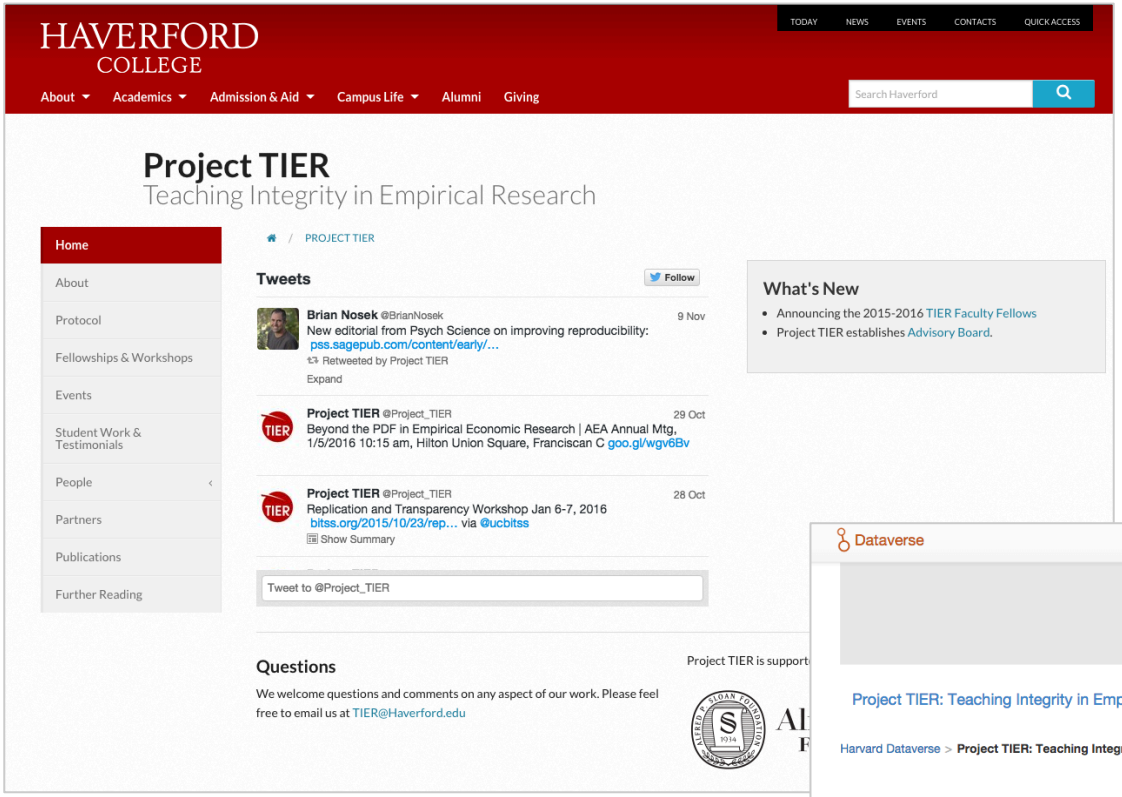
Alaska electoral data:
1,500 data downloads

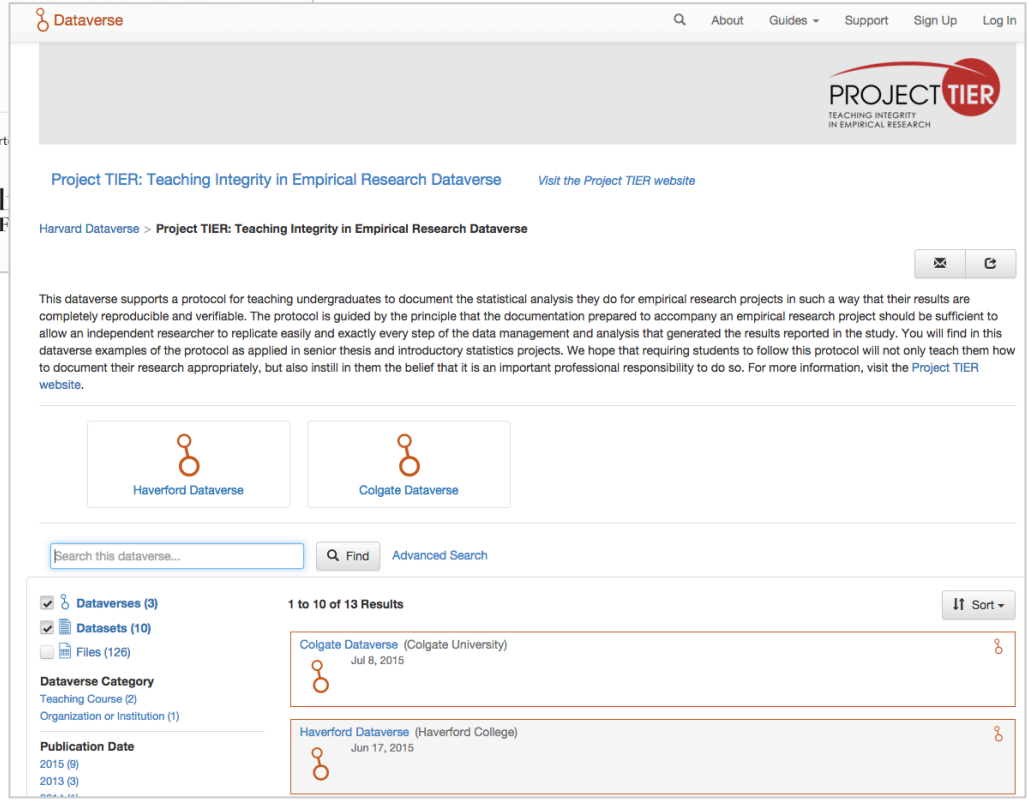


Project TIER

Richard Ball, Norm Medeiros (Haverford College)

Teaching empirical research with reproducibility in mind to future scholars





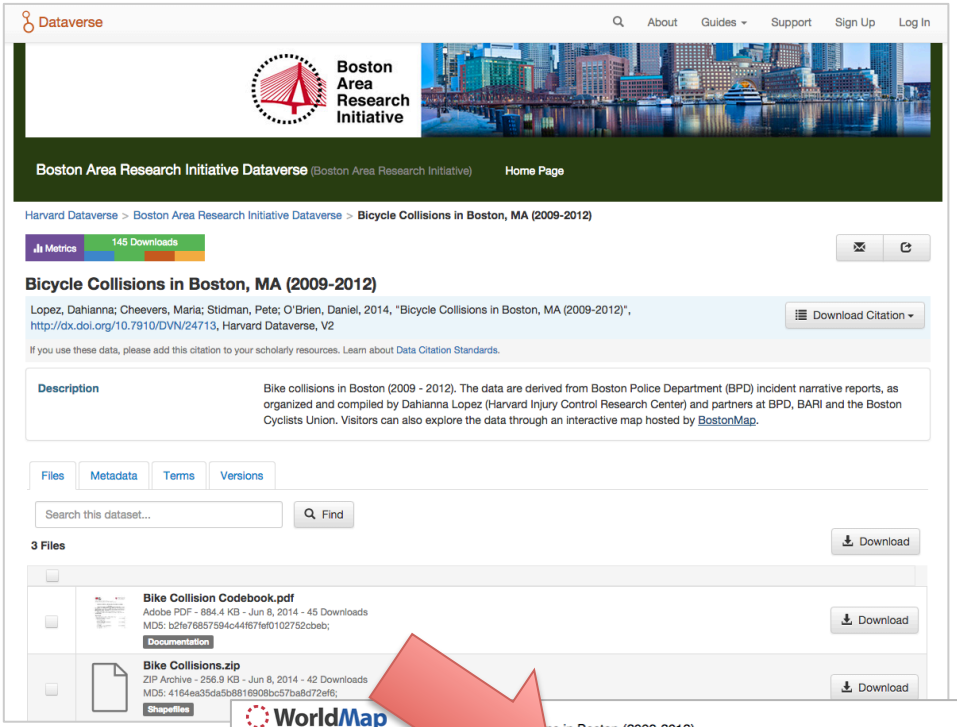
Provides a protocol to document all steps in data management and analysis:

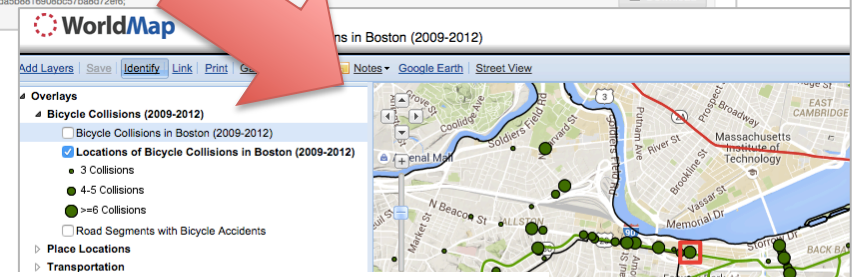
- Data Files
- Metadata Files
- Computing Command Files
- Readme File

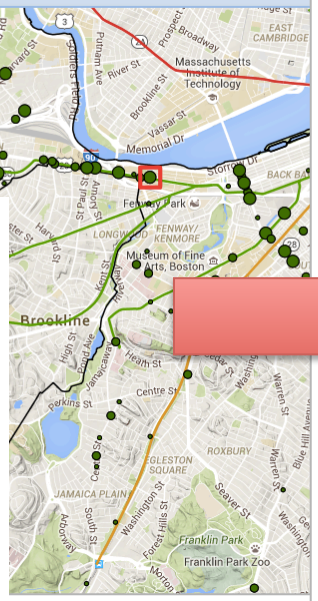
Boston Area Research Initiative

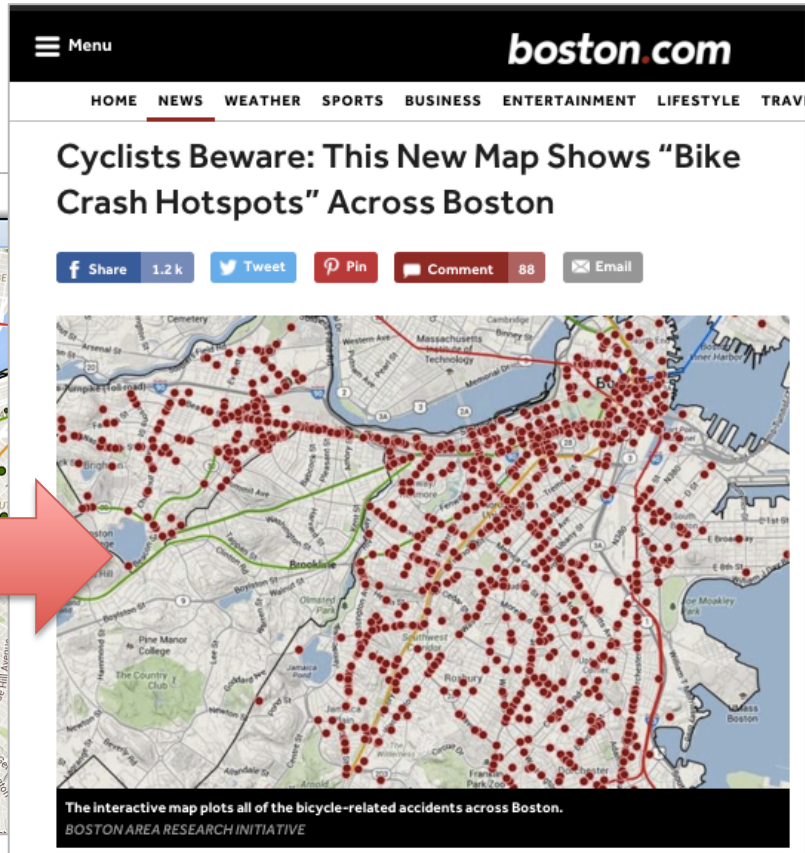
Daniel O'Brien (Northeastern), Robert Sampson, Christopher Winship (Harvard)

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



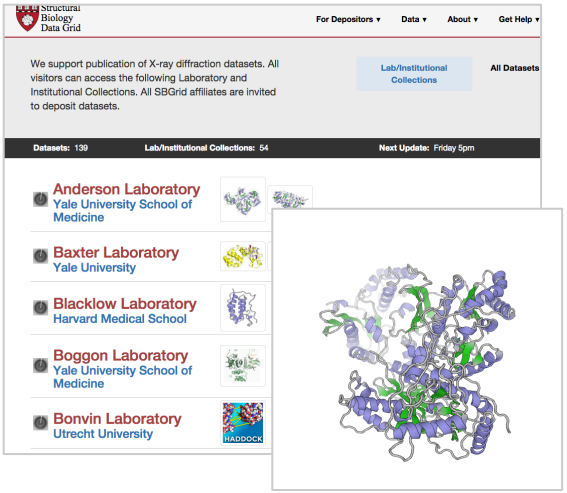


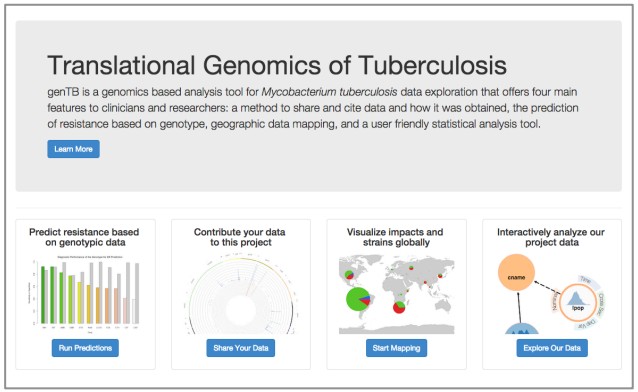


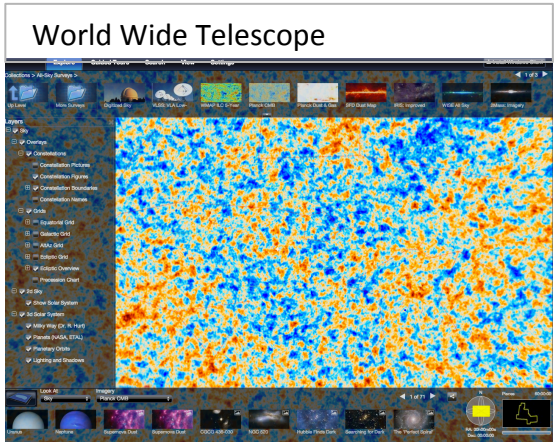


- Dataset of Bicycle Collisions in Boston (in collaboration with Boston Police, Harvard School of Public Health, and Cyclists Union)
- Data visualization with WorldMap

Future impact on other research communities: Biomedical and Astronomy







Structural Biology Data

OME-TIFF Files

Tuberculosis Genomics Data

R Data Frames

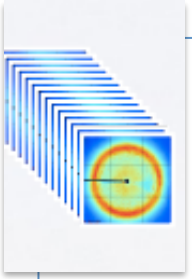
Astronomy Data

FITS Files

The Dataverse Project

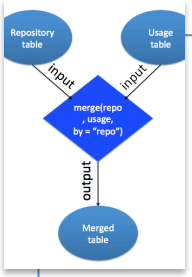
- Data archival
- Conversion to standard formats
- Extraction of file-level metadata

Current Collaborations: Addressing the Next Challenges in Data Sharing



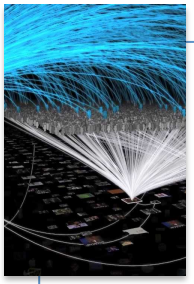
Structural Biology Grid Data Repository (Sliz, HMS, Crosas, IQSS)

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST



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

NSF



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

ALFRED P. SLOAN FOUNDATION 1934

Privacy Tools to share sensitive data (SEAS, Berkman Center, Privacy Lab, IQSS, MIT)

NSF

ALFRED P. SLOAN FOUNDATION 1934

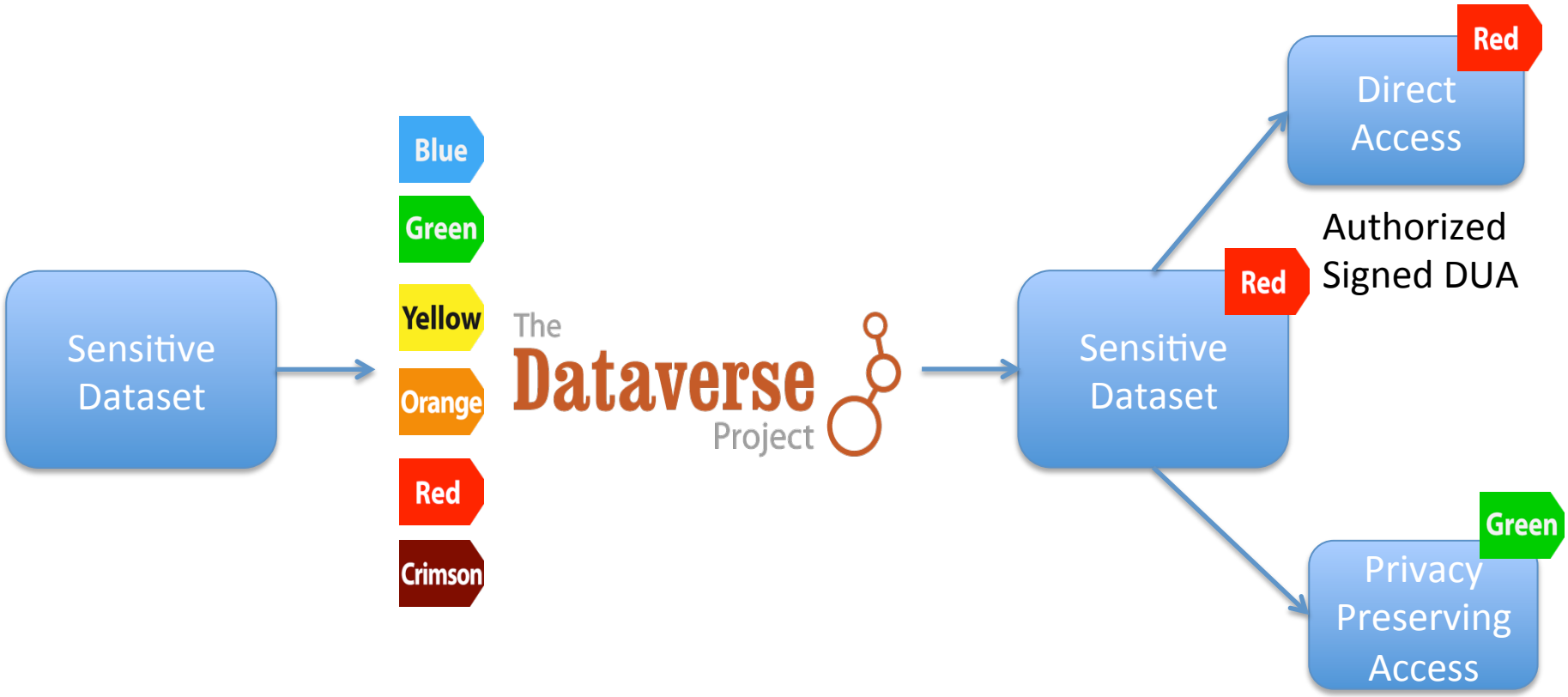
Sharing Sensitive Data with Confidence: DataTags System

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

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



DATA PRIVACY LAB

IQSS
The Institute for Quantitative Social Science

BERKMAN CENTER FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

CRCS Center for Research on Computation and Society
at Harvard School of Engineering and Applied Sciences

Program on Information Science
MIT Libraries

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

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

THANKS