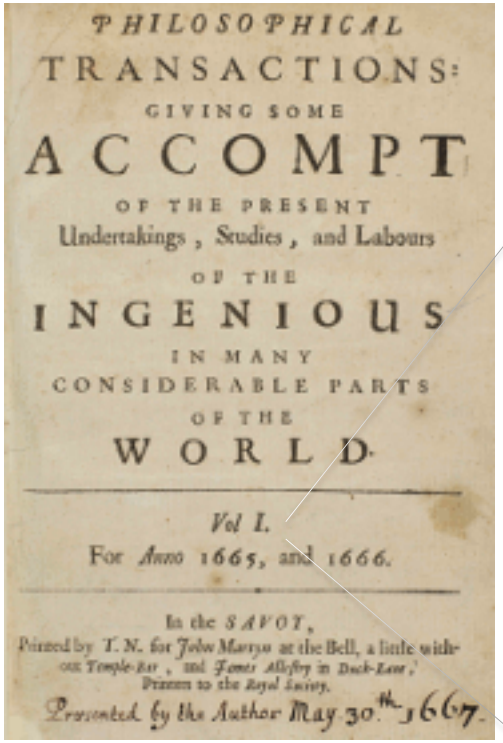
Mercè Crosas, Ph.D.
Chief Data Science and Technology Officer
Institute for Quantitative Social Science
Harvard University
@mercecrosas

DATA PUBLISHING

Scholarly Publication Then



PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane, Printers to the Royal Society.
Presented by the Author May 30th 1667.

No digital data – but data are described in detail in the article

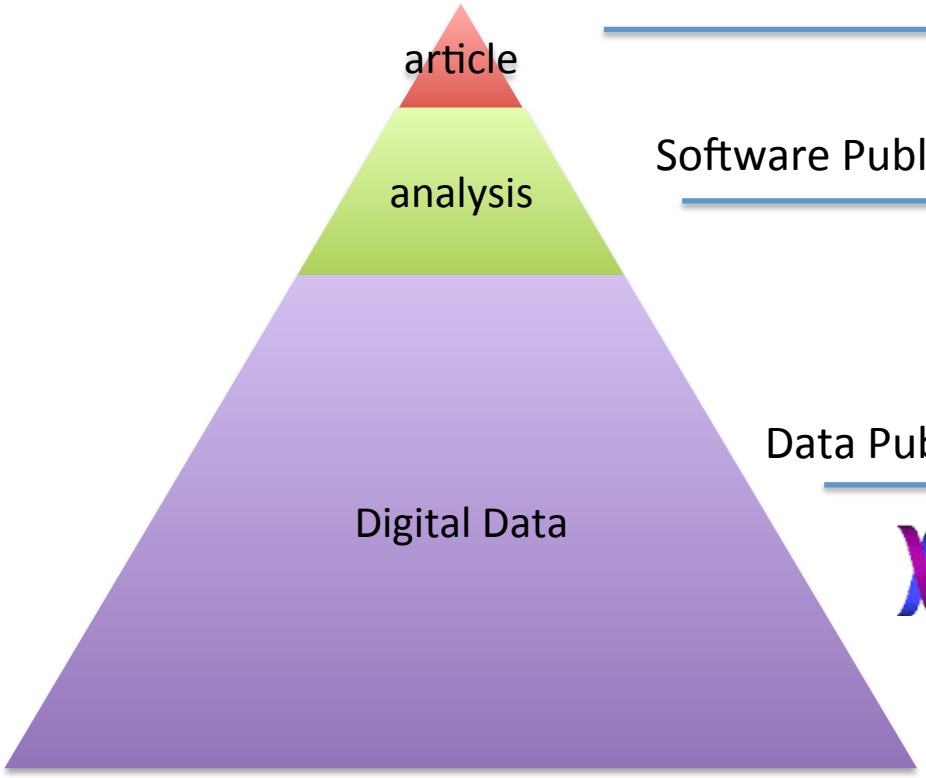
Observations
About Shining Worms *in Oysters.*

These Observations occur in the *French journal* of *April* 12. 1666. in two letters, written by M. *Auzout* to M. *Dela Voye*; whereof the substance may be reduced to the following particulars.

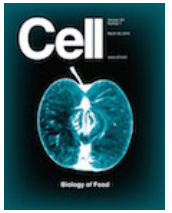
1. That M. *Dela Voye* having observed, as he thought, some

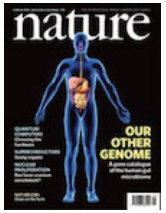
350 years ago, the first issue of Philosophical Transactions was published by the Royal Society, under the motto ***"Nullius in verba"*** (or "*Take nobody's word for it*")

Scholarly Publication Now



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.

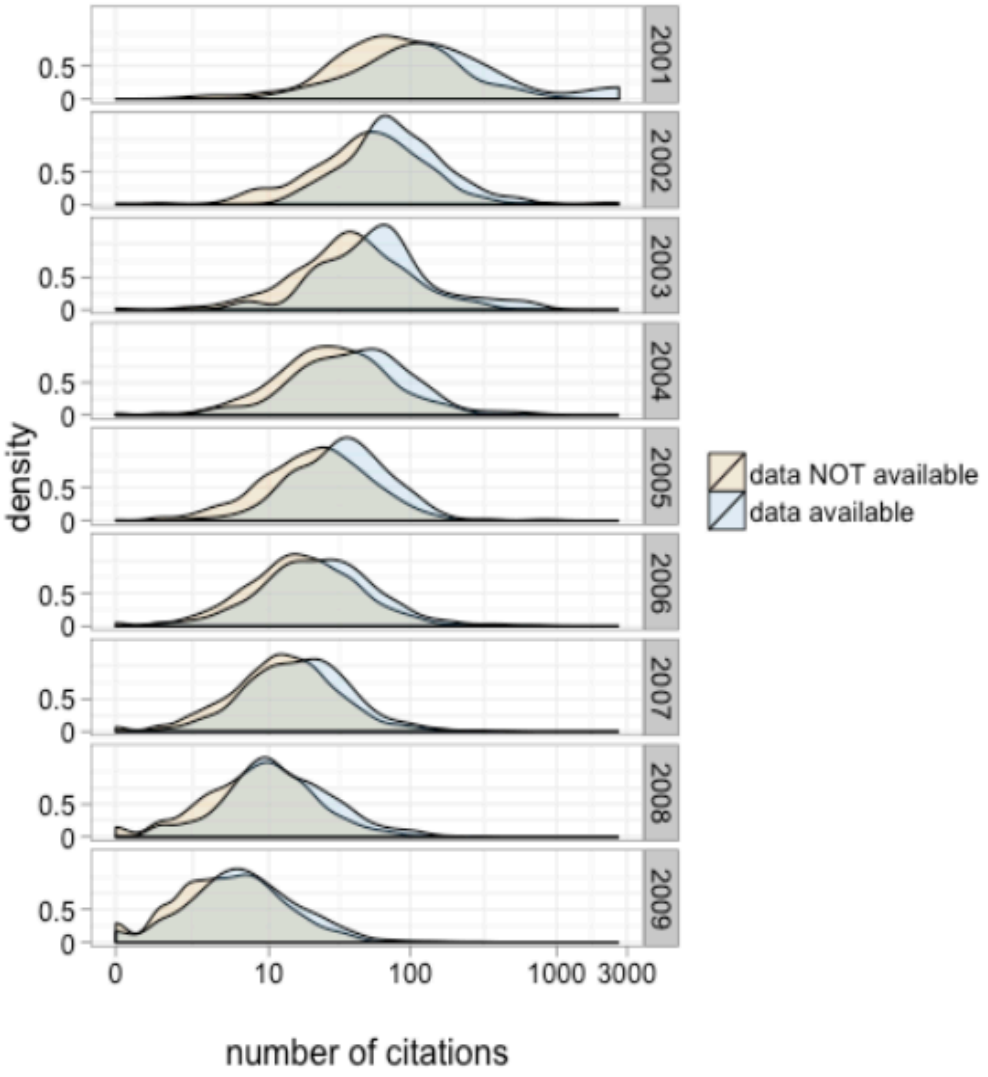
Data publishing: It's good for you and good for the world

Sharing Data Increases Citations

From 10,555 studies with gene expression microarray data:

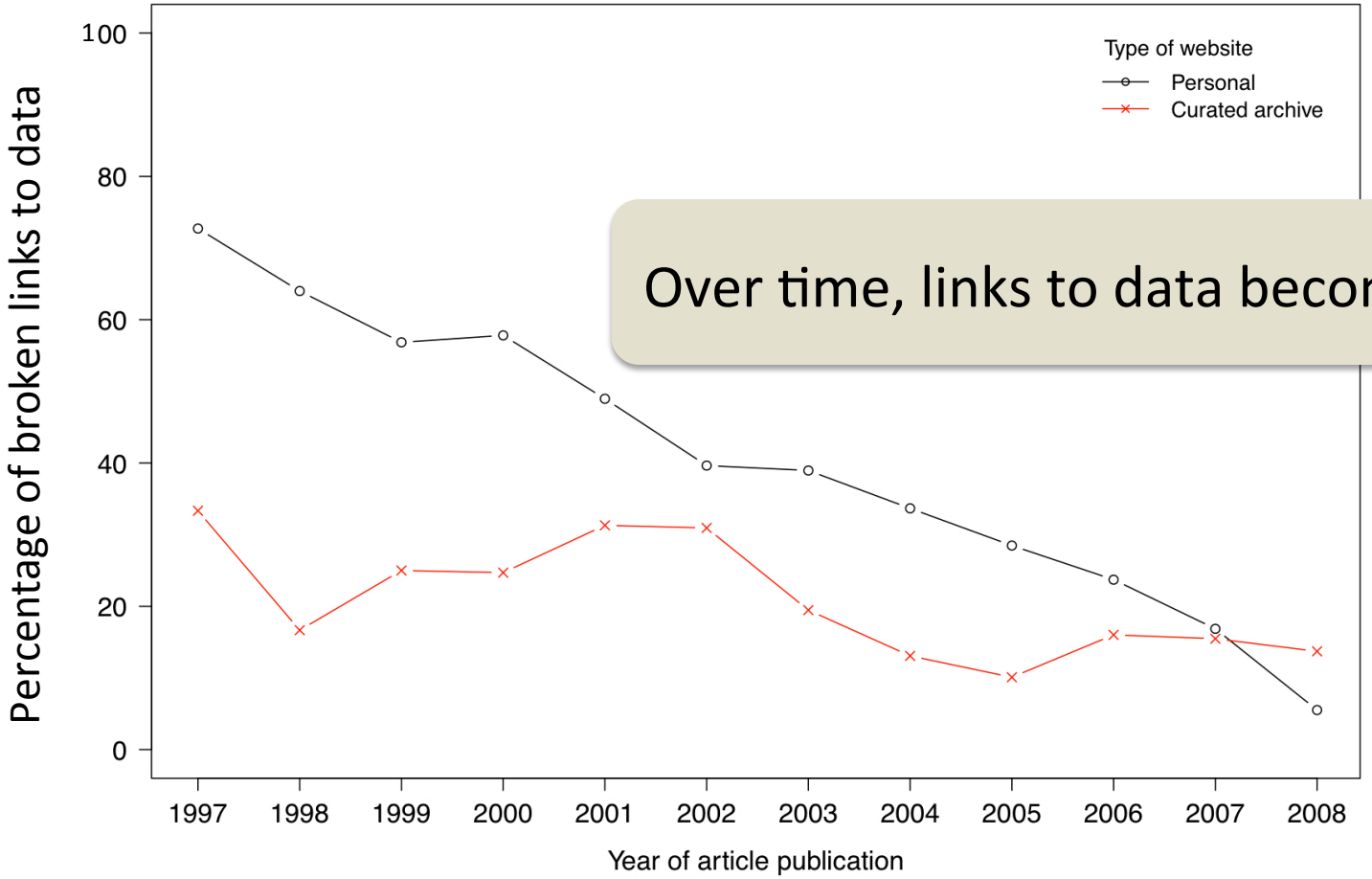
- Studies that shared data received 9% **more citations**

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



Piwowar and Vision (2013), Data reuse and the open data citation advantage. PeerJ 1:e175; DOI 10.7717/peerj.175

Long-Term Accessibility must be Considered

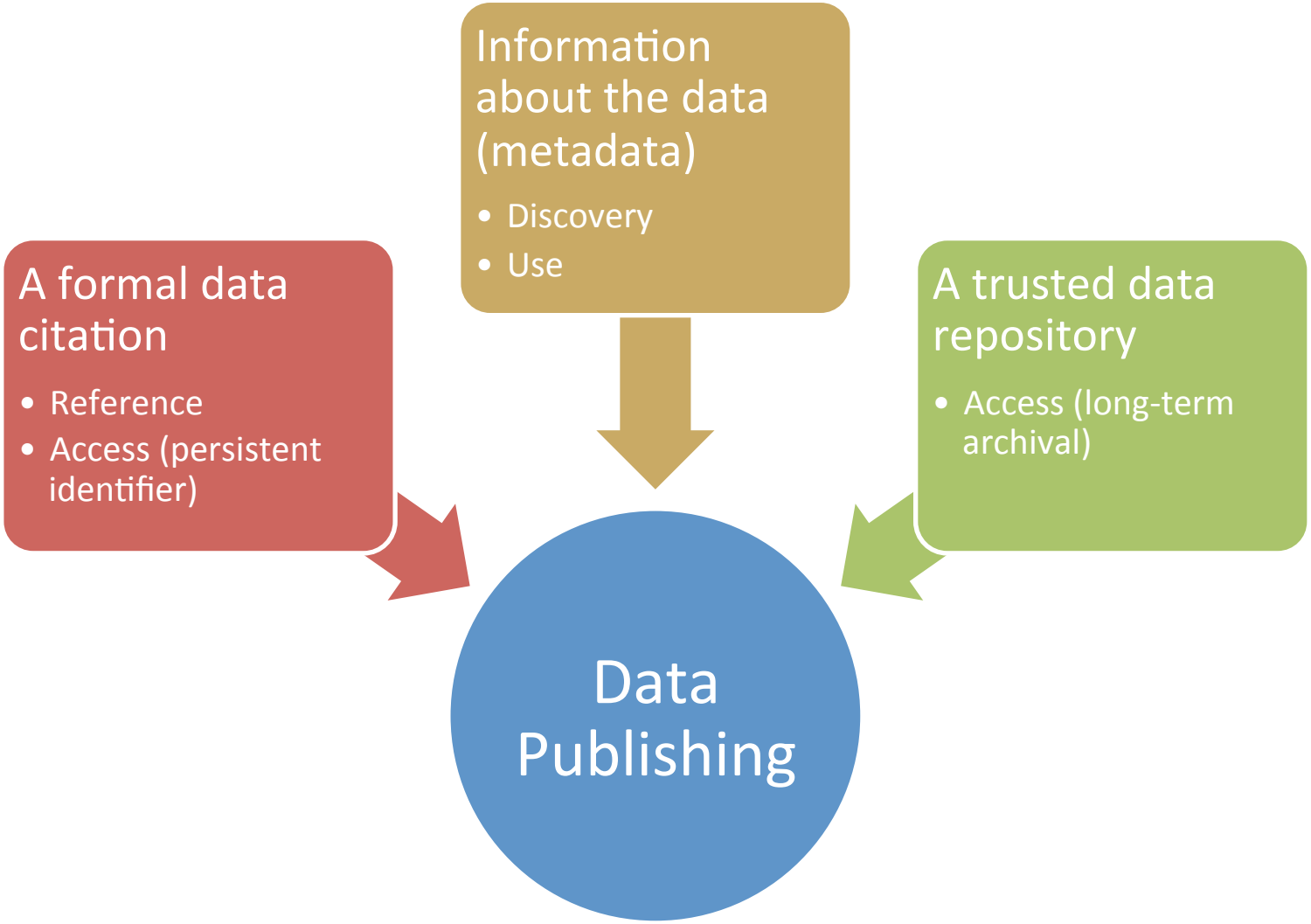


Over time, links to data become invalid

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

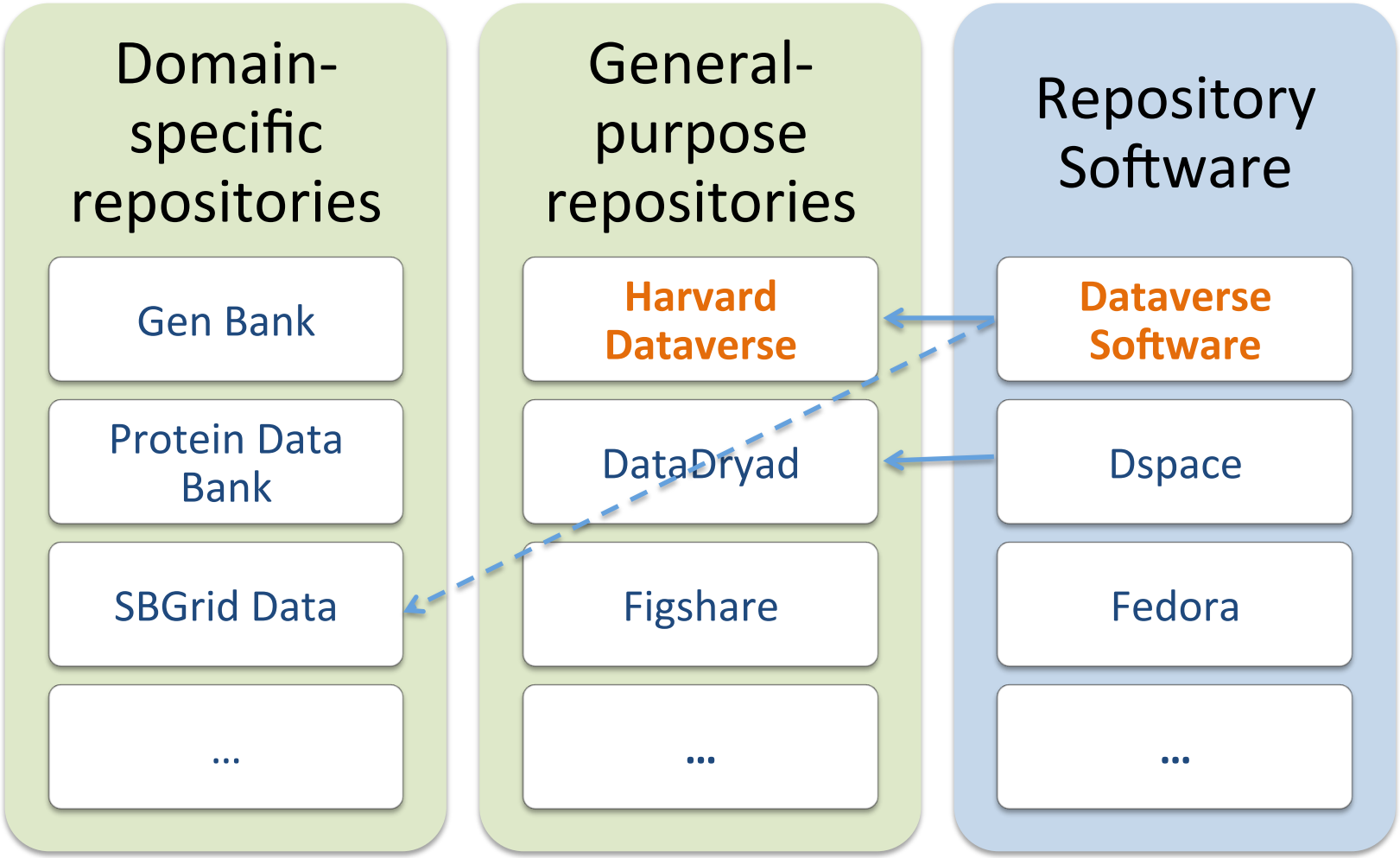


Data Citation Principles

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

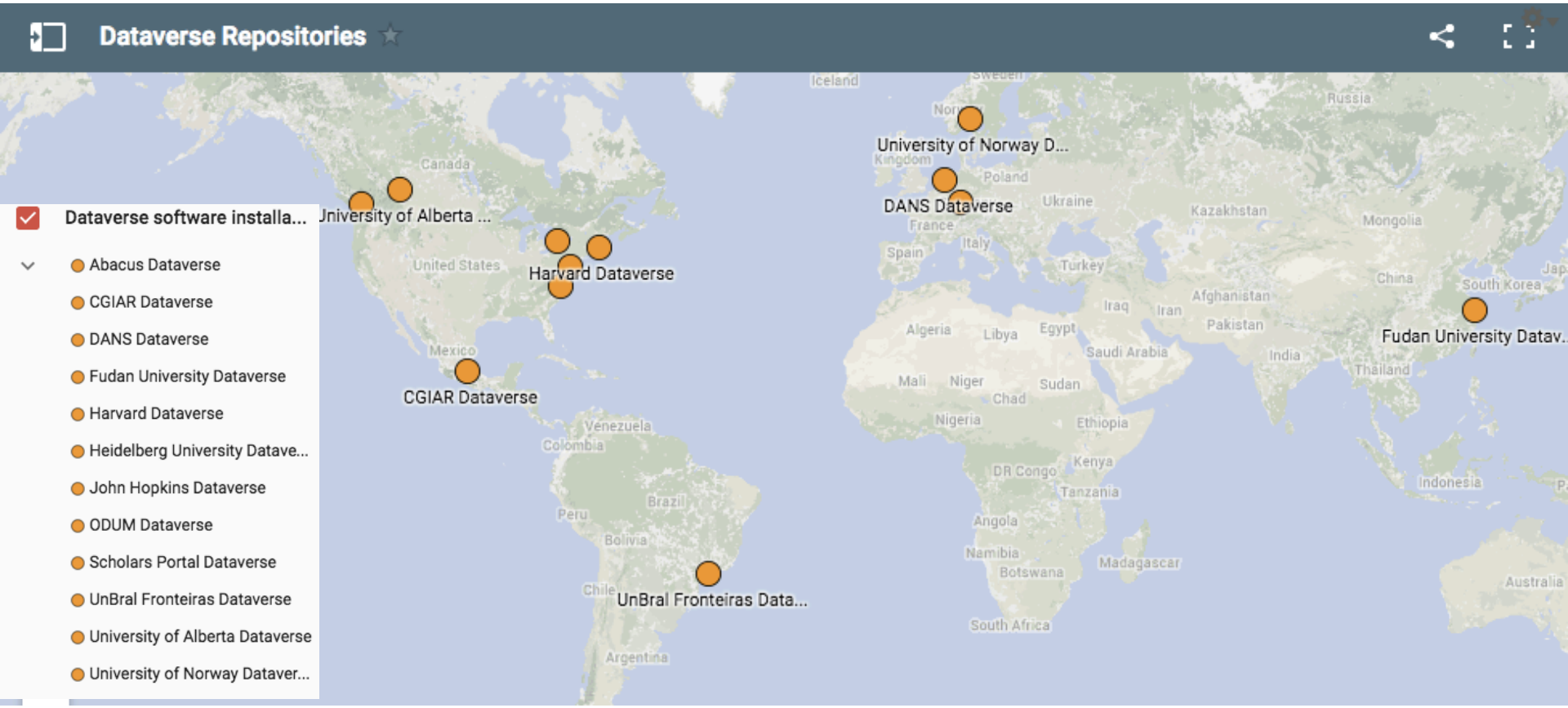
Data Repositories vs Repository Software

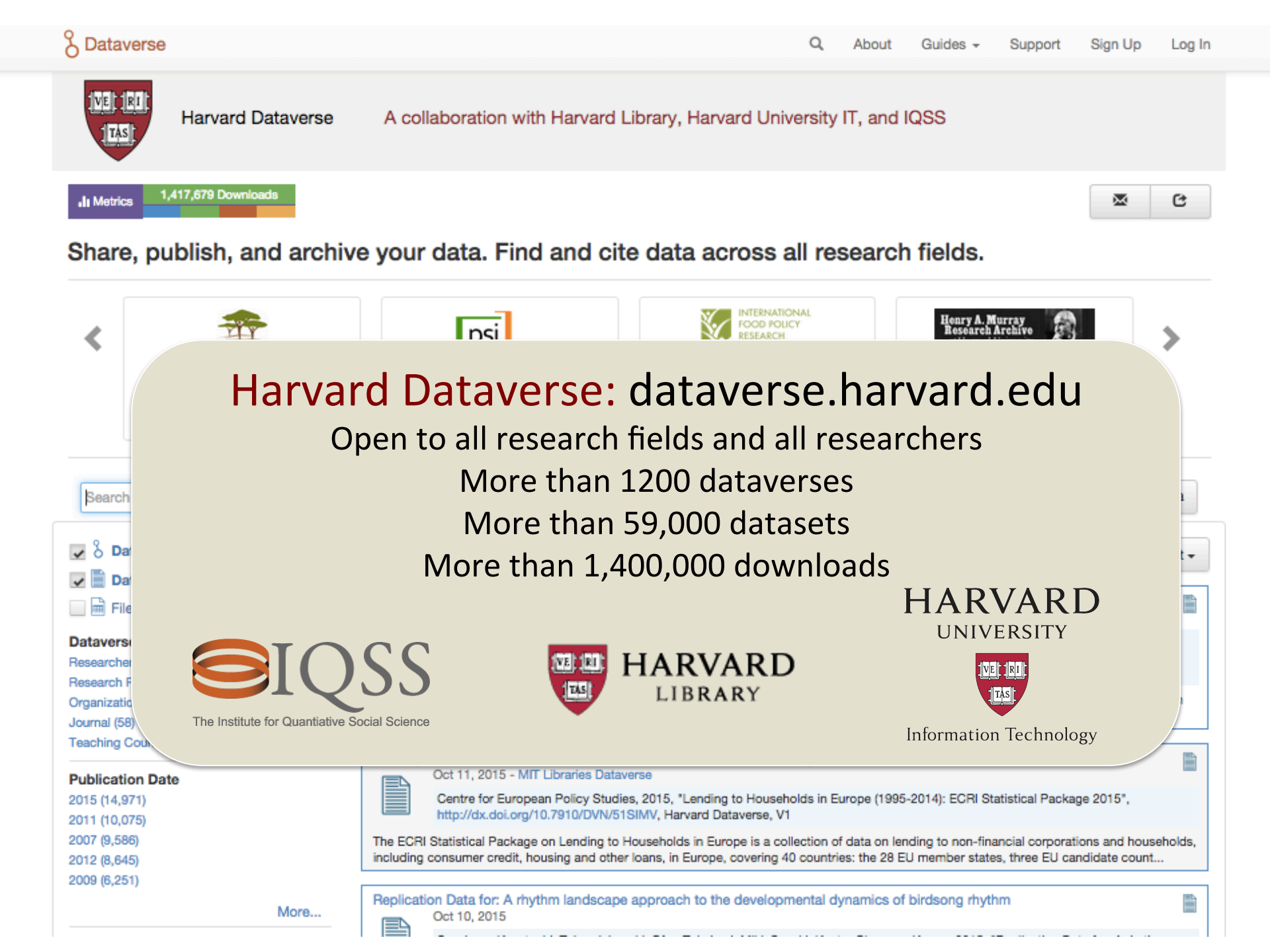


The Dataverse Project

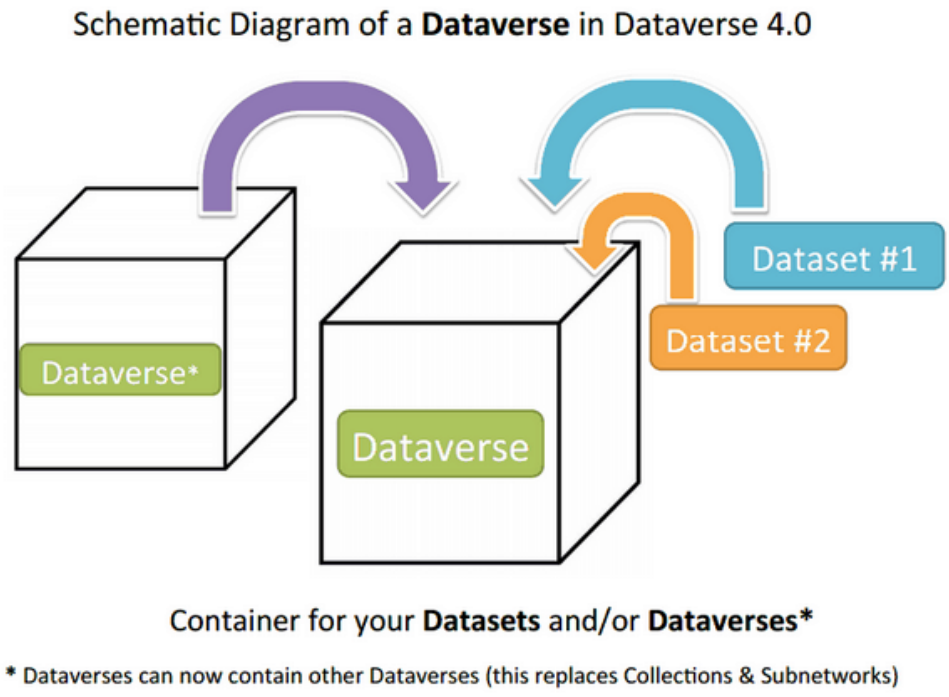
dataverse.org

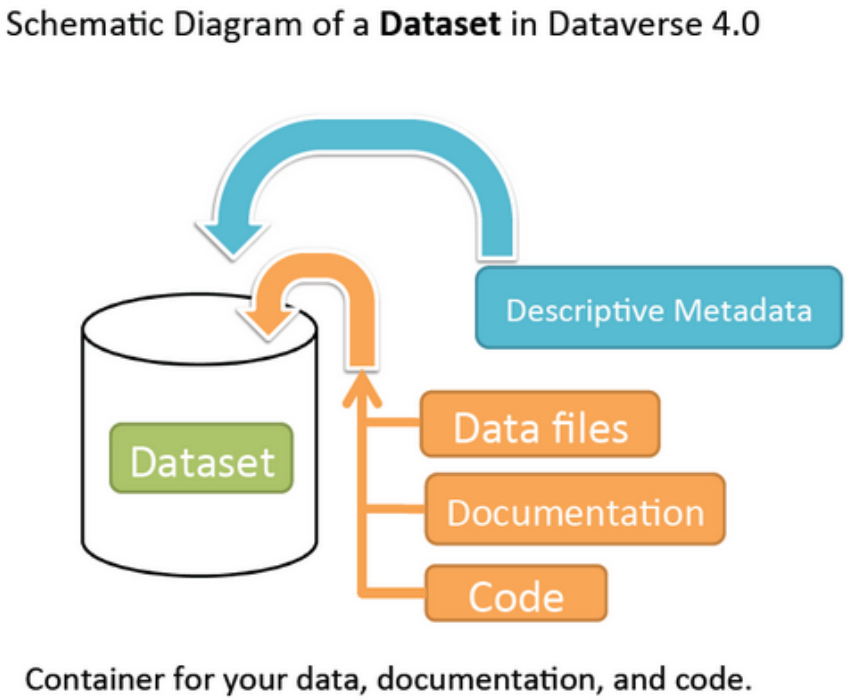
Open-source software developed at Harvard' IQSS since 2006
Installed in 12 sites world wide
Serving 100s of universities and organizations




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
Open to all research fields and all researchers
More than 1200 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

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

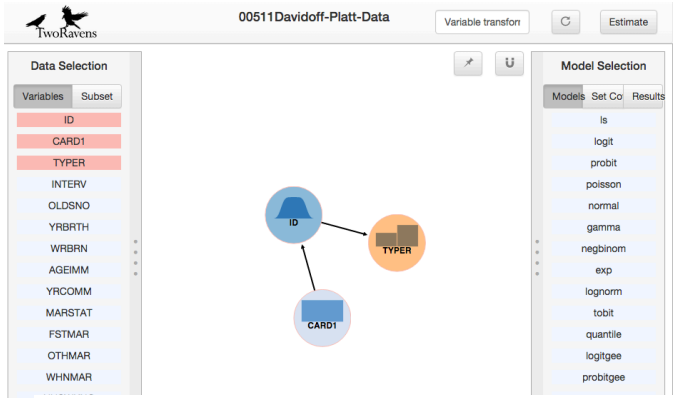
- Versioning
- Conversion of tabular data files to standard format
- Automatic extraction of file metadata (R, STATA, SPSS, XSD, FITS)

Interoperability through APIS

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

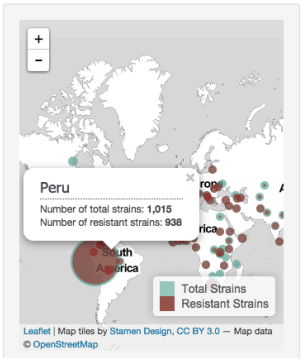
What you can do with file-level metadata and APIs Now

Anti-slavery petitions data

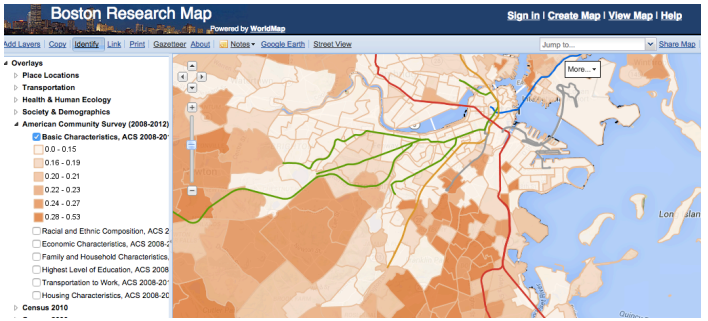


Statistical analysis with TwoRavens

The Dataverse Project



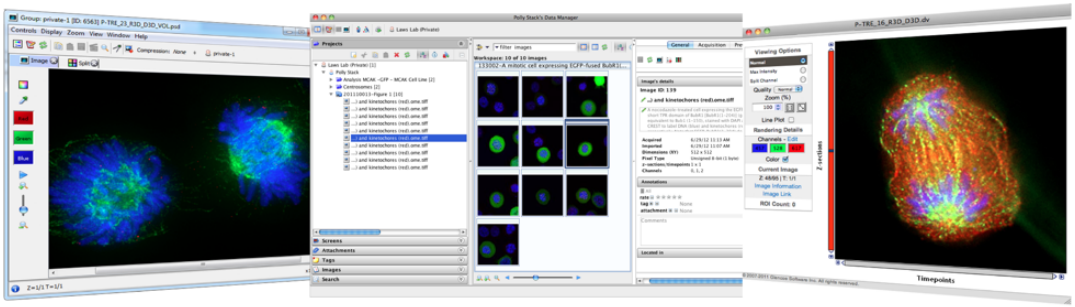
Tuberculosis Genomics data



Boston Area Research Initiative data visualization in WorldMap

What you will be able to do with Image Data in Dataverse

OME-TIFF Files



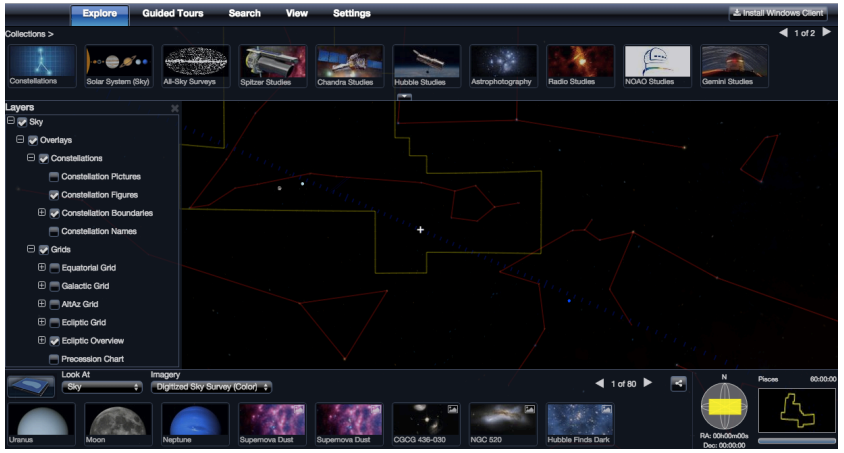
OMERO

Conversion to standard formats

+

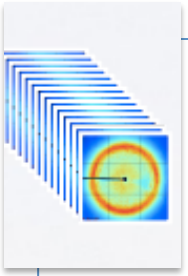
Extraction of file-level metadata

FITS Files



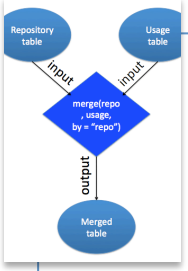
WORLD WIDE TELESCOPE

Current Collaborations



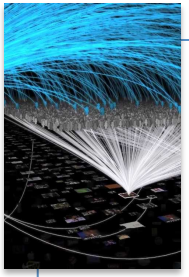
SB Grid Data Repository (HMS, IQSS)

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST



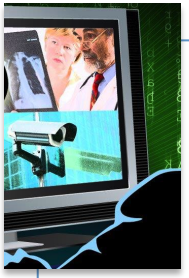
Data Provenance (SEAS, IQSS)

NSF



Social Science Big Data (IQSS)

ALFRED P. SLOAN FOUNDATION 1934



Privacy Tools to share sensitive data (SEAS, Berkman, 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

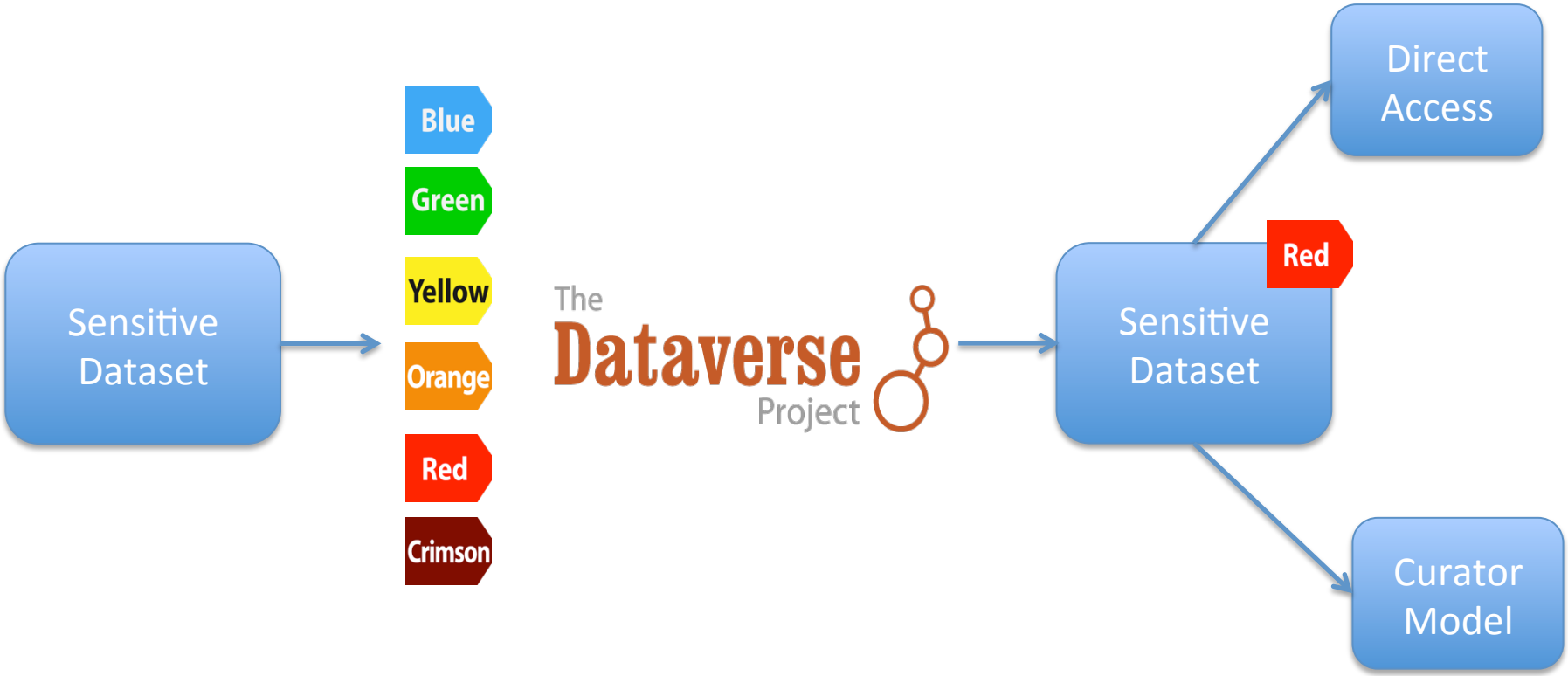
DataTag: A set of security features and access requirements for file handling

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.

Sweeney L, Crosas M, Bar-Sinai M. Sharing Sensitive Data with Confidence: The DataTags System. Technology Science. 2015101601. October 16, 2015. http://techscience.org/a/2015101601

Data Publishing Workflow for Sensitive Data



DATA PRIVACY LAB

IQSS
The Institute for Quantitative Social Science

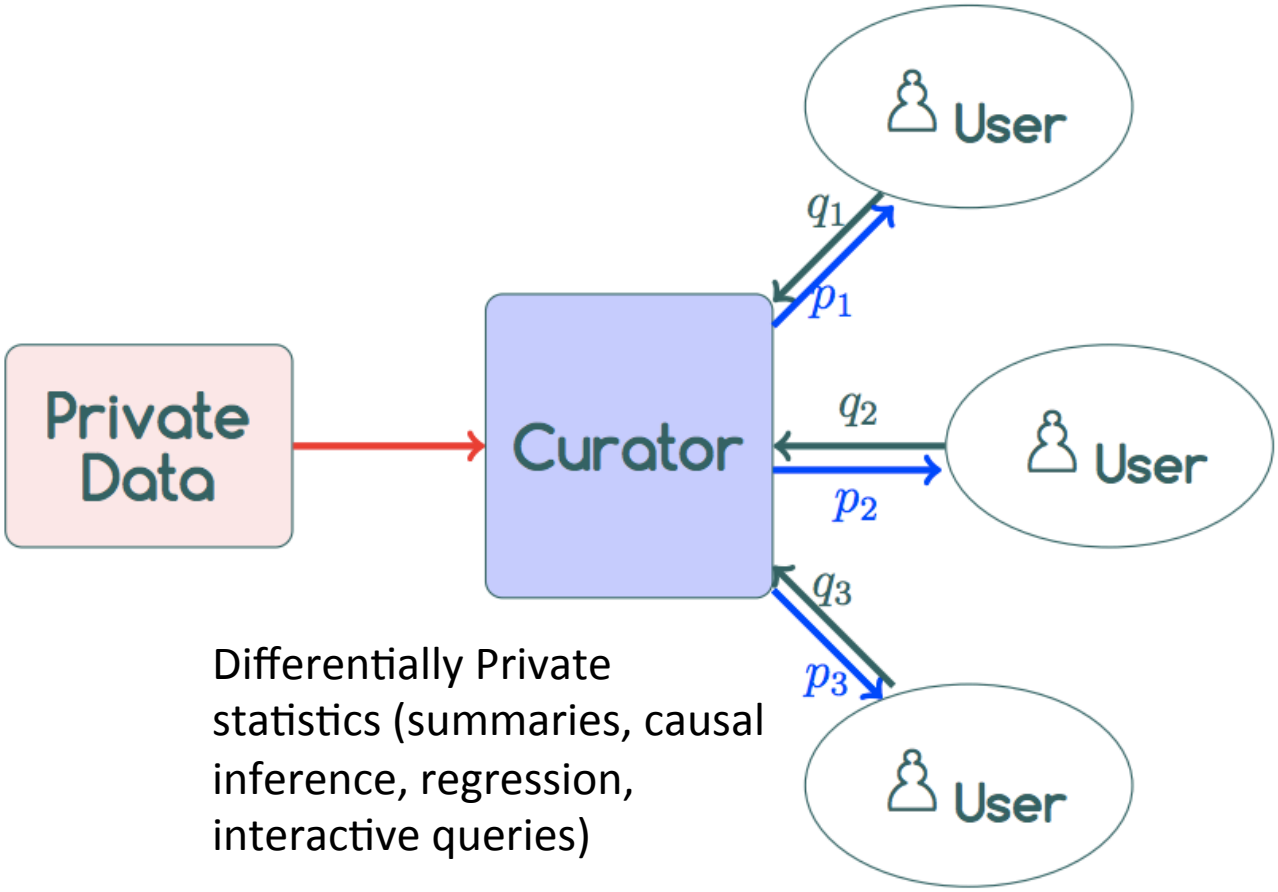
BERKMAN CENTER FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

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

Program on Information Science
MIT Libraries

Datatags.org

A Curator Model for Privacy-Preserving Analysis



Acknowledgement: Honaker, J. and Nissim, K., Data Privacy Tools Project

DEMO

https://beta.dataverse.org/custom/DifferentialPrivacyPrototype/

Acknowledgement: Latanya Sweeney, James, Honaker, Eleni Castro, Margo Seltzer, Piotrek Sliz, Christine Choirat, Garth Griffin, and the Dataverse team for graphics and slides

THANKS