DATA SHARING FOR BETTER SCIENCE

THE DATAVERSE PROJECT

Mercè Crosas, Institute for Quantitative Social Science, Harvard University

@mercecrosas

MAX PLANCK INSTITUTE FOR RADIOASTRONOMY, SEPTEMBER 12, 2017

THIS TALK

- **Importance of Data Sharing**
  - **Reproducibility** to verify science
  - **Reuse** to advance science and evidence-based policy
- **Enabling Data Sharing**
  - **Data Policies** from journals and funding agencies
  - **Data Citation** to find datasets, give credit to data authors
  - **Data Repositories** as publishers of data

DATA SHARING, DATA PUBLISHING

Data sharing is "the release of research data, associated metadata, accompanying documentation, and software code for re-use and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way."

Data Publishing Group, 201 5

RESEARCH DATA ALLIANCE

Since the Beginning of Modern Science ...

NULLIUS IN VERBA

"TAKE NOBODY'S WORD FOR IT"

(motto of the Royal Society, founded in 1660,
launched first scientific journal in 1665)

AUG 22 2017

LEAVE A COMMENT

BY JOHN BORGHI

BEST PRACTICES

WHAT WE TALK ABOUT WHEN WE TALK ABOUT REPRODUCIBILITY

Campbell's Soup Cans (1962) by Andy Warhol. Created by replicating an existing object and then reproducing the process at least 32 times.

University of California Curation Center, DataPub blog, August 2017

REPRODUCIBILITY

REPRODUCIBILITY AND REPLICATION (BY THE NATIONAL SCIENCE FOUNDATION):

The ability of a researcher to duplicate the results of a prior study

... using the same materials and procedures used by the original investigator. **(reproducibility)**

... if the same procedures are followed but new data are collected. **(replication)**

EMPIRICAL, COMPUTATIONAL, AND STATISTICAL REPRODUCIBILITY (STODDEN, 2014):

Empirical: data and collection details are made freely available

Computational: code, software, hardware and implementations details are provided

Statistical: details on choice of statistics tests, model parameters are provided

REPRODUCIBILITY CRISIS?

TRUST, BUT VERIFY

The Economist | World politics | Business & finance | Economics | Science & technology | Cult

Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs to change itself

Oct 19th 2013 | From the print edition

Like 15k | Tweet 1,120

Science AAAS

Home | News | Journals | Topics | Careers

Science | Science Advances | Science Immunology | Science Robotics | Science Signaling | Science Translational Medicine

SHARE

0

19

RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration*,†

+ See all authors and affiliations

Science 28 Aug 2015:
Vol. 349, Issue 6251, aac4716
DOI: 10.1126/science.aac4716

Article | Figures & Data | Info & Metrics | eLetters | PDF

You are currently viewing the abstract. | View Full Text

6 (11%) out of 53 landmark cancer biology studies could be reproduced.

39 out of 100 psychology studies could be reproduced.

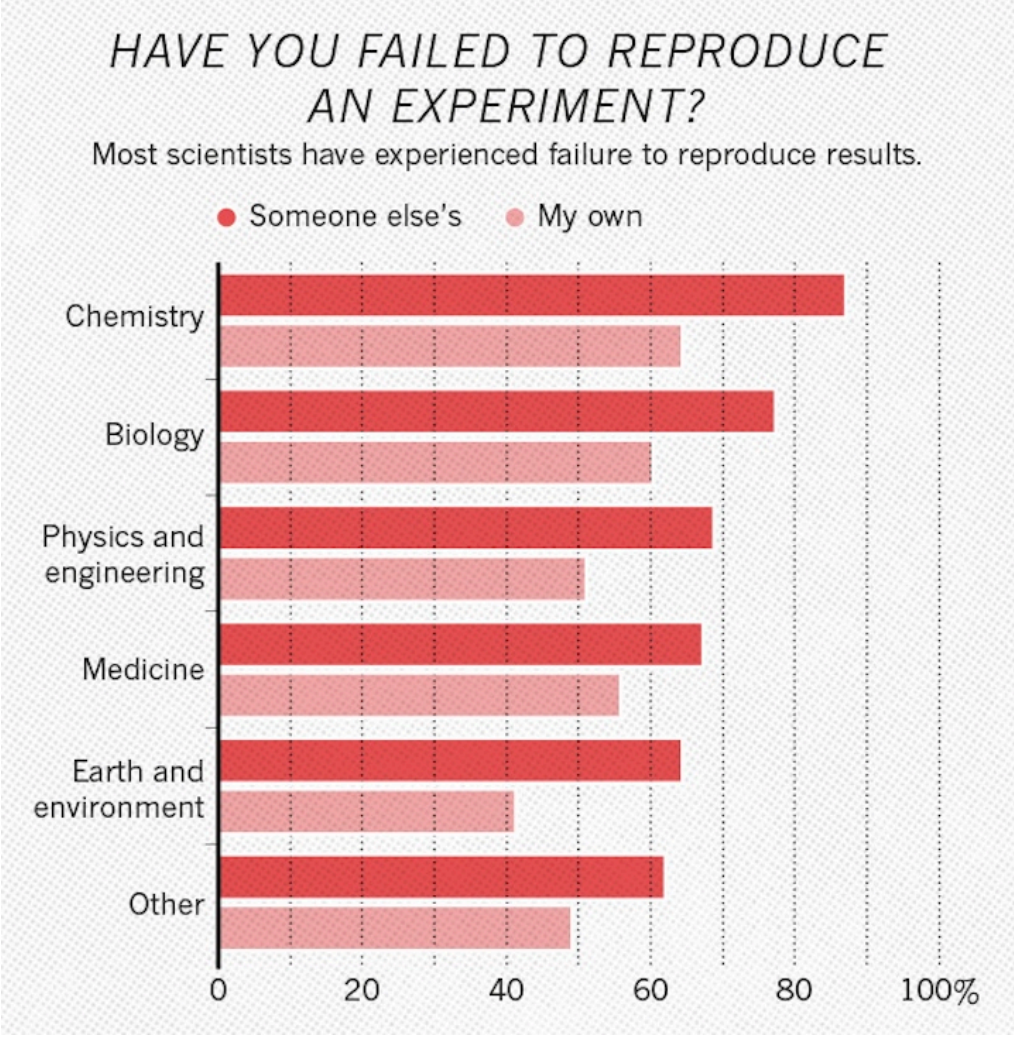
The Washington Post

Democracy Dies in Darkness

Speaking of Science

No, science's reproducibility problem is not limited to psychology

Washington Post, Joel Achenbach, August 28, 2015



***Nature*'s survey of 1,576 researchers:**

703 Biology

106 Chemistry

95 Earth and Environmental

203 Medicine

236 Physics and Engineering

233 Other

Nature, 2016, "1,500 scientists lift the lid on reproducibility", vol 533, Issue 734

Digital Science: reproducibility and visibility in Astronomy

J.E. Ruiz[1], L. Verdes-Montenegro[1], S. Sánchez[1], J.D. Santander-Vela[1], and J. Garrido[1]

[1] Instituto de Astrofísica de Andalucía – CSIC

Highlights of Spanish Astrophysics VII, Proceedings of the X Scientific Meeting of the Spanish Astronomical Society held on July 9 - 13, 2012, in Valencia, Spain. J. C. Guirado, L.M. Lara, V. Quilis, and J. Gorgas (eds.)

"In the Wf4Ever project we propose to improve the quality of science with metrics based on reproducibility and reuse, preserving decomposable thoroughly curated digital artefacts that enhances reproducibility and visibility of the experiment, as well as allowing more accurate mechanisms for credit attribution."

SHARING DATA, CODE, AND WORKFLOWS FACILITATES REPRODUCIBILITY AND REDUCES DUPLICATION

BUT DATA SHARING IS MORE THAN POSTING YOUR DATA IN YOUR WEBSITE

PLOS | ONE
TENTH ANNIVERSARY

OPEN ACCESS PEER-REVIEWED

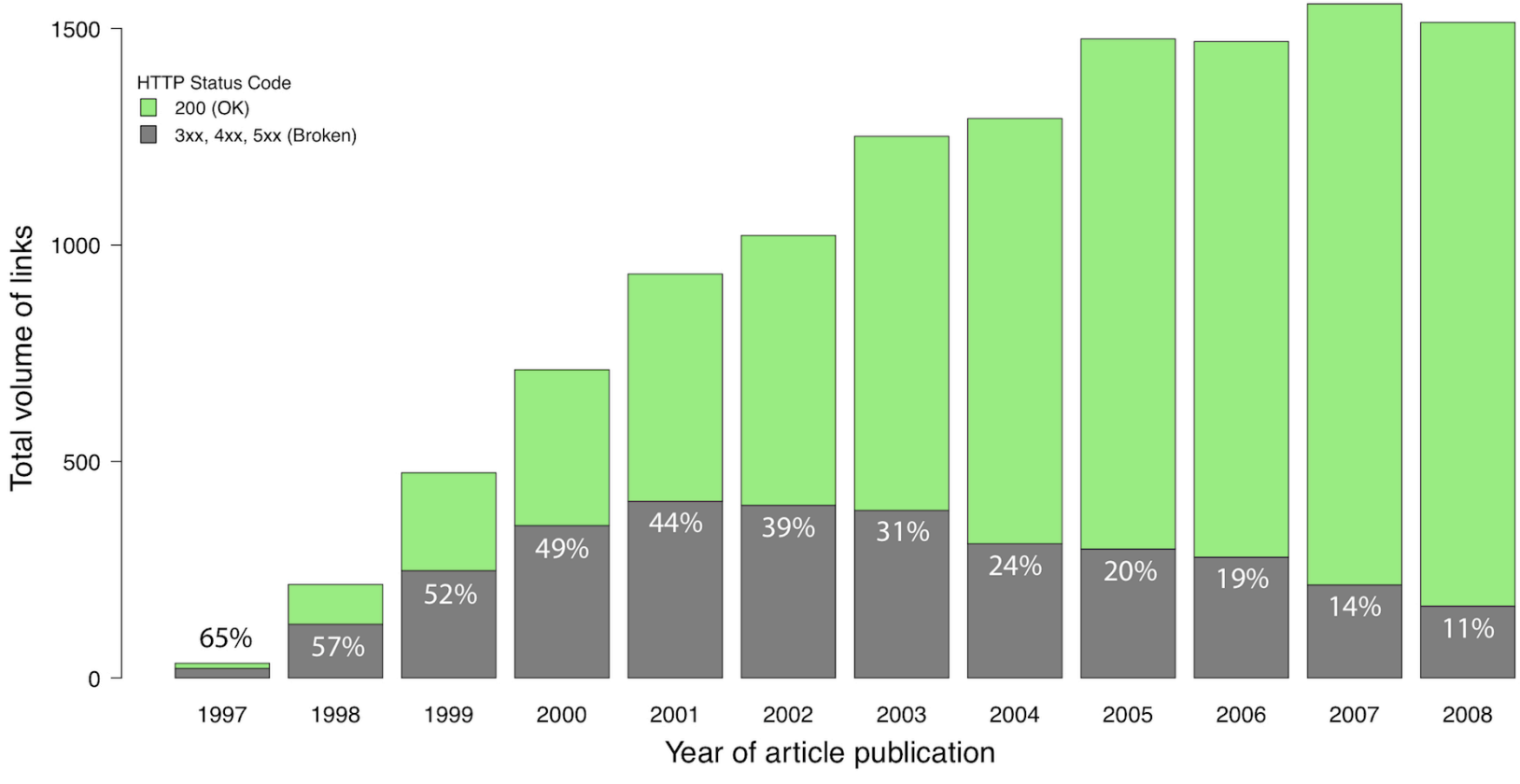
RESEARCH ARTICLE

How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers

Alberto Pepe, Alyssa Goodman, August Muench, Merce Crosas, Christopher Erdmann

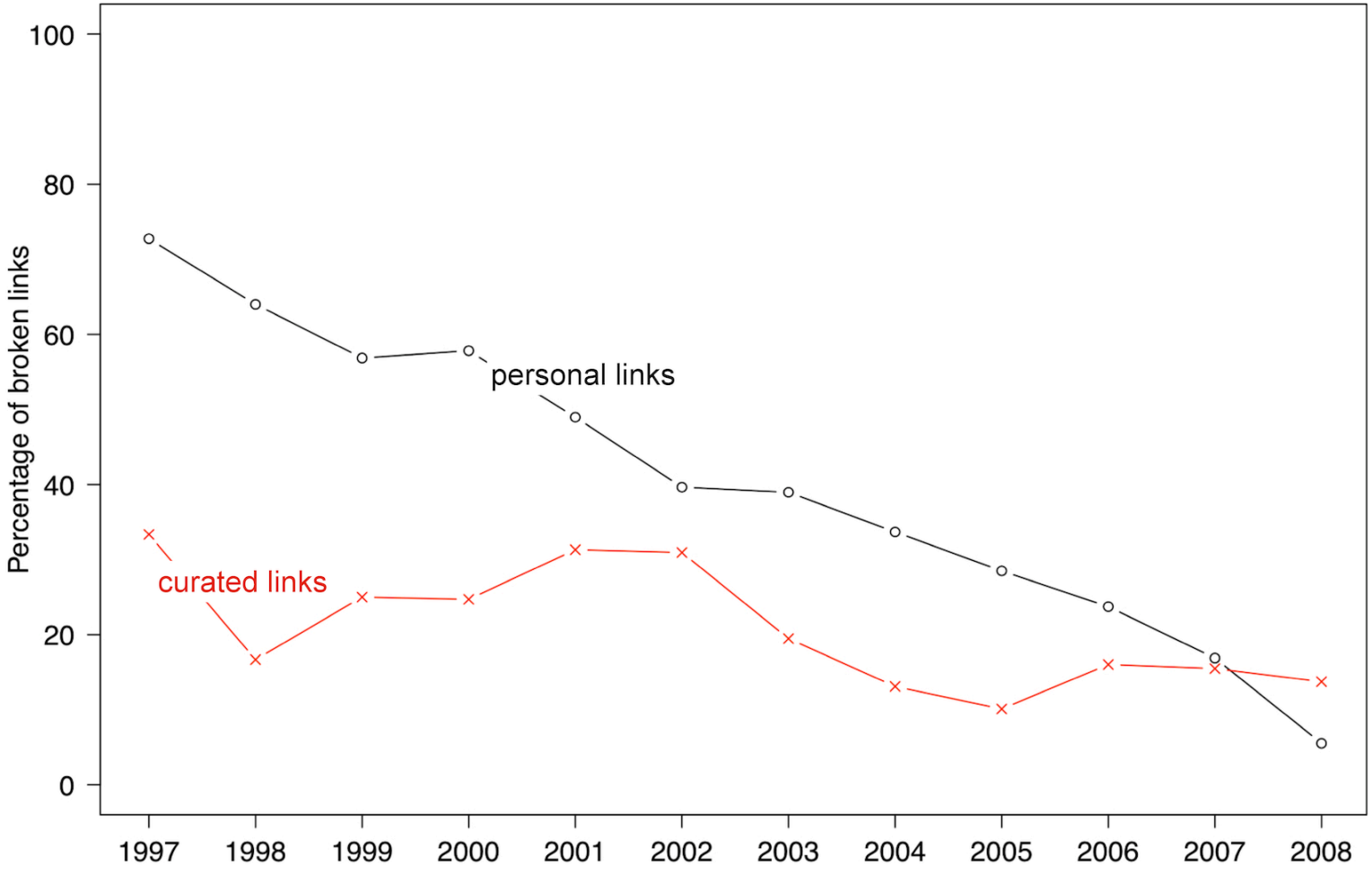
Published: August 28, 2014 • https://doi.org/10.1371/journal.pone.0104798

MORE THAN HALF OF LINKS TO DATA IN ARTICLES FROM 15 YEARS AGO ARE BROKEN



External links in all articles published between 1997 and 2008 in the four main astronomy journals published by the American Astronomical Society.

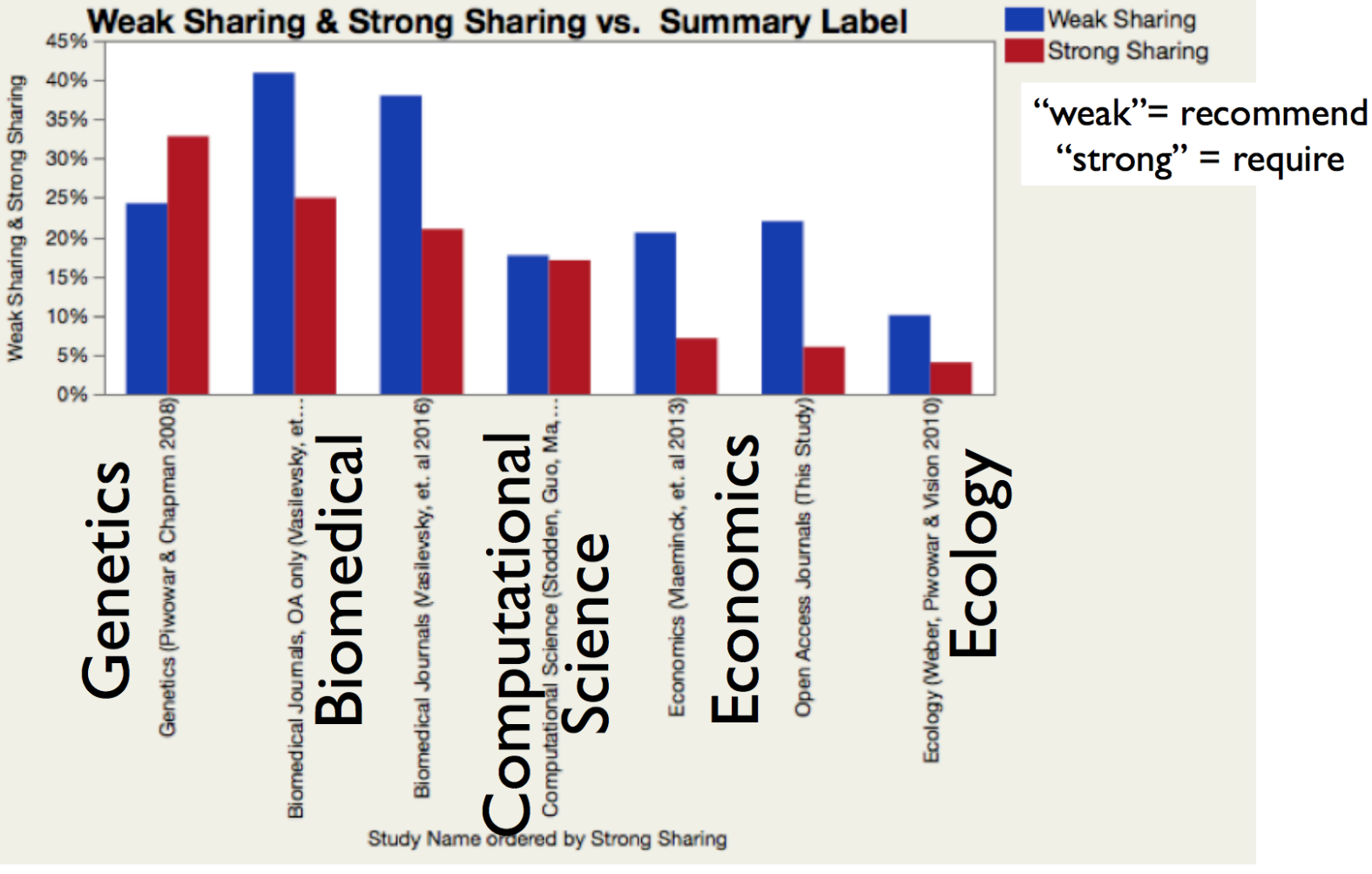
70% OF LINKS TO PERSONAL WEBSITES FROM ARTICLES PUBLISHED IN 1997 ARE BROKEN



HOW CAN WE IMPROVE DATA SHARING?

- New Norms
- New Incentives
- New Technology

FORMAL DATA-SHARING POLICIES ARE APPLIED IN JOURNALS ACROSS DISCIPLINES



Castro, Crosas, Garnett, Sheridan, Altman, 2017, Journal of Scholarly Publishing

MANY FUNDERS REQUIRE DATA SHARING & OPEN DATA

PRIVATE RESEARCH FUNDERS

- Bill and Melinda Gates Foundation Information Sharing Approach
- Sloan Foundation Data Sharing Policy
- Wellcome Trust Data Sharing Policy
- Arnold Foundation
- Moore Foundation
- Robert Wood Johnson Foundation
- HHMI Policy on the Sharing of Publication-Related Materials, Data and Software

PUBLIC RESEARCH FUNDERS

- Department of Agriculture
- Department of Commerce
- Department of Defense
- Department of Education
- Department of Energy
- Department of Health and Human Services
  - Agency for Healthcare Research and Quality (AHRQ)
  - Assistant Secretary for Preparedness and Response (ASPR)
  - Center for Disease Control and Prevention (CDC)
  - Food and Drug Administration (FDA)
  - National Institutes of Health (NIH)
- Department of Homeland Security
- Department of Housing and Urban Development
- Department of Interior
- Department of Labor
- Department of Transportation
- Department of Veterans Affairs
- Environmental Protection Agency (EPA)

Are reproducibility and open science starting to matter in tenure and promotion review?

July 14th, 2017, Brian Nosek

Tags: openscience, reproducibility

The NEW ENGLAND JOURNAL of MEDICINE

SOUNDING BOARD

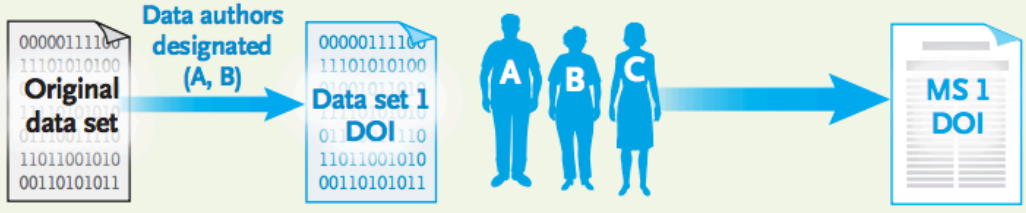
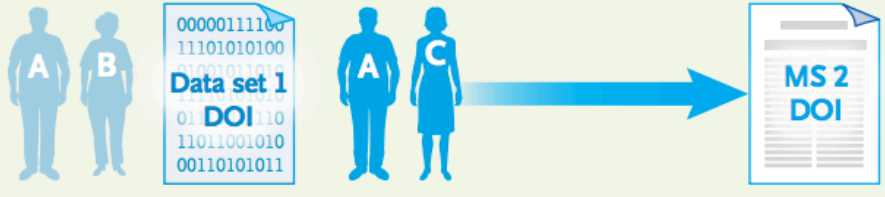
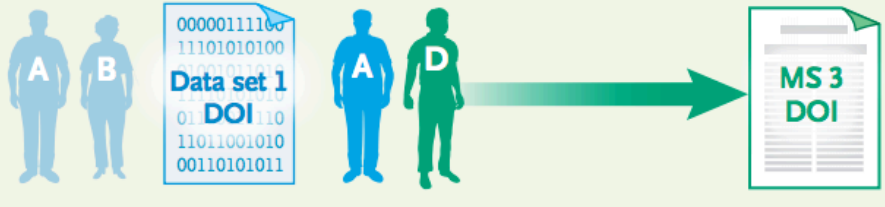
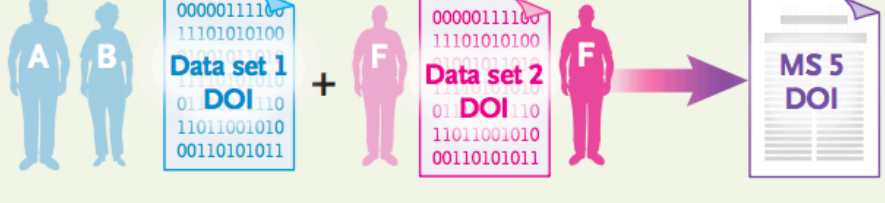
Data Authorship as an Incentive to Data Sharing

Barbara E. Bierer, M.D., Mercè Crosas, Ph.D., and Heather H. Pierce, J.D., M.P.H.

N Engl J Med 2017; 376:1684-1687 | April 27, 2017 | DOI: 10.1056/NEJMsb1616595

Data from well-designed and well-executed research not only are useful for the original purpose and secondary analyses by the original researchers but also can be repurposed for a variety of applications, including independent replication, avoidance of duplicative studies, generation or testing of hypotheses, and meta-analyses. Although the use of research data by persons other than those who originally gathered the data, termed "data sharing," is currently encouraged or mandated by a number of organizations, it remains a hypothetical or occasional occurrence. The Committee on Data of the International Council of Science has recently encouraged medical journal leadership through the draft data-sharing policy of the International Committee of Medical Journal Editors, charitable foundations such as the Wellcome Trust and the Bill and Melinda Gates Foundation, and the National Institutes of Health (NIH) in its recent request for information on data management and sharing strategies. Data sharing, whether elective or required, creates an obligation for the original investigators who obtain funding, design studies, collect and analyze data,

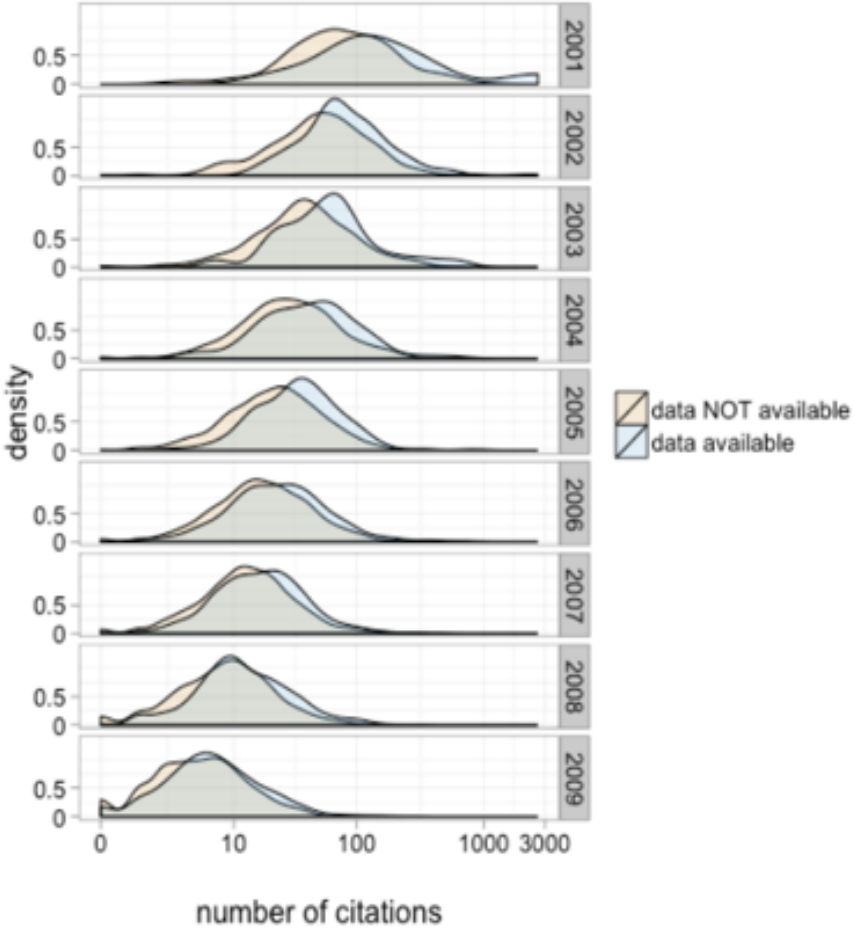
" We believe that both as a matter of fairness and as a matter of providing an incentive for data sharing, the persons who initially gathered the data should receive appropriate and standardized credit that can be used for academic advancement, for grant applications, and in broader situations."

1	**Primary authors generate data set and designate "data authors," DOI is assigned, and primary publication occurs**		3 Authors **(A, B, C)** 2 Data authors: **Data set 1 DOI (A, B)** 1 Manuscript DOI: **MS 1 DOI**
2	**Secondary analysis by members of primary group**		2 Authors **(A, C)** 2 Data authors: **Data set 1 DOI (A, B)** 1 New Manuscript DOI: **MS 2 DOI**
3	**Secondary analysis with collaborator**		2 Authors **(A, D)** 2 Data authors: **Data set 1 DOI (A, B)** 1 New Manuscript DOI: **MS 3 DOI**
4	**Secondary analysis by independent investigator**		1 Author **(E)** 2 Data authors: **Data set 1 DOI (A, B)** 1 New Manuscript DOI: **MS 4 DOI**
5	**Combination of data set with other data for new analysis by independent investigator**		1 Author **(F)** 3 Data authors: **Data set 1 DOI (A, B)** **Data set 2 DOI (F)** 1 New Manuscript DOI: **MS 5 DOI**

DATA SHARING INCREASES CITATIONS

From 10,555 studies with gene expression microarray data:

- Studies that shared data received 9% more citations
- Data reuse by other researchers continued for 6 years



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

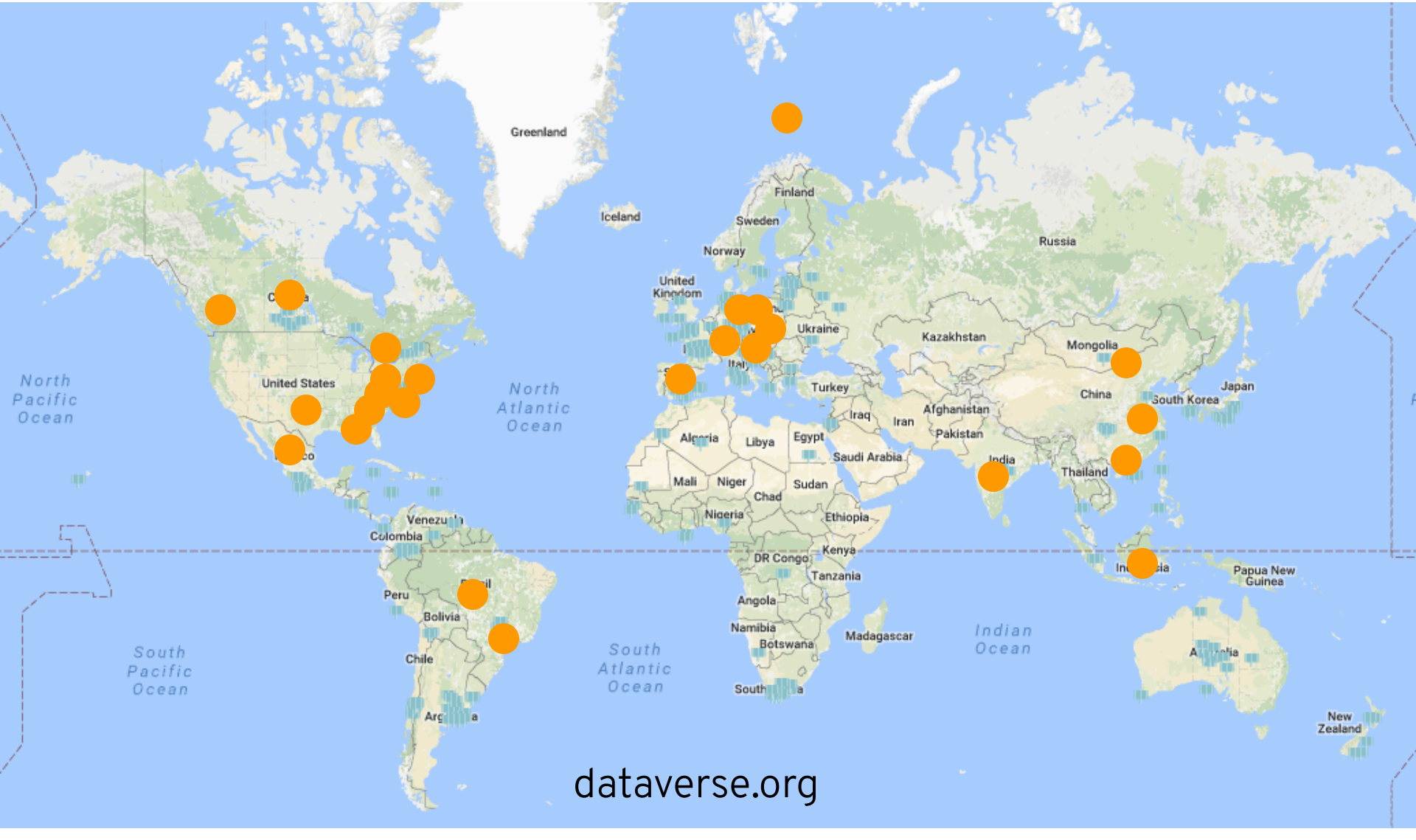
OUR INSTITUTE PROVIDES A TECHNOLOGY SOLUTION TO DATA SHARING

An open-source software to share, cite, and find data.

Developed at Harvard's Institute for Quantitative Social Science

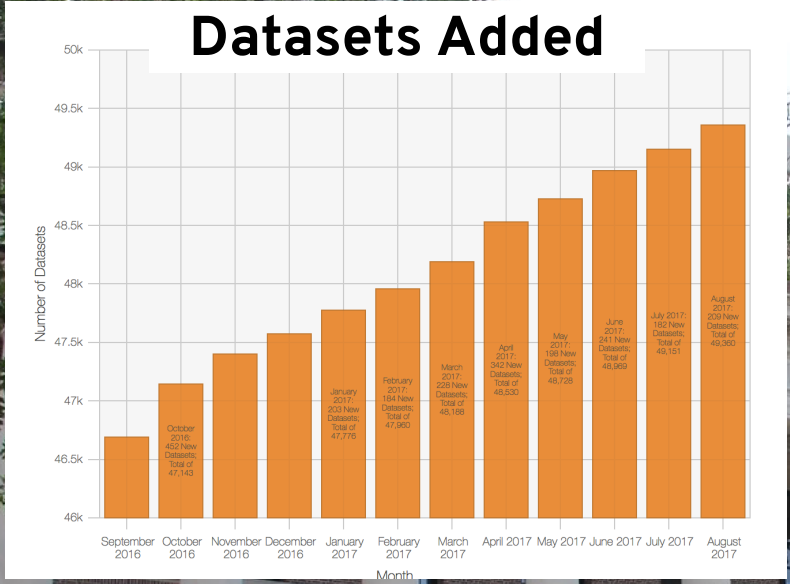
2006 (we started)

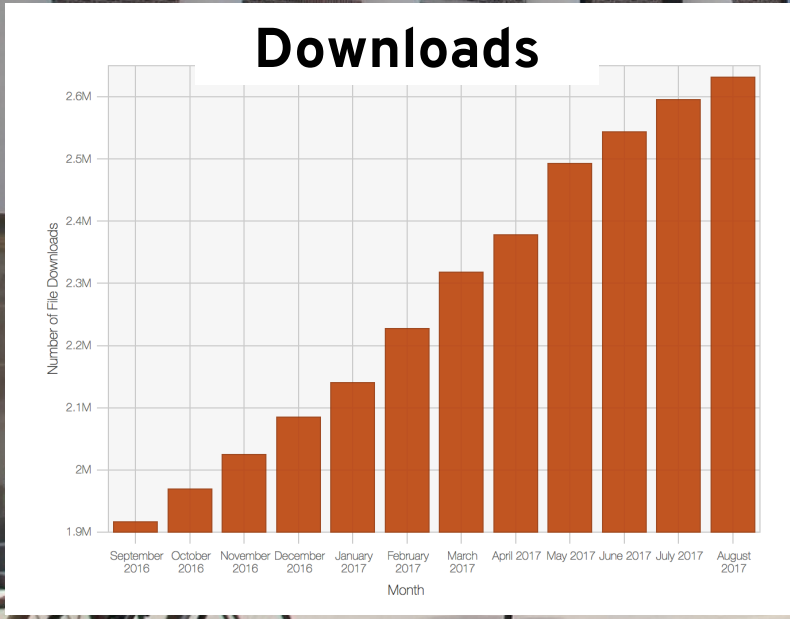
2017



dataverse.org

HOW RESEARCHERS SHARE & USE DATA WITH DATAVERSE





Harvard Dataverse Repository

> 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

OUR CONTRIBUTIONS TO ENHANCE DATA SHARING

King, 1995, Replication, Replication

Altman et al, 2001, A Digital Library for the Dissemination and Replication of Quantitative Social Science

Altman and King, 2007, A Proposed Standard for the Scholarly Citation of Quantitative Data

King, 2007, An Introduction to the Dataverse Network as an Infrastructure for Data Sharing

Crosas, 2012, The Dataverse Network: an open source application for sharing, discovering, and preserving research data

Crosas, 2013, A Data Sharing Story

Altman and Crosas, 2013, The Evolution to Data Citation: from principles to implementation

2014, Joint Declaration of Data Citation Principles

Pepe et al, 2014, How Do Astronomers Share Data?

Goodman et al, 2014, Ten Simple Rules for the Care and Feeding of Scientific Data

Crosas, Honaker, King, Sweeney, 2015, Automating Open Science for Big Data

Castro et al, 2015, Achieving Human and Machine Accessibility of Cited Data

Sweeney, Crosas, Bar-Sinai, 2015, Sharing Sensitive Data with Confidence: The DataTags System

Meyer et al. 2016, Data Publication with the Structural Biology Data Grid Supports Live Analysis

Wilkinson et al, 2016, The FAIR Guiding Principles for Scientific Data Management and Stewardship

Bierer, Crosas, Pierce, 2017, Data Authorship as an Incentive to Data Sharing

2017

Data should be ...

FINDABLE

ACCESSIBLE

INTERPOPERABLE

REUSABLE

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

Nature Scientific Data

FAIR DATA IN DATAVERSE

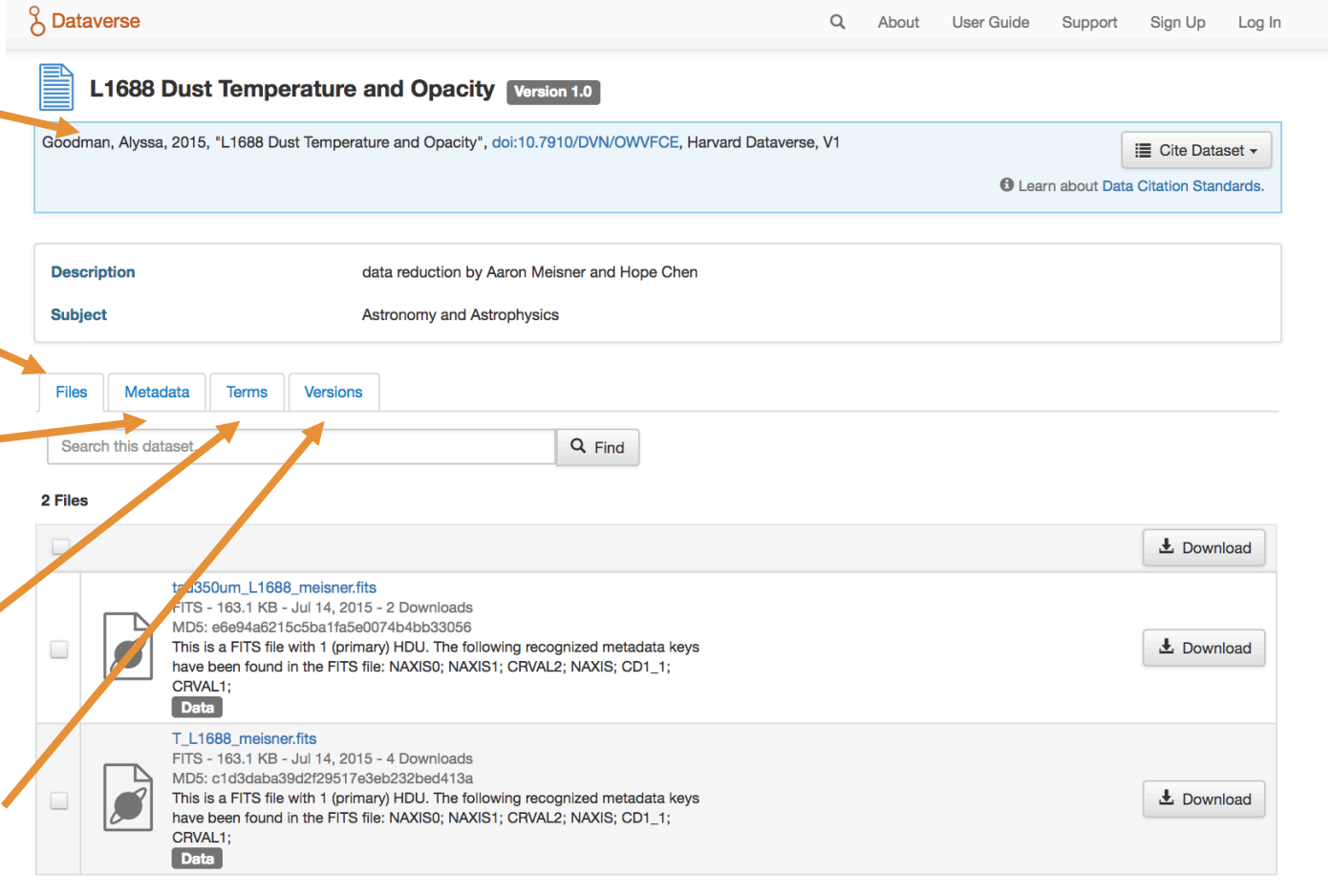
Data Citation with Persistent Identifier (DOI)

Data Files

Metadata

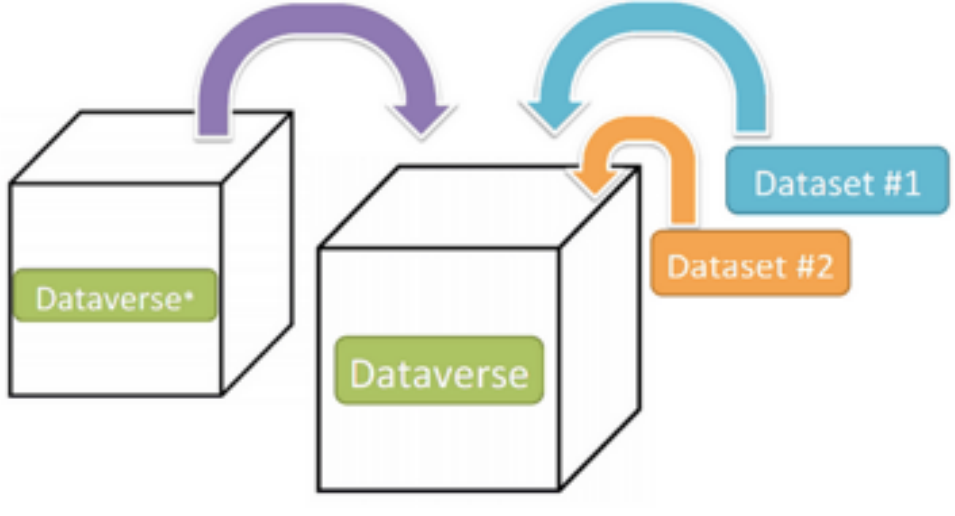
Data Licenses, User Agreements

Dataset Versions



A DATAVERSE IS A CONTAINER OF DATASETS AND A DATASET IS A CONTAINER OF DATA FILES, DOCUMENTATION, AND CODE

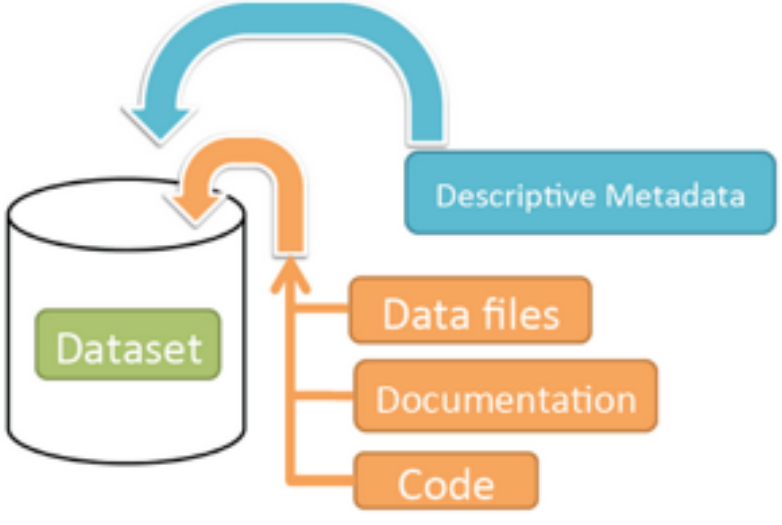
Schematic Diagram of a **Dataverse** in Dataverse 4.0



Container for your **Datasets** and/or **Dataverses***

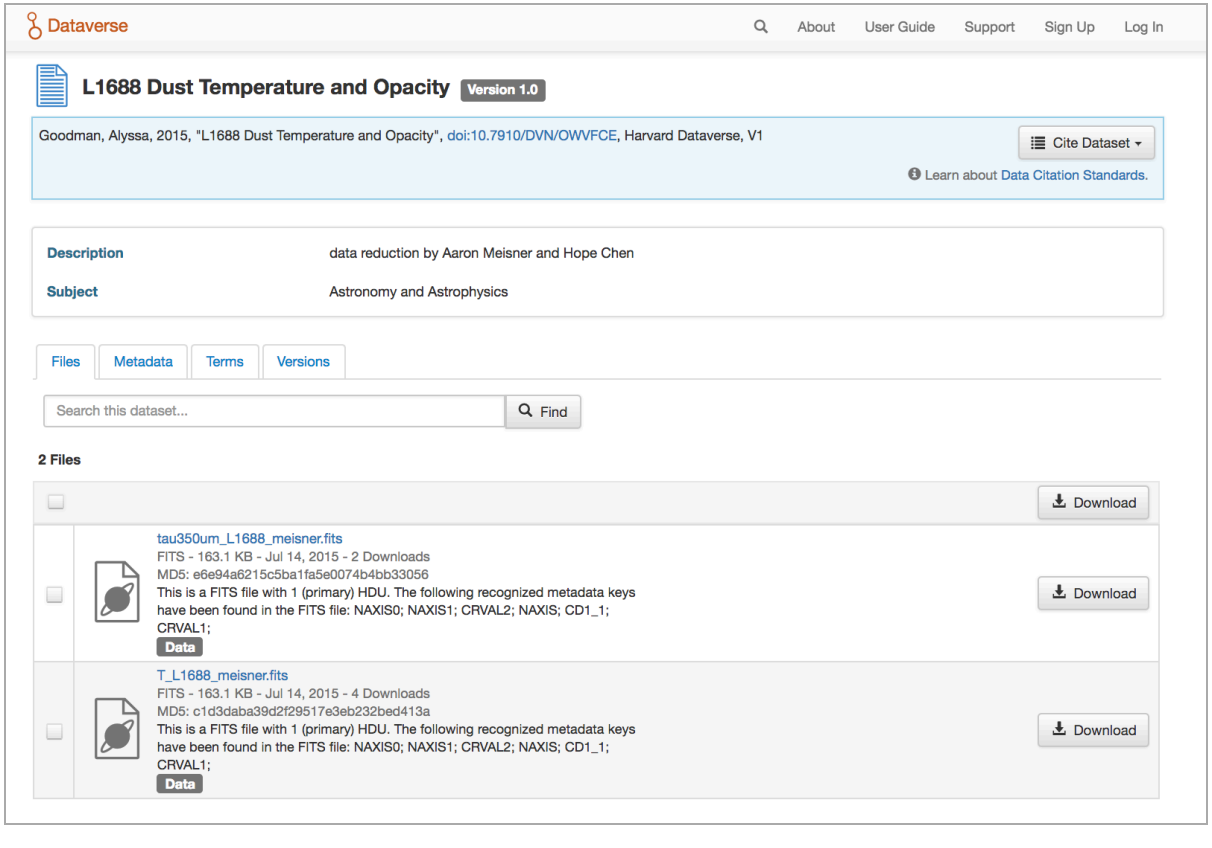
* Dataverses can now contain other Dataverses (this replaces Collections & Subnetworks)

Schematic Diagram of a **Dataset** in Dataverse 4.0



Container for your data, documentation, and code.

DATAVERSE SUPPORTS ASTRONOMY DATA



- **Supports default astronomy metadata fields (based on virtual observatory schema)**
- **Extracts header metadata from FITS files upon ingest**

DATAVERSE USED BY MAX-PLANCK INSTITUTE ...

Dataverse

About User Guide Support Sign Up Log In

J. Michael's Analysis Projects Dataverse (Max-Planck Institute for Extraterrestrial Physics)

Harvard Dataverse > **J. Michael's Analysis Projects Dataverse**

Contact Share

Data products, analysis results, and code associated with my publications to be used for replication.

Search this dataverse... Find Advanced Search

- [x] Dataverses (0)
- [x] Datasets (1)
- [ ] Files (133)

Publication Date

2017 (1)

Subject

Astronomy and Astrophysics (1)

1 to 1 of 1 Result

Sort

Replication Data for: Is the Spectral Width of GRBs a Reliable Measure of GRB Emission Physics?

May 15, 2017

Burgess, J. Michael, 2017, "Replication Data for: Is the Spectral Width of GRBs a Reliable Measure of GRB Emission Physics?", doi:10.7910/DVN/BDC2GS, Harvard Dataverse, V1

This data set includes the processed GBM PHA/BAK/RSP files used in the publication. The files are readable by XSPEC or 3ML's OGIPLike plugin. Additionally, the 3ML analysis results FITS files are included. These results include the Bayesian posteriors from the fits. These can be...

DATAVERSE IN THE ASTRONOMY NEWS ...

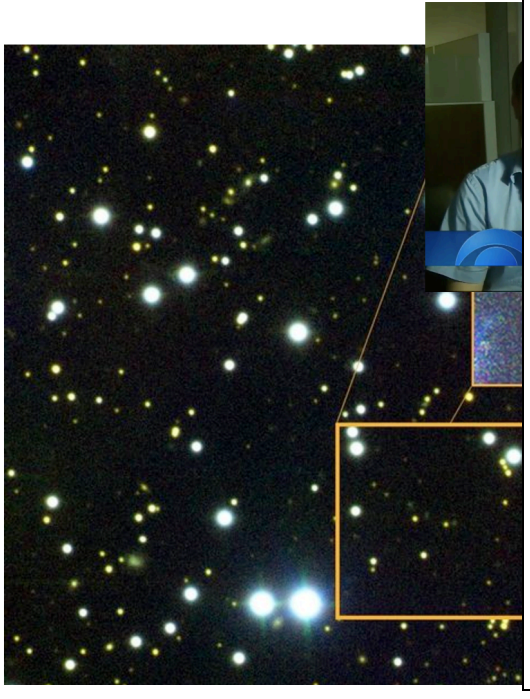
HUFFPOST US EDITION

WEIRD NEWS 08/31/2017 05:15 am ET

Scientists Detect Mysterious Radio Signals From Deep Space

Strange bursts are coming from 3 billion light years away.

By Ed Mazza



More than 12,000 downloads!

Dataverse

About User Guide Support Sign Up Log In

Metrics 12,520 Downloads

Contact Share

The Sound of Fast Radio Burst FRB 121102 Version 1.0

Law, Casey, 2016, "The Sound of Fast Radio Burst FRB 121102", doi:10.7910/DVN/QSWJE6, Harvard Dataverse, V1

Cite Dataset

Learn about Data Citation Standards.

Description	Sound files generated from nine radio transients detected by the Very Large Array toward cosmological radio transient, FRB 121102. The sound of FRB 121102 has a "chirp" that is caused by dispersion. Original publication is Chatterjee et al (2017), "The direct localization of a fast radio burst and its host". Original data is available at https://doi.org/10.7910/DVN/TLDKXG.
Subject	Astronomy and Astrophysics
Keyword	radio interferometry, astronomy, fast radio burst
Notes	These bursts were observed with the VLA from 2.5 to 3.5 GHz with a dispersion measure of approximately 570 pc/cm3.

Files Metadata Terms Versions

Search this dataset... Find

9 Files

Download

57623.mp3
audio/mp3 - 200.0 KB - Dec 31, 2016 - 6,256 Downloads
MD5: 9967df426bfaa8d239110dbd3c9ffbc0 Download

57633.scan13.mp3
audio/mp3 - 200.0 KB - Dec 31, 2016 - 1,930 Downloads
MD5: 0c864a33799d4f98a4e0f018f7ad59c3 Download

WHAT ARE WE WORKING ON NOW?

DATA PROVENANCE

TRACK THE ORIGINAL SOURCE OF A DATASET

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

Comment: If these data could talk

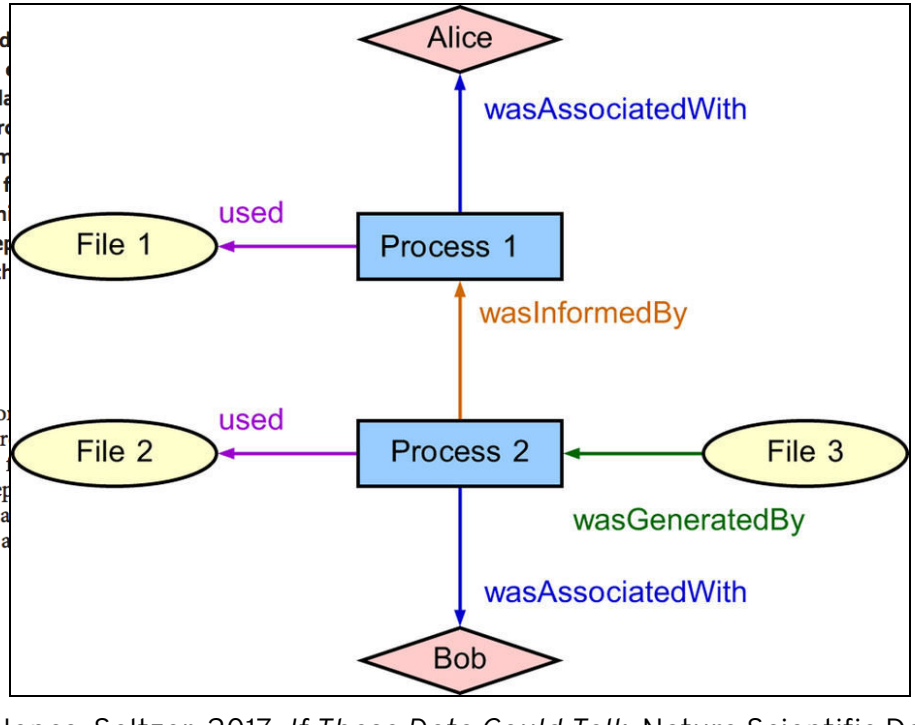
Thomas Pasquier[1], Matthew K. Lau[2], Ana Trisovic[3,4], Emery R. Boose[2], Ben Couturier[3], Mercè Crosas[5], Aaron M. Ellison[2], Valerie Gibson[4], Chris R. Jones[4] & Margo Seltzer[1]

Received: 12 April 2017
Accepted: 24 July 2017
Published: 5 September 2017

In the last few decades, data-driven method
Open data and open-source software have
manage and analyze the growing flood of da
fields exhibit distressingly low rates of repr
issue, we believe that there is a lack of form
from the data source to the analysis to the f
make their research and data accessible, th
reporting, which contributes to issues of rep
through *systematic* and *formal* records of th
publications and researchers.

Reproducibility

The success and power of science depends o
issues with reproducibility have surfaced acr
issues have emanated from fields ranging
including medicine[1]. Although the lack of rep
remains a worrisome issue. This comes at a
exponentially[3]. At the same time, the data a
computationally demanding.

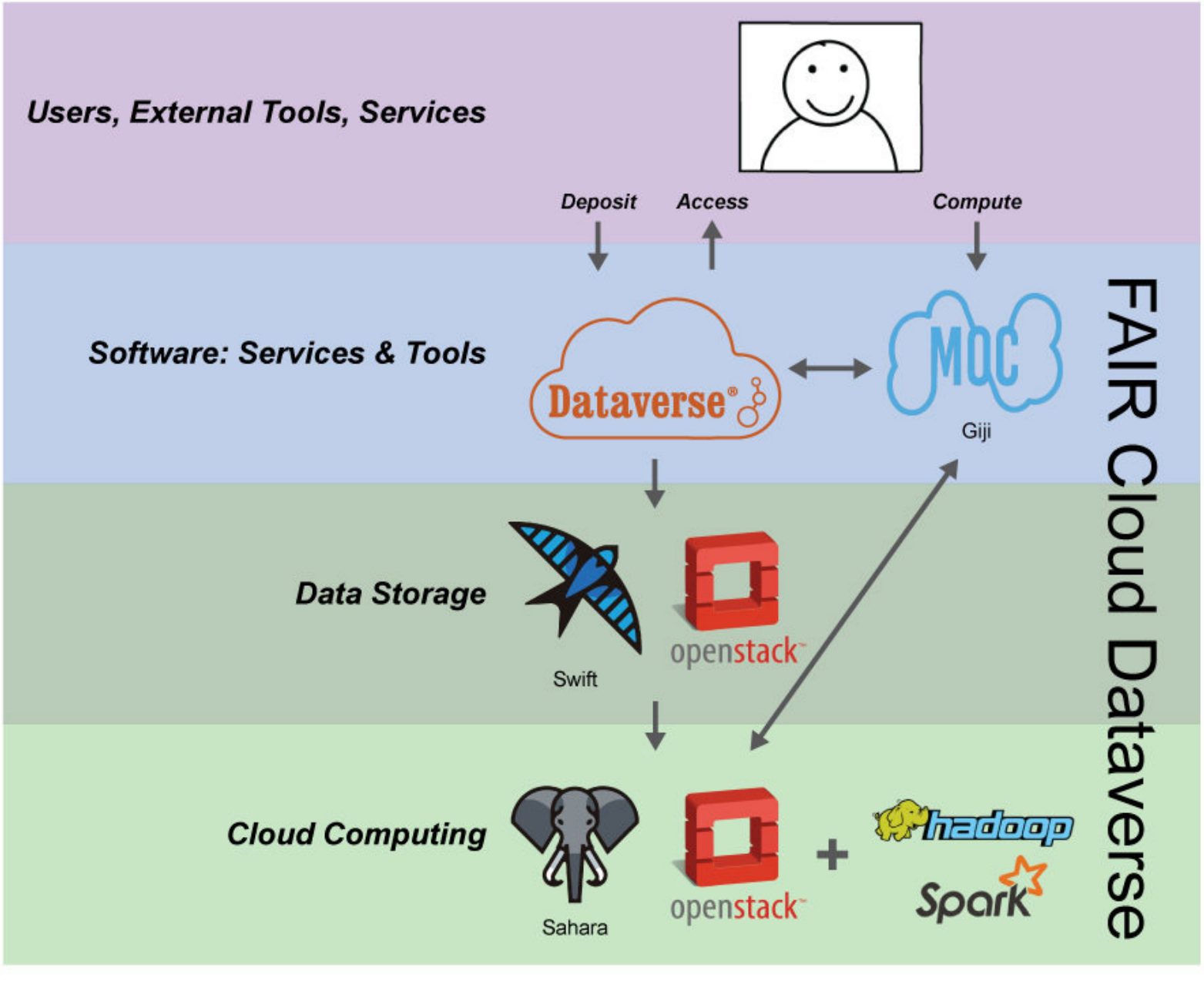


Pasquier, Lau, Trisovic, Boose, Coutierer, Crosas, Ellison, GIbson, Jones, Seltzer, 2017, *If These Data Could Talk*, Nature Scientific Data

(Data Provenance examples from CERN and Harvard Forest)

CLOUD DATAVERSE

COMBINE DATA REPOSITORIES WITH CLOUD COMPUTING


Users, External Tools, Services
Deposit
Access
Compute
Software: Services & Tools
Dataverse
MOC
Giji
Data Storage
Swift
openstack
Cloud Computing
Sahara
openstack
+
hadoop
Spark
FAIR Cloud Dataverse

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 auto-generated data user agreements (DUA)
- Review Board

orange **Direct Access**

Requires:

- User registration
- Approval needed for access
- Signed DUA

green **Privacy Preserving Access**

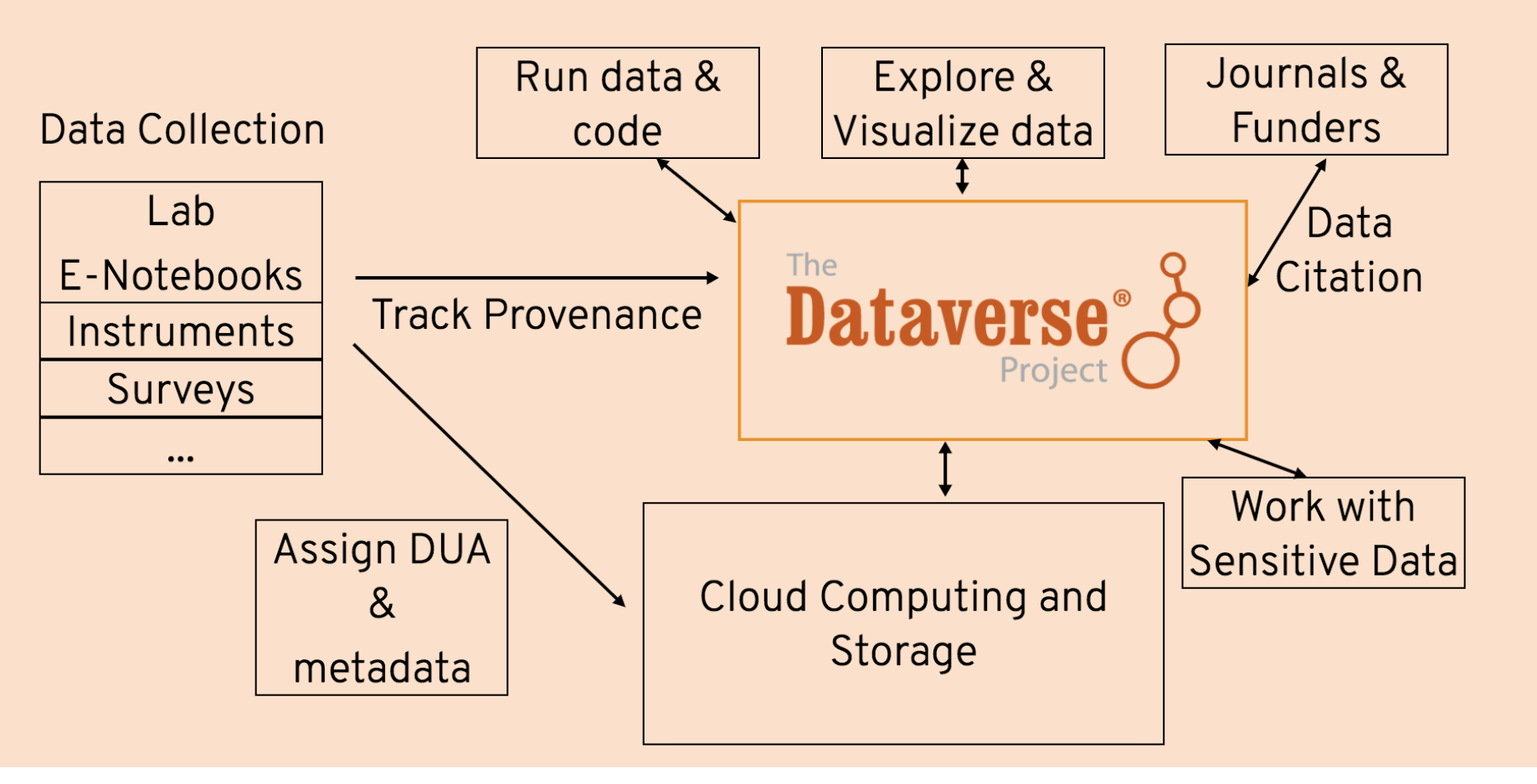
- Requires user registration
- Provides access to differentially private statistics using Private data Sharing Interface (PSI)

Harvard Data Privacy Tools Project: privacytools.seas.harvard.edu

DataTags Project: datatags.org

INTEGRATION WITH TOOLS

DATAVERSE AS PART OF THE DATA LIFECYCLE


Data Collection
Lab
E-Notebooks
Instruments
Surveys
...
Track Provenance
Run data & code
Explore & Visualize data
Journals & Funders
Data Citation
The Dataverse® Project
Work with Sensitive Data
Assign DUA & metadata
Cloud Computing and Storage


DATAVERSE
COMMUNITY

49 SOFTWARE CONTRIBUTORS

BI-WEEKLY COMMUNITY CALLS

235 ATTENDEES
26 ORGANIZATIONS/UNIVERSITIES
11 COUNTRIES

ANNUAL COMMUNITY MEETING

NEXT: JUNE 13, 14, 15, 2018

THANKS

@mercecrosas

scholar.harvard.edu/mercecrosas

dataverse.org