

# THE DATAVERSE PROJECT

**Mercè Crosas, Institute for Quantitative Social Science, Harvard University**



**@mercecrosas**

**RDA 10TH PLENARY, MONTREAL, SEPTEMBER 21, 2017**



# OUR INSTITUTE PROVIDES A TECHNOLOGY SOLUTION TO DATA SHARING

Institute for Quantitative Social Science, Harvard University  
@IQSS



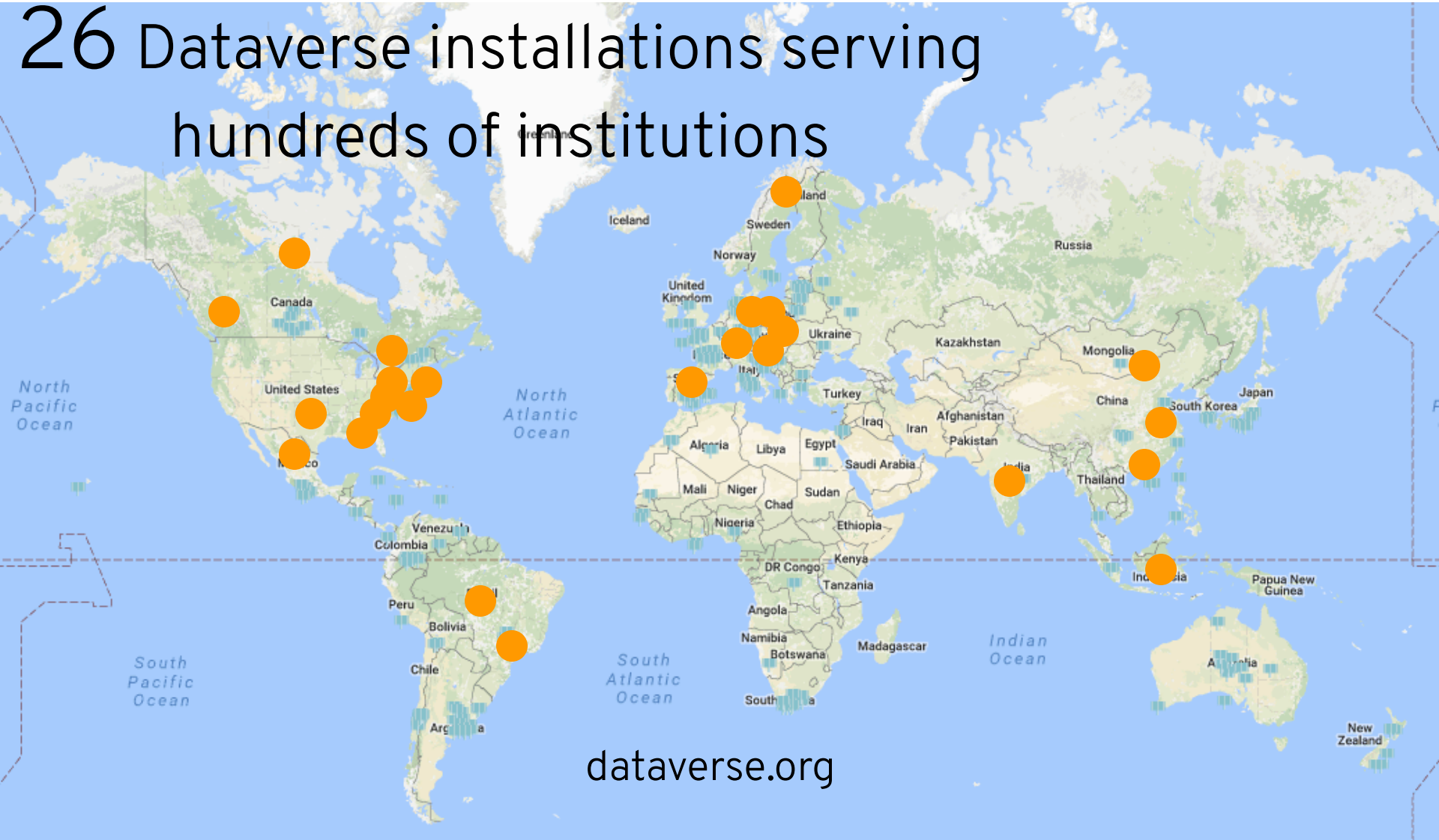
An open-source software to share, cite, and find data.  
Developed at Harvard's Institute for Quantitative Social Science  
with the contribution of an active and growing community.

2006 (we started)

2017



# 26 Dataverse installations serving hundreds of institutions

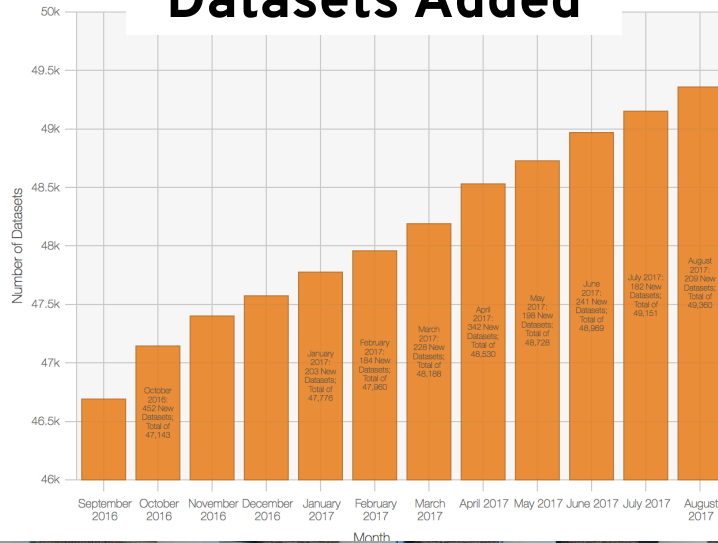


[dataverse.org](http://dataverse.org)



# HOW RESEARCHERS SHARE & USE DATA WITH DATAVERSE

## Datasets Added



## Harvard Dataverse Repository

A public repository for research data

> 70,000 datasets total

> 49,000 datasets uploaded to Harvard Dataverse repository

200 datasets/month

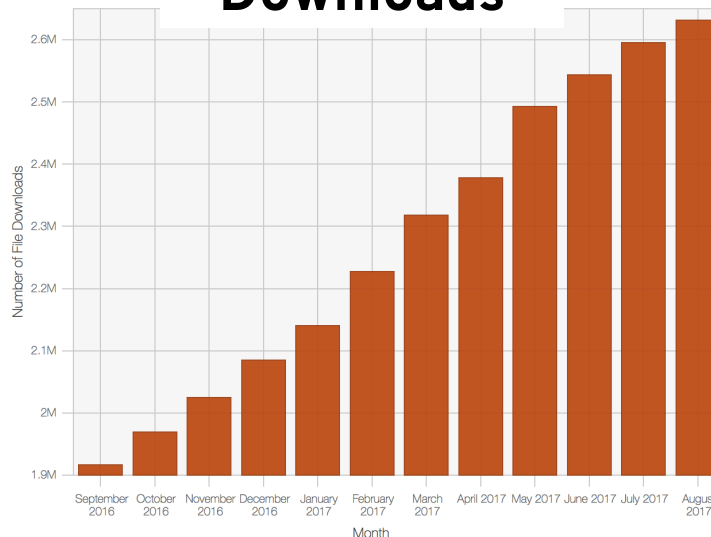
> 340,000 files

4,000 files/month

> 2.5 M downloads

60,000 downloads/month

## Downloads



[dataverse.harvard.edu](http://dataverse.harvard.edu)

# OUR CONTRIBUTIONS TO ENHANCE DATA SHARING

## King, 1995, Replication, Replication

## 2014, Joint Declaration of Data Citation Principles

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

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

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

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

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

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

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

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

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

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

Crosas, 2013, A Data Sharing Story

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

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

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



2017

Data should be ...

**F**INDABLE

**A**CCESSIBLE

**I**NTERPOPERABLE

**R**EUSABLE

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

Dataverse Q About User Guide Support Sign Up Log In

## L1688 Dust Temperature and Opacity Version 1.0

Goodman, Alyssa, 2015, "L1688 Dust Temperature and Opacity", doi:10.7910/DVN/OWVFCE, Harvard Dataverse, V1

 Cite Dataset 

 Learn about Data Citation Standards.

Data Files

**Description** data reduction by Aaron Meisner and Hope Chen  
**Subject** Astronomy and Astrophysics

Metadata

[Files](#) [Metadata](#) [Terms](#) [Versions](#)





Search this dataset... Q Find

2 Files

Data Licenses,  
User Agreements,  
Restrictions

Versions

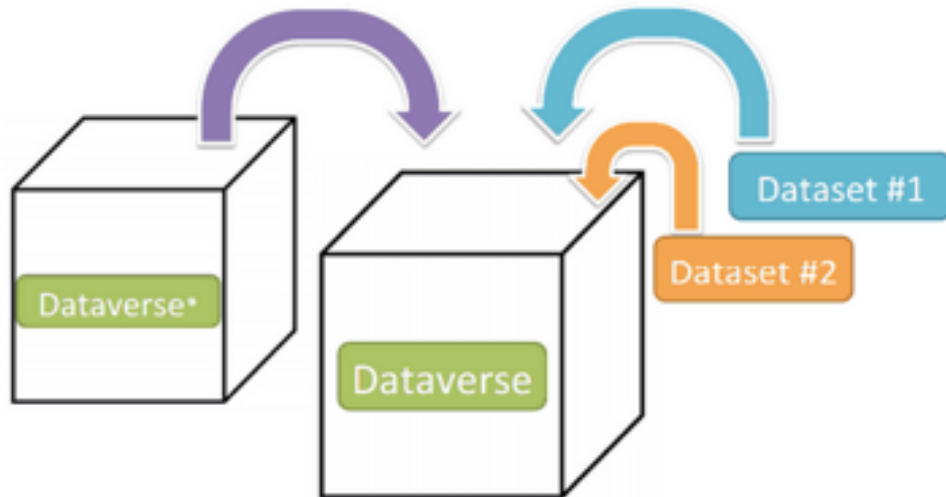
APIs

	<a href="#">t3350um_L1688_meisner.fits</a>	
	<p>FITS - 163.1 KB - Jul 14, 2015 - 2 Downloads MD5: e6e94a6215c5ba1fa5e0074b4bb33056 This is a FITS file with 1 (primary) HDU. The following recognized metadata keys have been found in the FITS file: NAXIS0; NAXIS1; CRVAL2; NAXIS; CD1_1; CRVAL1;</p> <p><b>Data</b></p>	
	<a href="#">T_L1688_meisner.fits</a>	
	<p>FITS - 163.1 KB - Jul 14, 2015 - 4 Downloads MD5: c1d3daba39d2f29517e3eb232bed413a This is a FITS file with 1 (primary) HDU. The following recognized metadata keys have been found in the FITS file: NAXIS0; NAXIS1; CRVAL2; NAXIS; CD1_1; CRVAL1;</p> <p><b>Data</b></p>	



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

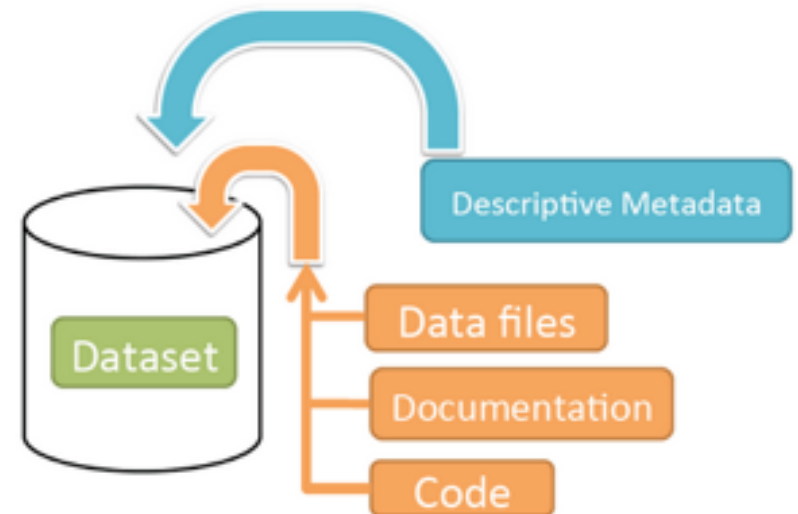
Schematic Diagram of a **Dataverse** in Dataverse 4.0



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 RICH SUPPORT FOR DATA

Harvard DataVerse > **New aspects of the innate immune system in the infant born prematurely.**

 Metrics

8 Downloads

 [Contact](#)  [Share](#)



## New aspects of the innate immune system in the infant born prematurely. Version 2.0

Kwinta, Przemko, 2017, "New aspects of the innate immune system in the infant born prematurely.", doi:10.7910/DVN/AHUUCD, Harvard DataVerse, V2, UNF:6:6d/FPnt4BGjyb91Jv1/33g==

### Description

The aim of this project is to assess the function of selected innate immu

### Subject

Medicine, Health and Life Sciences

### Keyword

preterm delivery, innate immunity, inflammasome

[Files](#)

[Metadata](#)

[Terms](#)

[Versions](#)

Search this dataset...

 Find


1 File




[Database PD-1.tab](#)

Tabular Data - 13.3 KB - Sep 21, 2017 - 4 Downloads


18 Variables, 102 Observations - UNF:6:6d/FPnt4BGjyb91Jv1/33g==


 Explore

 Download ▾

- Extract variable metadata from tabular data files
- Visualize geospatial files in a map
- Extract header metadata from FITS files
- Reformat

# DATAVERSE CUSTOMIZATION AND BRANDING

Q 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...

Q Find

[Advanced Search](#)

-  **Dataverses (0)**
-  **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 INTEGRATION WITH JOURNALS

Data citation from article to data, review workflow, replication code

**AJPS AMERICAN JOURNAL of POLITICAL SCIENCE**

American Journal of Political Science

© Midwest Political Science Association

July 2017  
Volume 61, Issue 3  
Pages 509–763  
[Previous Issue](#)

Select All [Save to profile](#) [Export citation](#)

**ISSUE INFORMATION**

[Issue Information - Table of Contents \(pages 509–542\)](#)  
Version of Record online: 3 JUL 2017 | DOI: 10.1111/ajps.12345  
[Abstract](#) | [Article](#) | [PDF\(448K\)](#) | [Request Permissions](#)

**ARTICLES**

[Race, Representation, and the Voting Rights Act \(pages 509–542\)](#)  
Sophie Schuit and Jon C. Rogowski  
Version of Record online: 12 DEC 2016 | DOI: 10.1111/ajps.12346  
[Abstract](#) | [Article](#) | [PDF\(264K\)](#) | [References](#)

[The Fulfillment of Parties' Election Pledges: A Comparison of Power Sharing \(pages 527–542\)](#)  
Robert Thomson, Terry Royed, Elin Naurin, Joaquín A. Ennsner-Jedenastik, Mark Ferguson, Petia Kostadinova

**Dataverse** [About](#) [User Guide](#) [Support](#) [Sign Up](#) [Log In](#)

**AJPS AMERICAN JOURNAL of POLITICAL SCIENCE**

American Journal of Political Science (AJPS) Dataverse (Michigan State University) [ajps.org](#)

[Harvard Dataverse](#) > [American Journal of Political Science \(AJPS\) Dataverse](#)

[Contact](#) [Share](#)

The *American Journal of Political Science* is committed to significant advances in knowledge and understanding of citizenship, governance, and politics, and to the public value of political science research. To find out more about our data integrity policies, please visit [our website](#).

Search this dataverse... [Find](#) [Advanced Search](#) [Add Data](#)

[Dataverses \(0\)](#)  
 [Datasets \(281\)](#)  
 [Files \(3,180\)](#)

**1 to 10 of 281 Results** [Sort](#)

**Publication Date**  
2014 (67)  
2013 (64)  
2015 (61)  
2016 (37)  
2012 (27)

**Author Name**  
Broockman, David (5)

**Replication Data for: Leader Influence and Reputation Formation in World Politics**  
Aug 24, 2017  
Dafoe, Allan; Renshon, Jonathan; Huth, Paul, 2017, "Replication Data for: Leader Influence and Reputation Formation in World Politics", doi:10.7910/DVN/HIQ4BV, Harvard Dataverse, V1, UNF:6:jwmD+ivAhXY75lvpR7cRaQ==

The study of reputation is one of the foundational topics of modern international relations. However, fundamental questions remain, including the question of to whom reputations adhere: states, leaders, or both? We offer a theory of influence-specific reputations (ISR) that unifi...

**Replication Data for: Disloyal Brokers and Weak Parties**  
Jul 26, 2017  
Newson, Lucas M., 2017, "Replication Data for: Disloyal Brokers and Weak Parties"



**WHAT ARE WE WORKING ON NOW?**

# **DATA PROVENANCE**

**TRACK THE ORIGINAL SOURCE OF A DATASET**



# SCIENTIFIC DATA

## OPEN Comment: If these data could talk

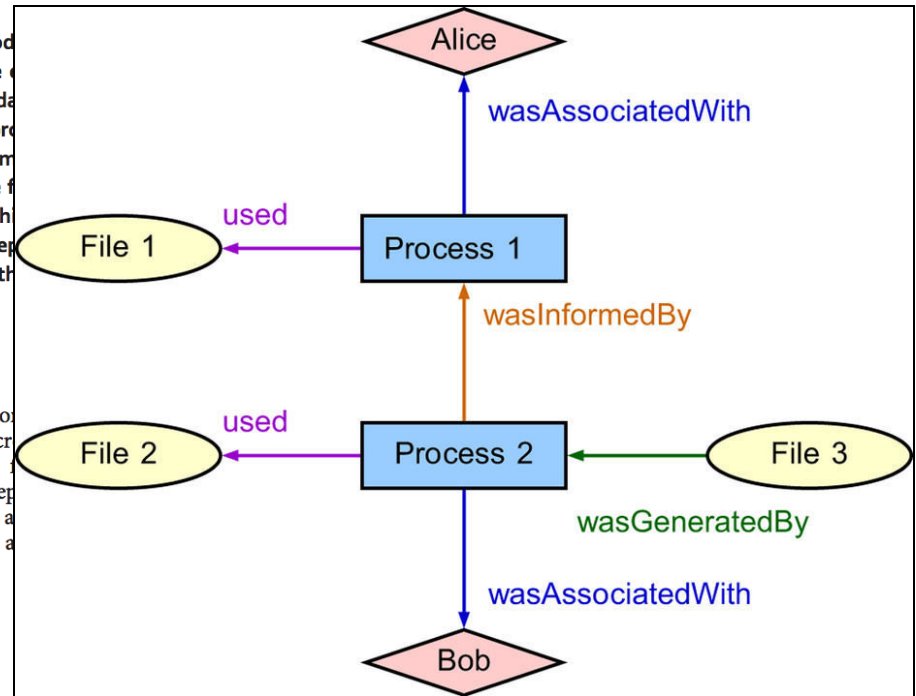
Thomas Pasquier<sup>1</sup>, Matthew K. Lau<sup>2</sup>, Ana Trisovic<sup>3,4</sup>, Emery R. Boose<sup>2</sup>, Ben Couturier<sup>3</sup>, Mercè Crosas<sup>5</sup>, Aaron M. Ellison<sup>2</sup>, Valerie Gibson<sup>4</sup>, Chris R. Jones<sup>4</sup> & Margo Seltzer<sup>1</sup>

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

In the last few decades, data-driven methods have revolutionized science. Open data and open-source software have helped researchers manage and analyze the growing flood of data. However, some fields exhibit distressingly low rates of reproducibility. In this issue, we believe that there is a lack of formal records from the data source to the analysis to the final publication. To make their research and data accessible, they need to report through *systematic* and *formal* records of their data through publications and researchers.

### Reproducibility

The success and power of science depends on the ability to reproduce results. Issues with reproducibility have surfaced across many fields, including medicine<sup>1</sup>. Although the lack of reproducibility remains a worrisome issue. This comes at a time when data is exponentially<sup>3</sup>. At the same time, the data is becoming computationally demanding.



Pasquier, Lau, Trisovic, Boose, Couturier, 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**



Giji

**Data Storage**



Swift



openstack™

**Cloud Computing**



Sahara



openstack™



Spark

**FAIR Cloud Dataaverse**

# STRUCTURAL BIOLOGY DATA BANK

THE LEONA M. AND HARRY B.  
HELMSLEY  
CHARITABLE TRUST

---

## The SBGrid Data Bank

We support publication of X-ray diffraction, MicroED, LLSM datasets, as well as structural models. All visitors can access our Laboratory and Institutional Collections. All structural biologists are invited to deposit datasets.

[VIEW DATA](#)[DEPOSIT DATA](#)**359** Datasets**359** Institutions**313** Structures

### Deposit

Share your data with the community. Every dataset deposited



### Explore

Browse all published datasets and download via rsync



### Cite

Give credit to the data used in your research. Every dataset published

# 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](https://privacytools.seas.harvard.edu)

DataTags Project: [datatags.org](https://datatags.org)

# **INTEGRATION WITH TOOLS**

**DATVERSE AS PART OF THE DATA LIFECYCLE**



## Data Collection

Lab
E-Notebooks
Instruments
Surveys
...

Track Provenance

Assign DUA  
&  
metadata

Run data &  
code

Explore &  
Visualize data

Journals &  
Fundors

Data  
Citation

Work with  
Sensitive Data

Cloud Computing and  
Storage

The  
**Dataverse**<sup>®</sup>  
Project 



# DATAVERSE COMMUNITY

**380 MEMBERS IN  
OUR COMMUNITY  
GROUP**

<https://groups.google.com/forum/#!members/dataverse-community>

# BI-WEEKLY COMMUNITY CALLS

**235 ATTENDEES**

**26 ORGANIZATIONS/UNIVERSITIES**

**11 COUNTRIES**

**49 SOFTWARE  
CONTRIBUTORS**



# ANNUAL COMMUNITY MEETING

NEXT: JUNE 13, 14, 15, 2018



# THANKS

@mercecrosas

[scholar.harvard.edu/mercecrosas](https://scholar.harvard.edu/mercecrosas)

[dataverse.org](https://dataverse.org)

