

Best Practices in Data Infrastructures Workshop
Pittsburg, May 17-18

Dataverse and Related Projects

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

@mercecrosas



A data repository system for sharing and archiving research data

*A solution for publishing **FAIR** research data*

FAIR = Findable Accessible Interoperable Reusable

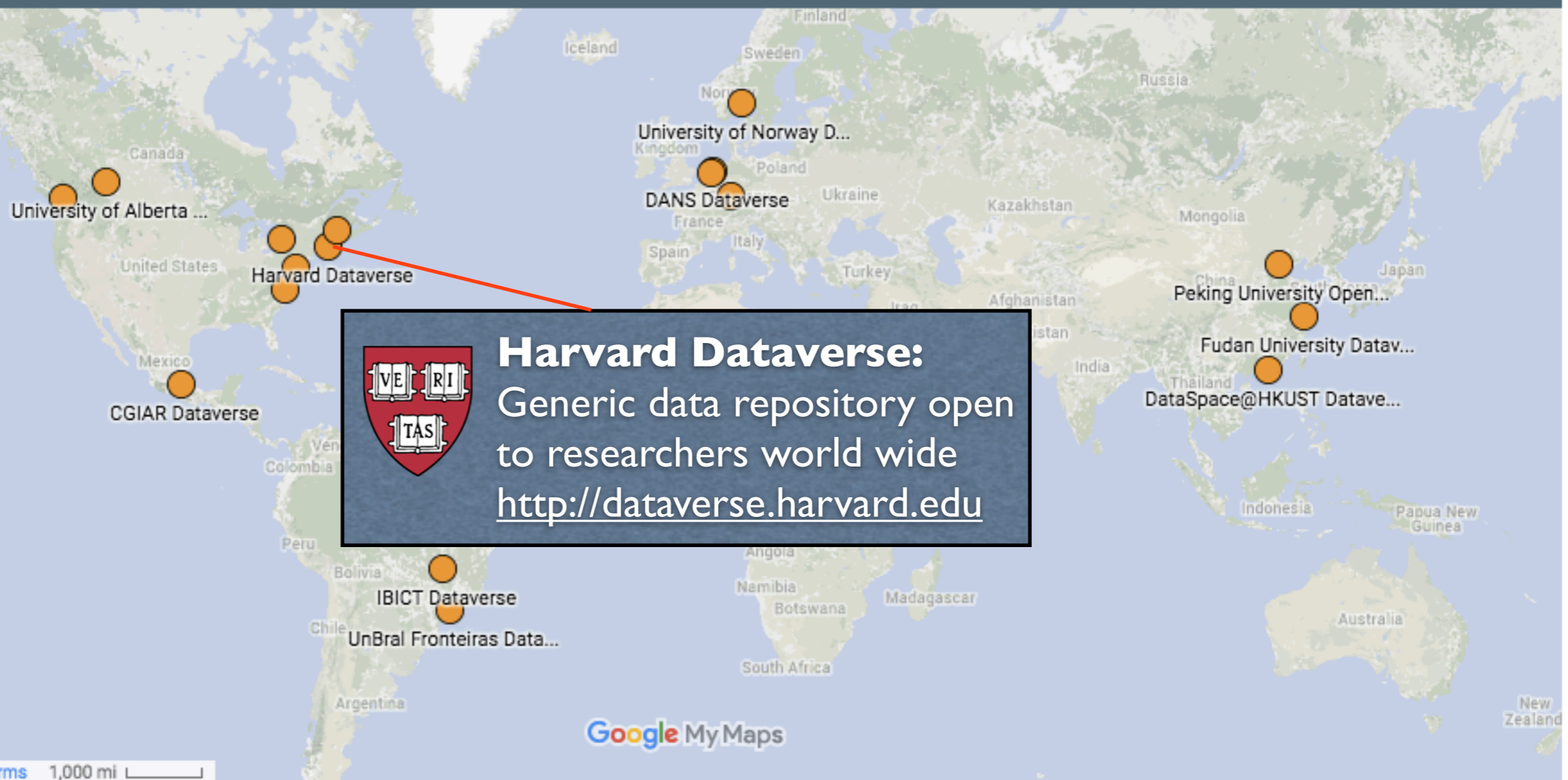
17 Installations

1,500+ Dataverses

65,000+ Datasets

1,700,000+ Downloads

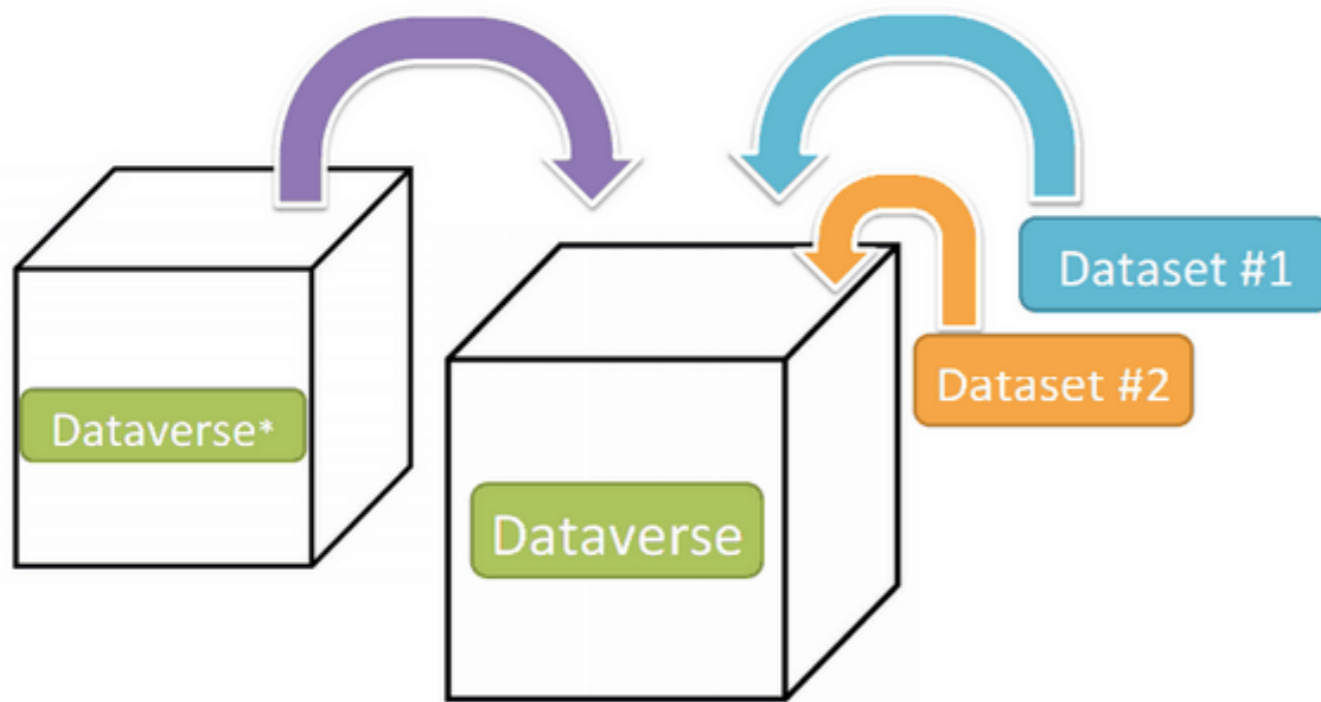
 Dataverse Repositories 



Dataverse repositories can serve a community, an institution, an archive, ...

Dataverses contain datasets and other dataverses, datasets contain metadata and data files

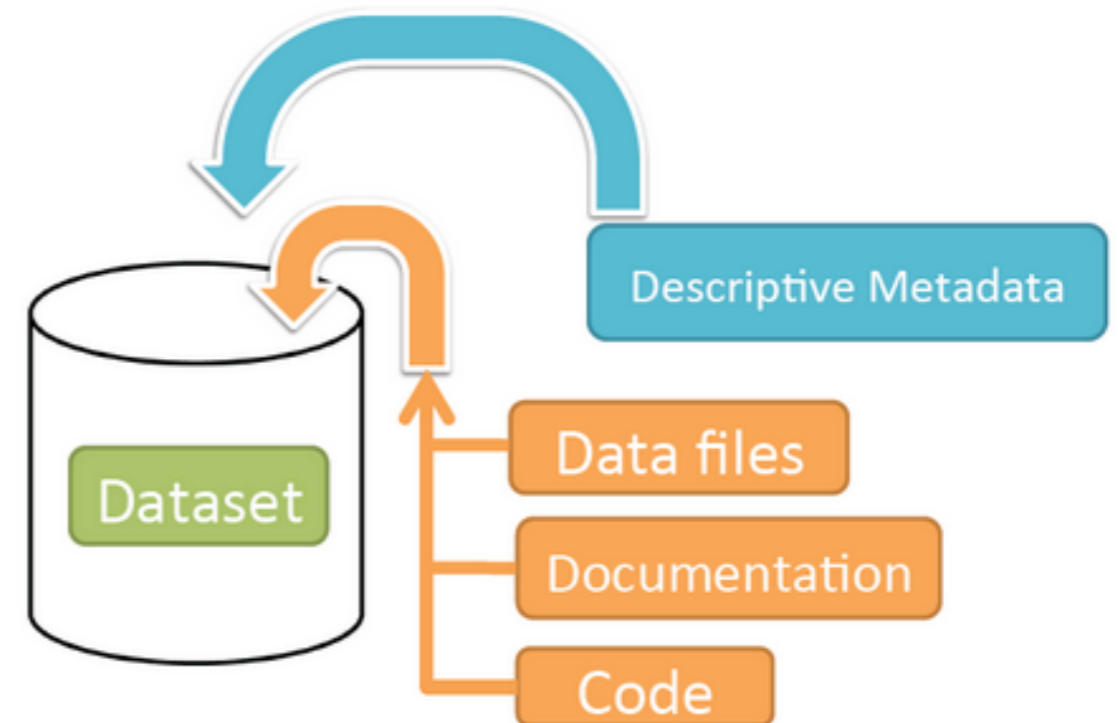
Schematic Diagram of a **Dataverse** in Dataverse 4.0



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.

[Home](#) ▶ [Comment](#) ▶ [Data Descriptor](#)SCIENTIFIC DATA | COMMENT **OPEN**

The FAIR Guiding Principles for scientific data management and stewardship

[Mark D. Wilkinson](#), [Michel Dumontier](#), [IJsbrand Jan Aalbersberg](#), [Gabrielle Appleton](#), [Myles Axton](#), [Arie Baak](#), [Niklas Blomberg](#), [Jan-Willem Boiten](#), [Luiz Bonino da Silva Santos](#), [Philip E. Bourne](#), [Jildau Bouwman](#), [Anthony J. Brookes](#), [Tim Clark](#), [Mercè Crosas](#), [Ingrid Dillo](#), [Olivier Dumon](#), [Scott Edmunds](#), [Chris T. Evelo](#), [Richard Finkers](#), [Alejandra Gonzalez-Beltran](#), [Alasdair J.G. Gray](#), [Paul Groth](#), [Carole Goble](#), [Jeffrey S. Grethe](#), [Jaap Heringa](#), [Peter A.C 't Hoen](#), [Rob Hooft](#), [Tobias Kuhn](#), [Ruben Kok](#), [Joost Kok](#), [Scott J. Lusher](#), [Maryann E. Martone](#), [Albert Mons](#), [Abel L. Packer](#), [Bengt Persson](#), [Philippe Rocca-Serra](#), [Marco Roos](#), [Rene van Schaik](#), [Susanna-Assunta Sansone](#), [Erik Schultes](#), [Thierry Sengstag](#), [Ted Slater](#), [George Strawn](#), [Morris A. Swertz](#), [Mark Thompson](#), [Johan van der Lei](#), [Erik van Mulligen](#), [Jan Velterop](#), [Andra Waagmeester](#), [Peter Wittenburg](#), [Katherine Wolstencroft](#), [Jun Zhao](#) & [Barend Mons](#) Show fewer authors

What is needed for *FAIR* Data Publishing?

Data Citation

- Persistent id to reference data uniquely
- Support for versions and fixity
- Attribution to authors and repository

Metadata

- Catalog to discover and locate the data
- Sufficient information to understand and reuse the data

Repository

- Digital access to metadata and data
- Archive and preservation for long-term access
- Interoperability through standards and APIs

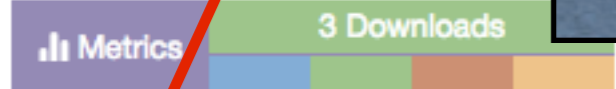
Data Citation in Dataverse

PKU Climate Dataverse

Authors

Published Year

Dataset Title



Evolution of East Asian summer and winter monsoons in the last 21,000 years

Wen, Xinyu, 2016, "Evolution of East Asian summer and winter monsoons in the last 21,000 years", <http://dx.doi.org/10.7910/DVN/BMAG9U>, Harvard Dataverse, V7

Download Citation

If you use these data, please add this citation to your scholarly resources. Learn about [Data Citation Standards](#).

Description

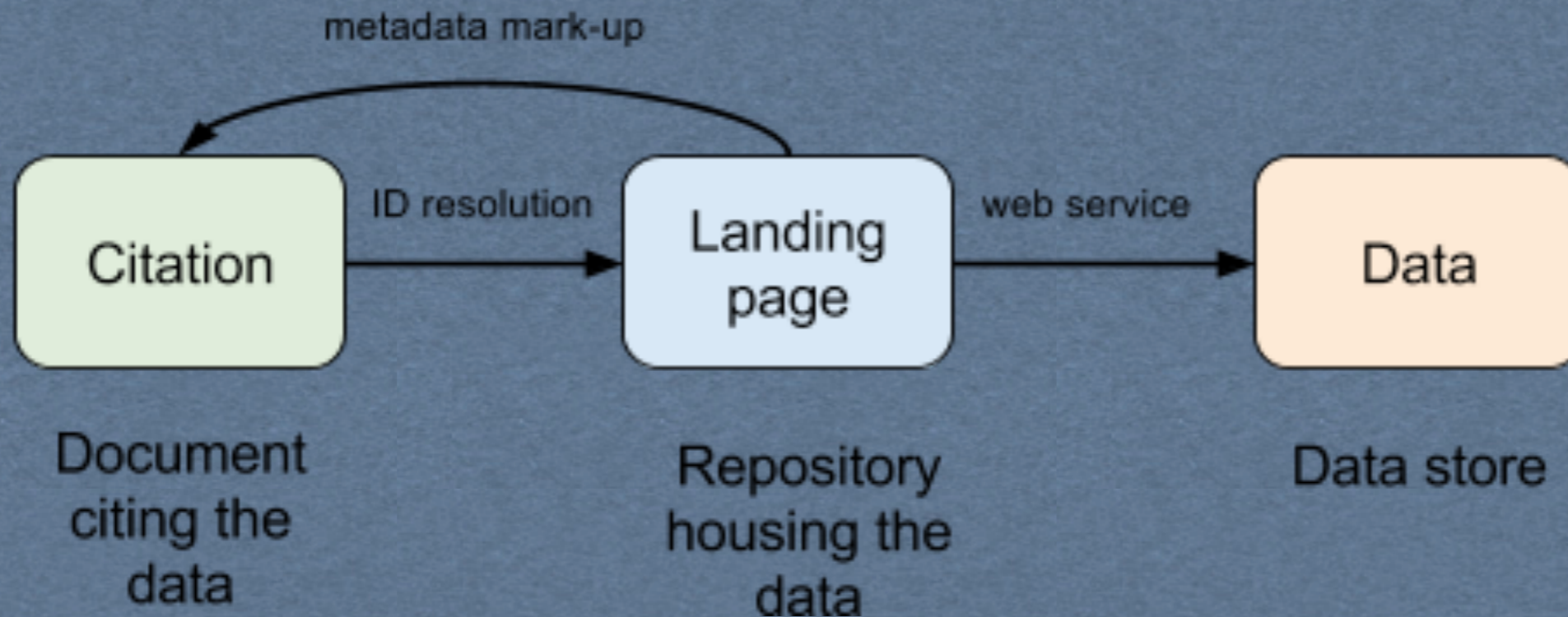
Global Persistent Identifier

Repository = Data Publisher

Version (or time range)

The data, scripts, and plots used in the paper entitled "Correlation and Anti-Correlation of the East Asian Summer Monsoon and the East Asian Winter Monsoon" (Journal of Climate Communications (2016)) are available here and make them available to the public. The data is available in the tarball of figure names.

Data Citation Basics



The dataset landing page is accessible and guaranteed by the repository (or data publisher), even when data are restricted or deaccessioned

Metadata in Dataverse

Metadata Level	Fields	Standards
Citation Metadata	author, title, repository, year published, version, etc	<ul style="list-style-type: none">• Dublin Core• DataCite
Domain-specific Metadata	data collection info (methods, organism, observation, survey, experiment, etc)	<ul style="list-style-type: none">• DDI (social sciences)• ISA-Tab BioCaddie (biomed)• Virtual Observatory (astro)• + <i>Custom metadata blocks</i>
File-level Metadata	metadata inside the data file (variables, instrument details, geospatial info, etc)	<ul style="list-style-type: none">• DDI (for variables),• + <i>more to be determined</i>

Information Extraction: Tabular Files

RData Stata SPSS Excel CSV		var 1	var 2	var 3
	obs 1	2	a	0
	obs 2	4	c	0
	obs 3	6	b	1
	obs 4	1	e	0
	obs 5	2	a	1
	obs 6	3	b	1

Variable Metadata:
Variable name, label,
type, stats, geospatial
coordinates

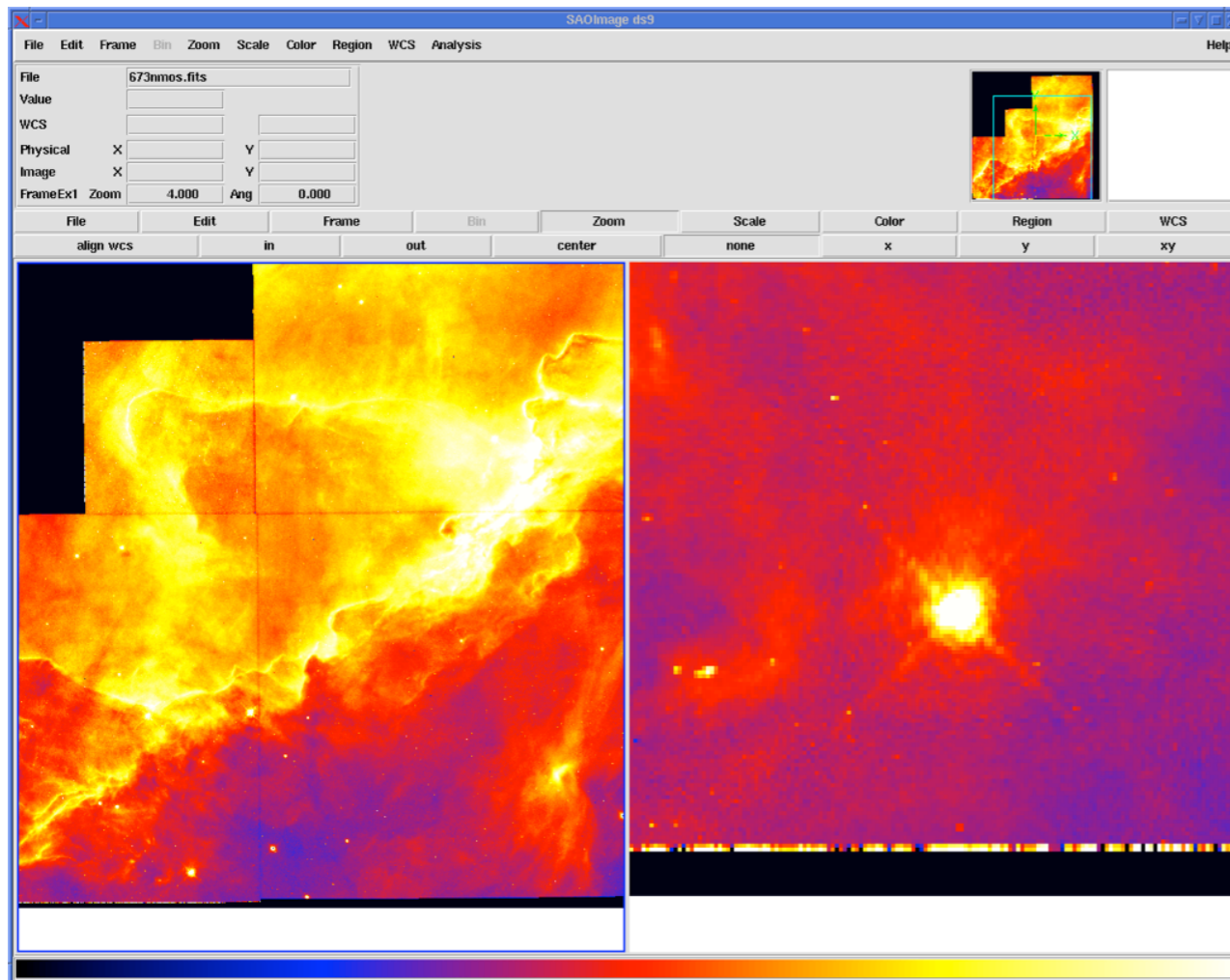
Data Values:
Independent of format

2	a	0
4	c	0
6	b	1
1	e	0
2	a	1
3	b	1

Universal Numerical Fingerprint (UNF):
checksum on data values, from canonical format

Information Extraction: FITS (astro) Files

Header Metadata:
coordinates (R.A.,
declination),
photometric info, ...



Data Objects:

- Image Files
- Spectra
- Data cubes
- Tables
- ...

In addition to data citation and metadata features, Dataverse has a **rich set of features** to support data publishing

Tiered Access

Metadata

Files

How to Access

Open (default): CC0	Open	Open	Click to Download
GuestBook	Open	Open	Fill in guestbook before download
Terms of Use	Open	Open	Click through terms of use before download
Data Restricted	Open	Restricted	Request Access via click through
Data Restricted	Open	Restricted	Request Access via application

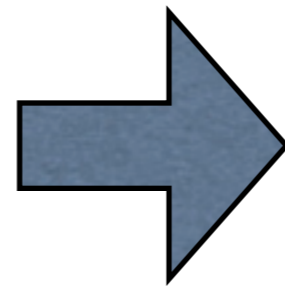
Data Publishing Workflows

Create Dataset
(landing page
restricted)

Review
(collaborators or
anonymous reviewers)

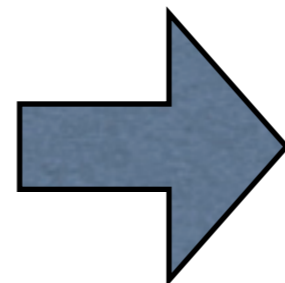
Publish v. 1

Minor change
(metadata only)



Publish v. 1.1

Major change
(might include
new data file)



Publish v. 2

Learn more at dataverse.org guides

User Guide

Installation Guide

API Guide

SWORD API

Search API

Data Access API

Native API

Client Libraries

Apps

Developer Guide

API Guide

We encourage anyone interested in building tools to interoperate with the Dataverse to utilize our APIs. In 4.0, we require to get a token, by simply registering for a Dataverse account, before using our APIs (We are considering making some of the APIs completely public in the future - no token required - if you use it only a few times).

Rather than using a production installation of Dataverse, API users should use <http://apitest.dataverse.org> for testing.

Contents:

- [SWORD API](#)
 - [Backward incompatible changes](#)
 - [New features as of v1.1](#)
 - [curl examples](#)
 - [Retrieve SWORD service document](#)
 - [Create a dataset with an Atom entry](#)
 - [Dublin Core Terms \(DC Terms\) Qualified Mapping - Dataverse DB Element Crosswalk](#)
 - [List datasets in a dataverse](#)
 - [Add files to a dataset with a zip file](#)
 - [Display a dataset atom entry](#)
 - [Display a dataset statement](#)
 - [Delete a file by database id](#)
 - [Replacing metadata for a dataset](#)
 - [Delete a dataset](#)
 - [Determine if a dataverse has been published](#)
 - [Publish a dataverse](#)
 - [Publish a dataset](#)

Biomedical Dataverse: Supporting Large-Scale Datasets



Data publication with the structural biology data grid supports live analysis

Peter A. Meyer, Stephanie Socias, Jason Key, Elizabeth Ransey, Emily C. Tjon, Alejandro Buschiazzi, Ming Lei, Chris Botka, James Withrow, David Neau, Kanagalaghatta Rajashankar, Karen S. Anderson, Richard H. Baxter, Stephen C. Blacklow, Titus J. Boggon, Alexandre M. J. J. Bonvin, Dominika Borek, Tom J. Brett, Amedeo Caflisch, Chung-I Chang, Walter J. Chazin, Kevin D. Corbett, Michael S. Cosgrove, Sean Crosson, Sirano Dhe-Paganon, Enrico Di Cera, Catherine L. Drennan, Michael J. Eck, Brandt F. Eichman, Qing R. Fan, Adrian R. Ferré-D'Amaré, J. Christopher Fromme, K. Christopher Garcia, Rachelle Gaudet, Peng Gong, Stephen C. Harrison, Ekaterina E. Heldwein, Zongchao Jia, Robert J. Keenan, Andrew C. Kruse, Marc Kvansakul, Jason S. McLellan, Yorgo Modis, Yunsun Nam, Zbyszek Otwinowski, Emil F. Pai, Pedro José Barbosa Pereira, Carlo Petosa, C. S. Raman, Tom A. Rapoport, Antonina Roll-Mecak, Michael K. Rosen, Gabby Rudenko, Joseph Schlessinger, Thomas U. Schwartz, Yousif Shamoo, Holger Sondermann, Yizhi J. Tao, Niraj H. Tolia, Oleg V. Tsodikov, Kenneth D. Westover, Hao Wu, Ian Foster, James S. Fraser, Filipe R. N. C. Maia, Tamir Gonen, Tom Kirchhausen, Kay Diederichs, **Mercè Crosas & Piotr Sliz**

Show fewer authors

[Affiliations](#) | [Contributions](#) | [Corresponding author](#)

Nature Communications 7, Article number: 10882 | doi:10.1038/ncomms10882

Received 16 October 2015 | Accepted 28 January 2016 | Published 07 March 2016

The Biomedical Dataverse at Harvard Medical School - also tested as a persistent repository for LINCS data (NIH Library of Integrated Network based Cellular Signatures)

The image shows two screenshots of the Dataverse web interface. The top screenshot displays the 'Laboratory of Systems Pharmacology Dataverse' page, which is currently unpublished. It features the logo of the Laboratory of Systems Pharmacology and a URL: <http://hits.harvard.edu/the-program/laboratory-of-systems-pharmacology/>. The bottom screenshot shows the 'HMS LINCS Center Dataverse' page, also unpublished, with the LINCS logo and Harvard Medical School crest. The URL is <http://incs.hms.harvard.edu>. The page includes a search bar, navigation links, and a message stating: 'This dataverse currently has no dataverses, datasets, or files. You can add to it by using the Add Data button on this page.' A sidebar on the left of the bottom screenshot shows filters for 'Dataverse', 'Dataset', and 'Files (0)', along with 'Publication Status' (Unpublished (2), Draft (1)) and 'Subject' categories like 'Computer and In', 'Mathematical Sci', and 'Medicine, Health'.

Collaboration with Piotr Sliz and Caroline Shamu (HMS)

Data Access Alliance

- A **National Data Service Pilot** project to replicate the structural biology data in the biomedical Dataverse to multiple sites:
 - Initial sites: SDSC, Argonne (Petrel)
 - Transfer data across Globus endpoints
 - Data close to computation resources (XSEDE)

**An Additional Challenge
for Data Publishing:
Sensitive Data**

The DataTags System



Technology Science

...how technology impacts humans. Home



Tweet

Published 2015-10-16. Views 3,490. Downloads 408. Suggestions 0.

Sharing Sensitive Data with Confidence: The Datatags System

Latanya Sweeney, Mercè Crosas, and Michael Bar-Sinai

Abstract

Introduction

Background

Methods

Results

Discussion

References

Download

Authors

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

Definitions for each of six ordered Blue to Crimson sample datatags.

- We introduce datatags as a means of specifying security and access requirements for sensitive data
- The datatags approach reduces the complexity of thousands of data-sharing regulations to a small number of tags
- We show implementation details for medical and educational data and for research and corporate repositories

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>

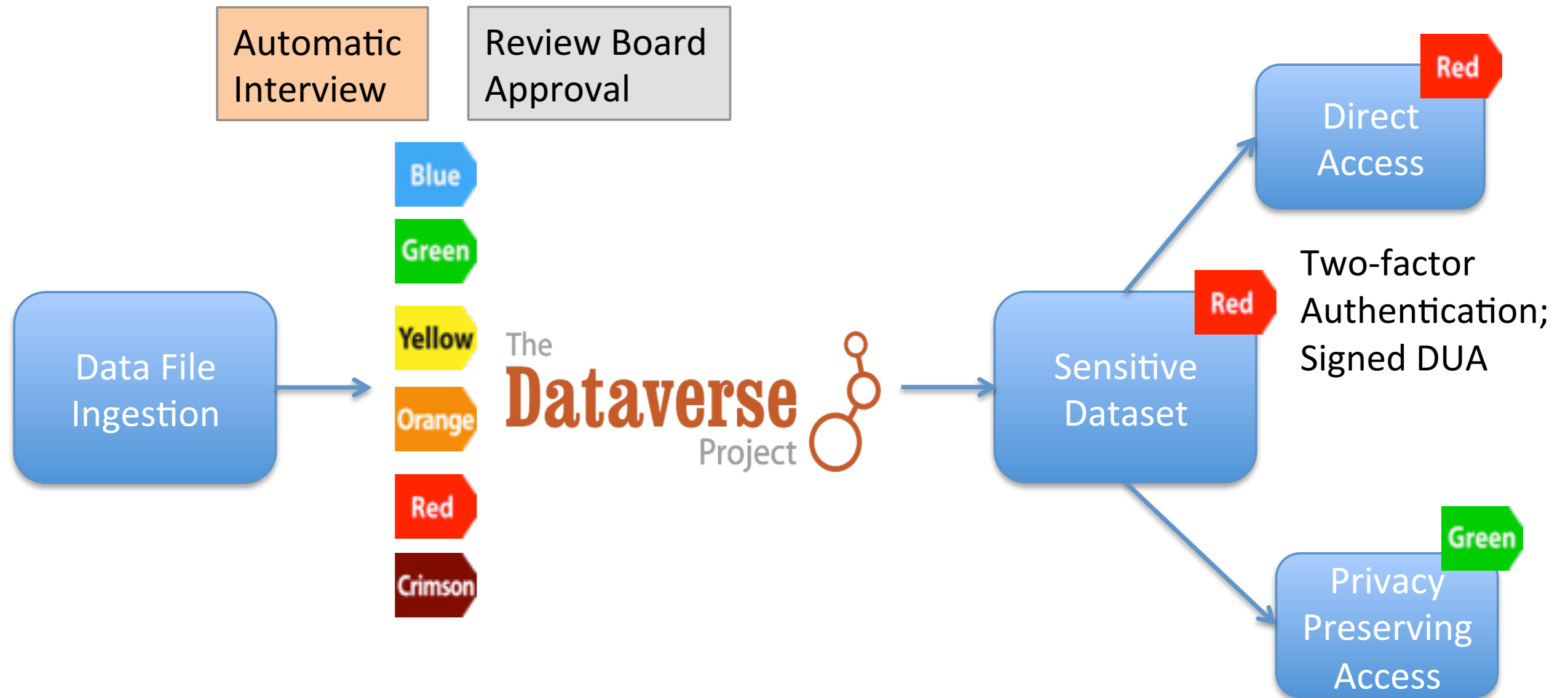
A **datatag** is a set of security features and access requirements for file handling.

A **datatags repository** is one that stores and shares data files in accordance with a standardized and ordered levels of security

Datatags Levels

Tag Type	Description	Security Features	Access Requirements
Blue	Public	Clear storage Clear transmission	Open
Green	Controlled public	Clear storage Clear transmission	Email, OAuth verified registration
Yellow	Accountable	Clear storage Encrypted transmit	Password, Registered , Approval, Click 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	MultiEncrypt store Encrypted transmit	Two-factor authentication, Approval, Signed DUA

DataTags Workflow in a Dataverse Repository (under development)



<http://datatags.org>

<http://privacytools.seas.harvard.edu>

Example of DataTags Interview: A sequence of questions from an expert-based system

The image displays three overlapping browser windows from the DataTags website, illustrating a sequence of questions from an expert-based system. Each window shows a question, an answer feed, and an engine trace.

Window 1 (Top): The question is "Do the data concern living people?". The answer feed shows "Yes" and "No" buttons. The engine trace shows a table with one entry:

Id	Type
\$0	ask

Window 2 (Middle): The question is "Do the data contain health information?". The answer feed shows "Yes" and "No" buttons. The engine trace shows a table with one entry:

Id	Type
medicalRecordsCompliance	

Window 3 (Bottom): The question is "Do the data contain information from a covered entity or business associate of a covered entity?". The answer feed shows "Yes", "Not Sure", and "No" buttons. The engine trace shows a table with four entries:

Id	Type
medicalRecordsCompliance	
\$4	
\$3	
\$0	

The detailed question in the bottom window includes definitions for "Business associate" and "Covered entity".

Business associate
A business associate is any person or organization, including a subcontractor, that acts on behalf of, or provides services to, a covered entity involving the use or disclosure of protected health information. This includes, but is not limited to, legal, actuarial, accounting, consulting, claim processing, data analysis, administration, utilization review, quality assurance, billing, benefit management, practice management, and re-pricing activities.

Covered entity
A covered entity is a health plan, health care clearinghouse, or health care provider that transmits any health information in electronic form.

- Health plans include health insurance companies, health maintenance organizations [HMOs], company health plans, and government programs that pay for health care, such as Medicare, Medicaid, and the military and veterans health care programs.
- Health care providers include doctors, clinics, psychologists, dentists, chiropractors, nursing homes, and pharmacies.
- Health care clearinghouses include entities that process nonstandard health information they receive from another entity into a standard, i.e. standard electronic format or data content, or vice versa.

The bottom window also shows an "Answer Feed" with a question about substance abuse diagnosis, referral, or treatment, and a "Current Tags" section showing a tag with the code "green".

Example of DataTags Interview: Final datatag human--readable and machine-actionable policy

www.datatags.org/interviews/questionnaireId/accept

DataTags Feedback

Dataset Can be Accepted

Your dataset is tagged as **Orange**

May include sensitive, identifiable personal information, shared with verified and/or approved recipients under agreement.

DataTags

Legal

- MedicalRecords
 - HIPAA **safeHarborDeidentified**
- EducationRecords
 - PPRA **protectedDeidentified** **consent**
- ContractOrPolicy **no**
- GovernmentRecords
 - DPPA **highlyRestricted**
- Code **orange**

Assertions

Summary

- **Dataverse** is an open-source software for building data repositories
- **Harvard Dataverse** is a generic data repositories, hosted at Harvard and open to all researcher world-wide
- **Data citation** and **rich metadata** support are key to Dataverse, and enable FAIR data publishing
- Dataverse also supports **tiered access** to data and data publishing **review and versioning workflows**
- Biomedical Dataverse is implementing **large-scale datasets** support for Dataverse
- DataTags generates human-readable and machine-actionable policies to support **sensitive datasets** in data repositories

Join us to this year's Dataverse Community Meeting

Dataverse Community Meeting 2016
July 11, 12, 13 at Harvard Medical School




[Location](#) [Meeting Agenda](#) [Privacy Workshop](#) [Registration](#) [Lodging](#)

Dataverse2016: Fostering the Dataverse Community

After a successful [first Dataverse Community Meeting](#) last year where we worked together to help define who and what makes up our community, this year's meeting (July 11-12) will focus on working together with stakeholders, users and contributors to continue *fostering the Dataverse Community* and its impact in the world of data sharing and archiving. We welcome researchers, librarians, archivists, publishers, funders, software developers and anyone interested in data repositories.

Tweets about #dataverse2016

 [dataverseorg](#) Registration for this year's Dataverse Community Meeting & Privacy Workshop is now open! t.co/qk1cYxOJ9k #dataverse2016
6 days 9 hours ago.

 [dataverseorg](#) Registration for this

References

@mercecrosas

<http://dataverse.org>

<http://dataverse.harvard.edu>

<http://datatags.org>

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

Altman, Borgman, Crosas, Martone, 2015, An Introduction to the Joint Data Citation Principles, Bulletin of the Association for Information Science and Technology

Starr et al, 2015, *Achieving Human and Machine Accessibility of Cited Data in Scholarly Publications*, PeerJ Computer Science

Meyer et al, 2016, *Data Publication with the Structural Biology Grid Supports Live Analysis*, Nature Communications

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