NDSR 2016 Symposium

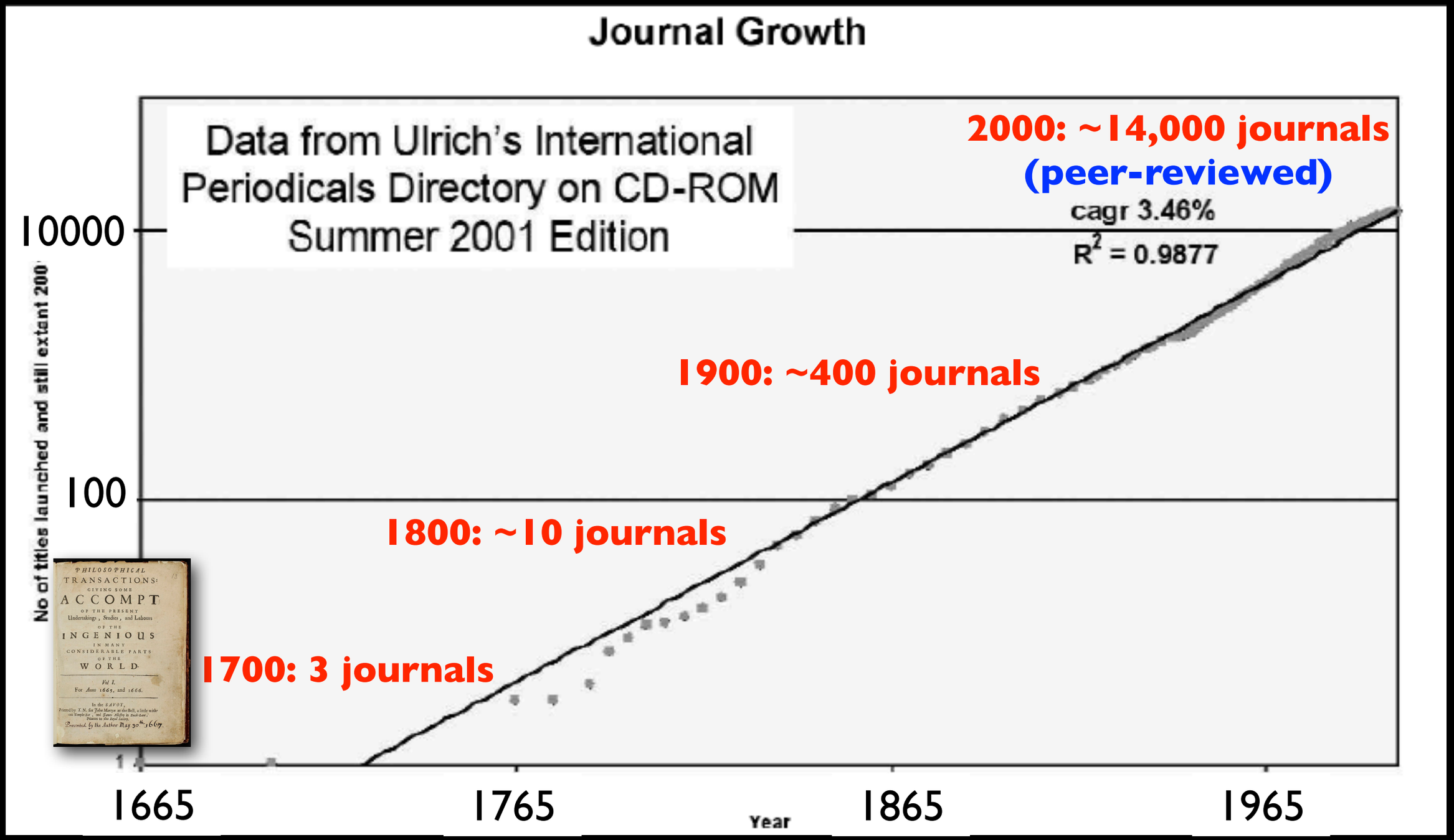
The Rise of Data Publishing in the Digital World (and how Dataverse and DataTags help)

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

@mercecrosas

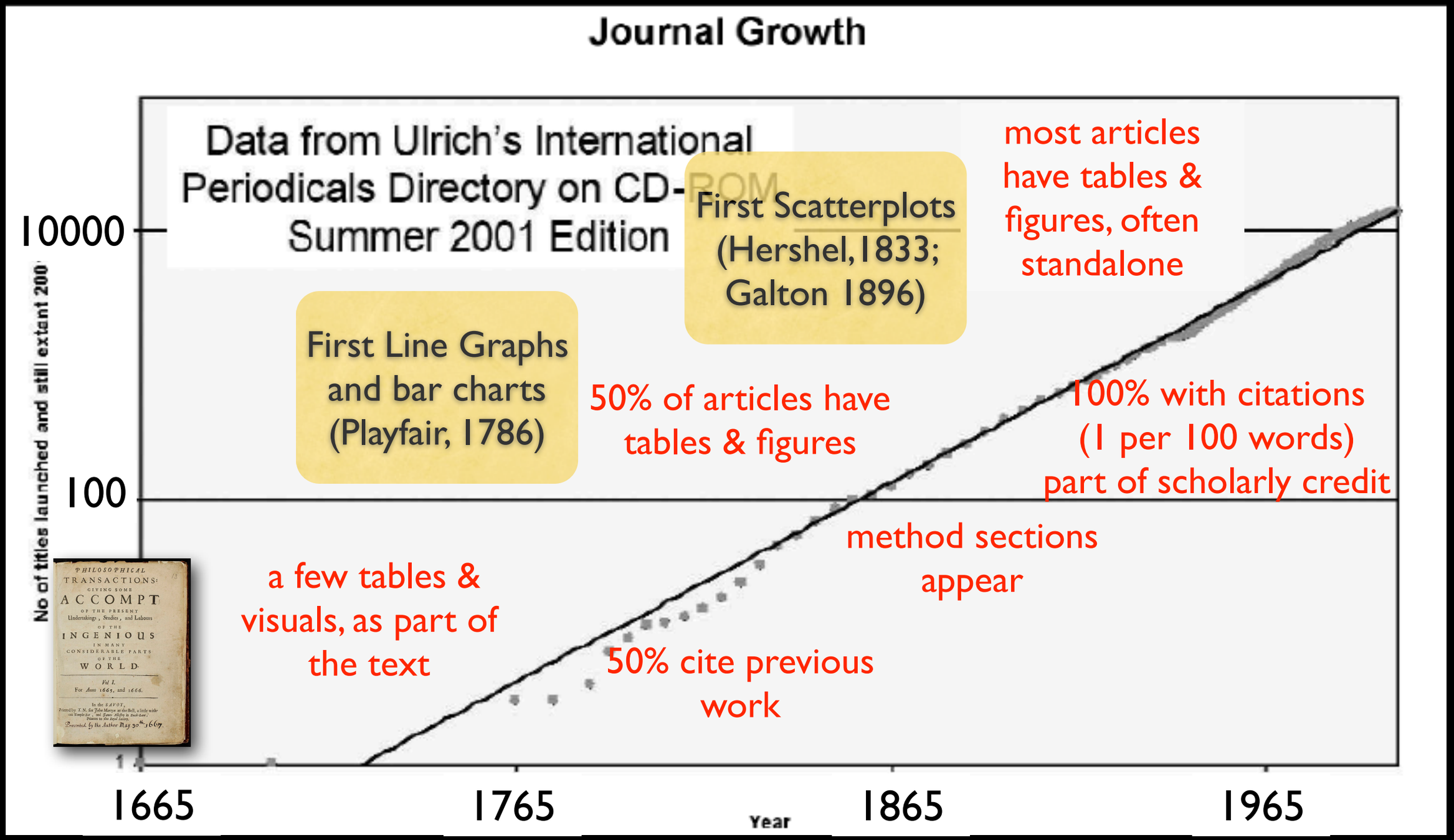
From 1665 to late 20th century:
A steady increase in size and complexity of research output

The number of journals doubles every 20 years since 1750s, with growth of number of scientists



Mabe, 2003

Data Tables and Visuals Become Increasingly Common, and part of the Scientific Argument



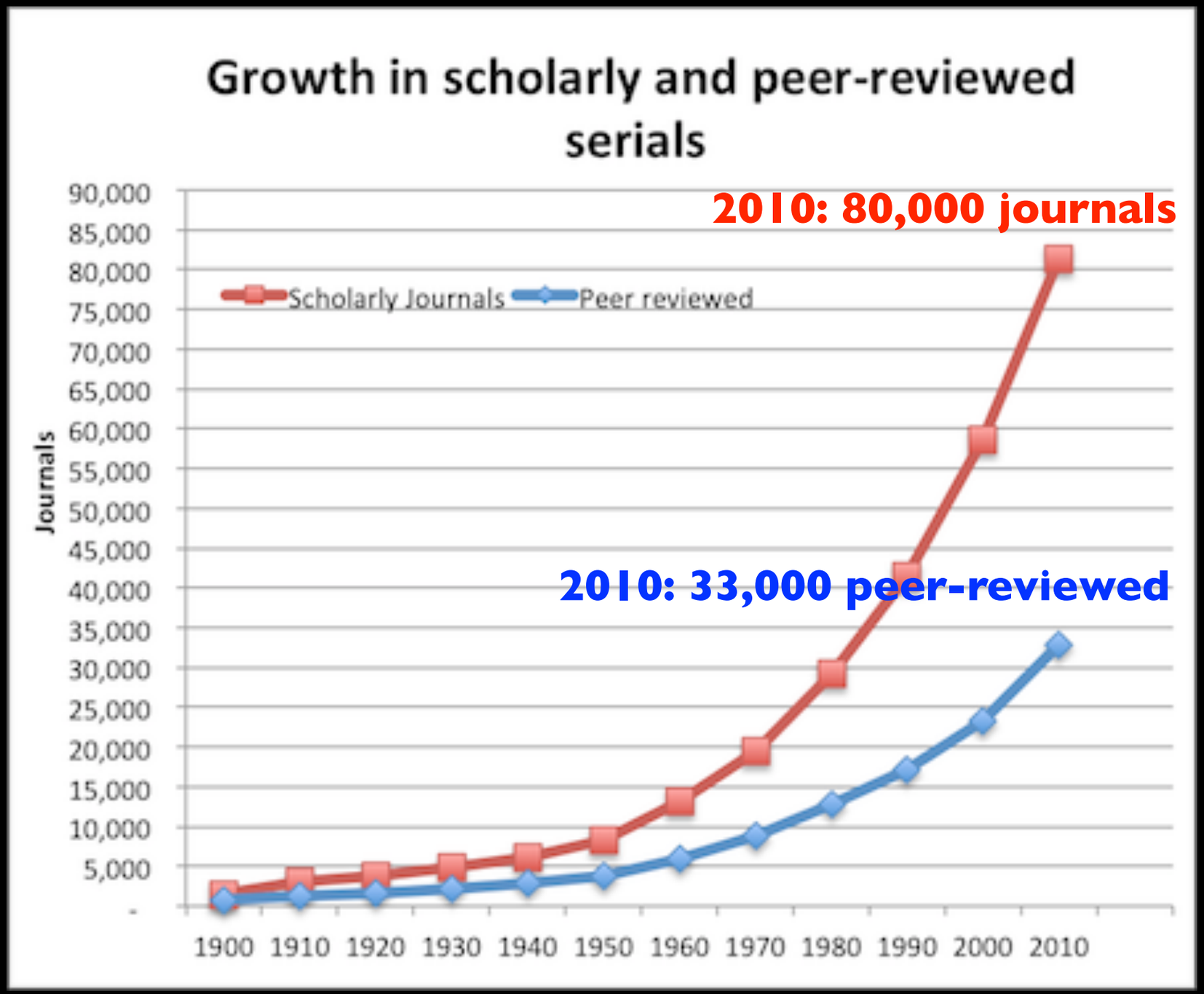
Scholarly Publishing Adapts to the Increase of Cognitive Complexity (Gross et al 2001)

- 18th century:
  - formal components appear in articles (introduction, conclusions, table, figures, citations)
- 19th century:
  - explain data instead of establish observations of facts
  - wide use of visuals, high citation density, methods section
- 20th century:
  - structured quantitative data with increased use of statistics
  - wide range of data types with new technologies
- Number of scientists increases from 100s to a few millions
- Science becomes extremely specialized:
  - from 1 journal to 14,000 peer-reviewed journals
  - one new journal for each 150 authors, read by 500

In the last decades, more and more publications and data

A Steeper Growth of Scholarly Output

Since 1950, the total number of journals doubles every ~15 years



An Outburst of Research Data and Specialization, Results into > 1000 Community Repositories

1920 - 1950s

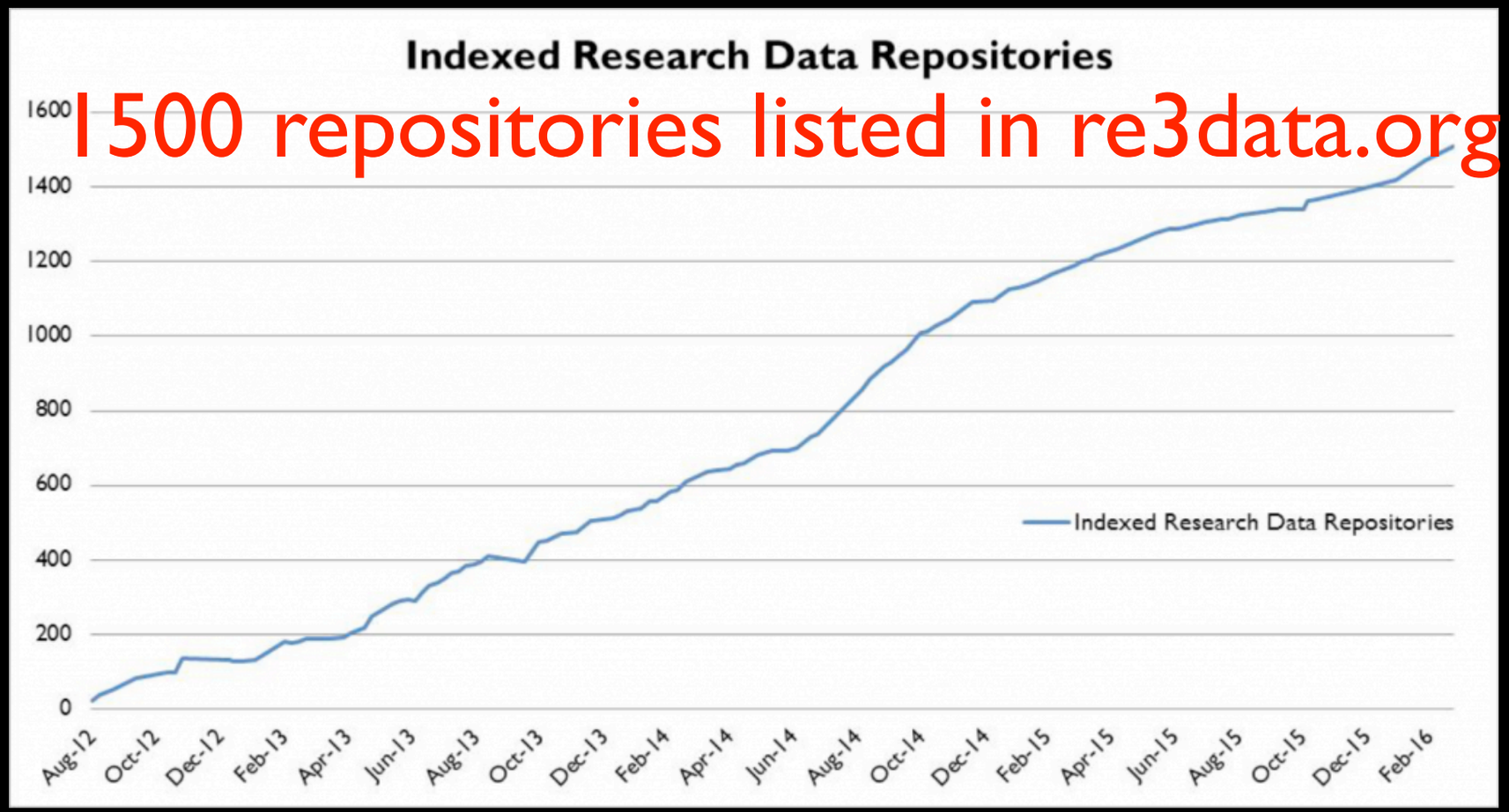
First Social Science Data Archives (ODUM, ICPSR, ...)

1970 - 1980s

First Biomedical Databases (PDB, GenBank, ...)

2016

A wide range of Research Data Repositories



Data Publishing Emerges as the Union of Scholarly Publishing and Data Archiving

Scholarly publishing:
Distribute research output

- Attribution and credit
- Dissemination
- Finding & Reuse

Data Archiving:
Long-term access to data

- Accessibility
- Preservation
- Finding & Reuse

Why Data Publishing now?

Extending Gross et al. thesis, *data publishing* accommodates the complexity of research input and output in the digital world.

- Data (and software) have become common input and output of research
- A scholarly article cannot hold or describe accurately these vast amounts of data and software
- As input and output of research, data must be citable and accessible to enable validation and reuse, with attribution

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

FAIR = Findable Accessible Interoperable Reusable

SCIENTIFIC DATA

Home | Archive | About | For Authors | For Referees | Data Policies | Collections

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

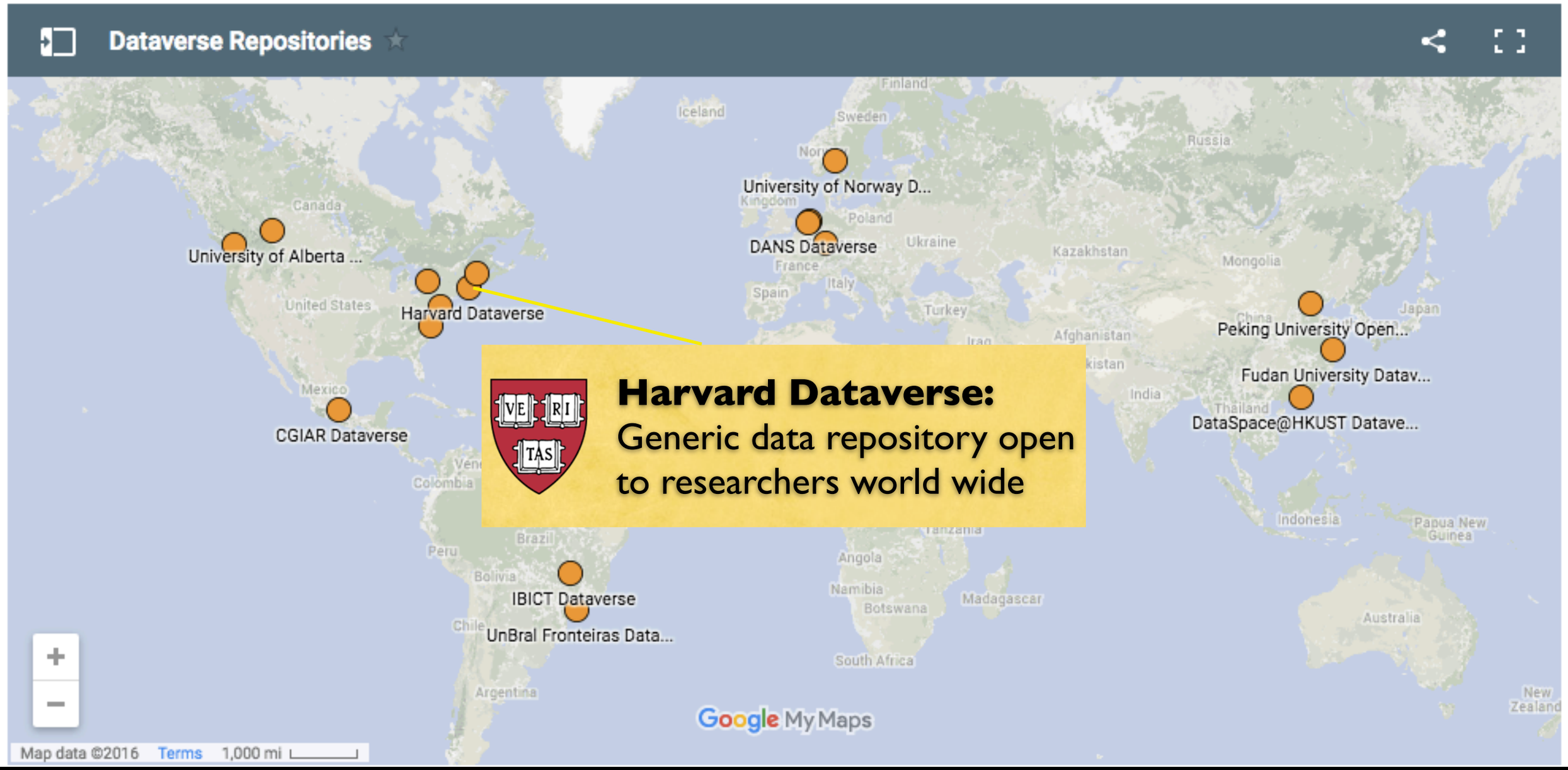
The **Dataverse** Project

A data repository system that serves as a solution for publishing FAIR research data

The Dataverse Project

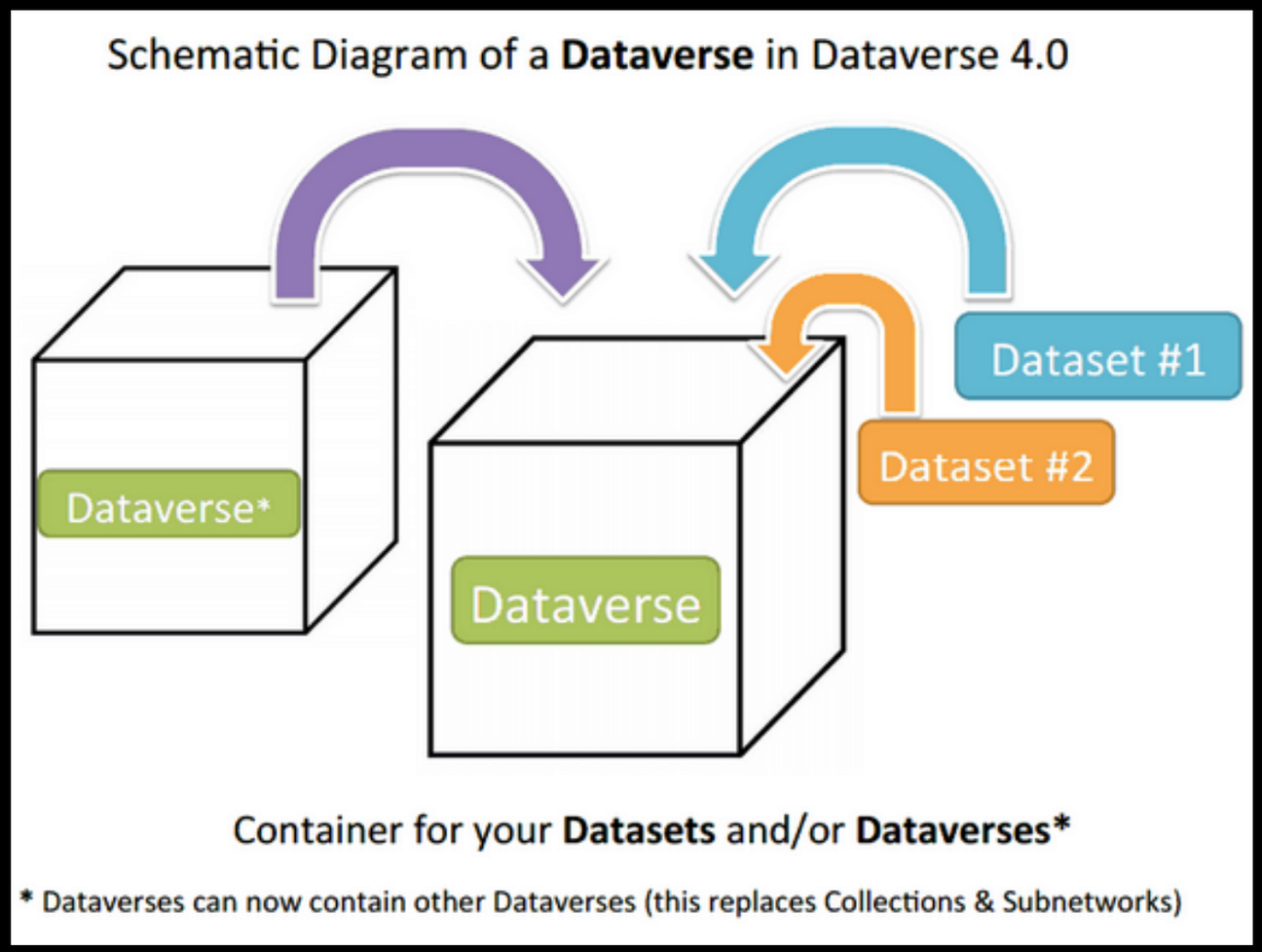
Around the World

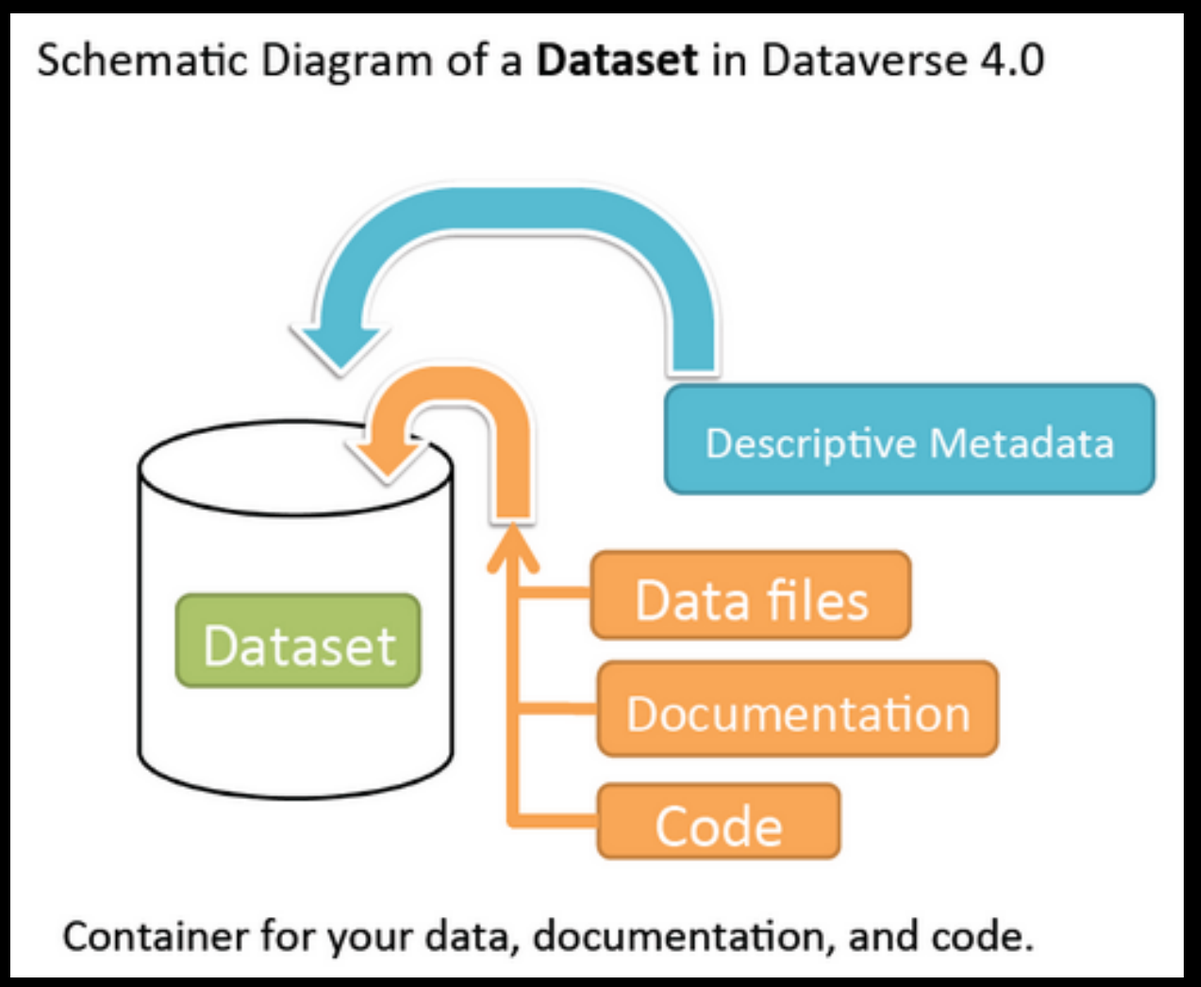
17 Installations | 1,500+ Dataverses | 65,000+ Datasets | 1,700,000+ Downloads



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

Dataverses contain datasets, datasets contain metadata and data files





Data Citation in Dataverse

Dataverse

About Guides Support Sign Up Log In

PKU Climate Dataverse (Peking University)

Authors

Published Year

Dataset Title

H Climate D f East Asia nonsoons in the last 21,000 years

Metrics 3 Downloads

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 The data, scripts, and plots used in the paper entitled "Correlation and Anti-Correlation of the East Asian ng the L ed to Nature Communications (2016) uld like puter code here and make them y. The in ata are packed in the tarball of nd plotti r those figure names.

Subjec

Keywo Monsoon, EASM, EAWM

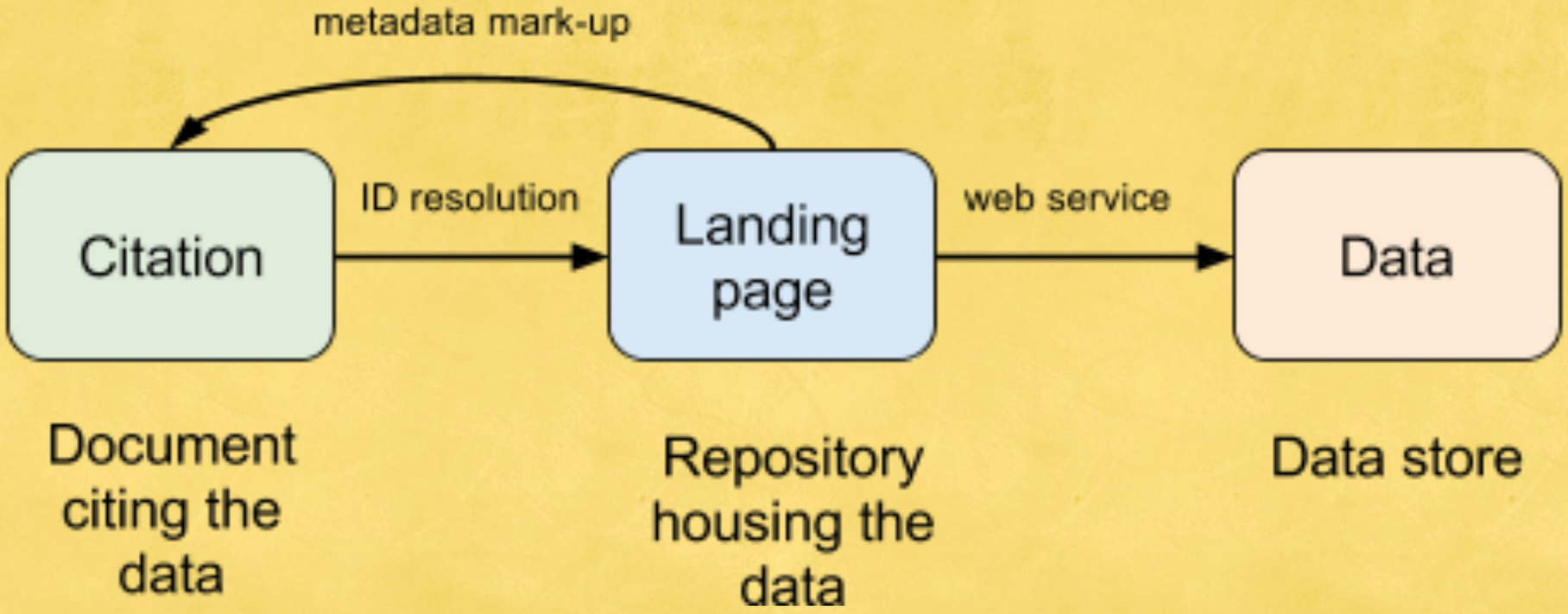
Global Persistent Identifier

Repository = Data Publisher

Version (or time range)

Files Metadata Terms Versions

Data Citation Basics



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

Force11, Joint Declaration of Data Citation Principles; Starr et al, 2015

Metadata In Dataverse

Metadata Level	Fields	Standards
Citation Metadata	author, title, repository, year published, version, etc	• Dublin Core • DataCite
Domain-specific Metadata	data collection info (methods, organism, observation, survey, experiment, etc)	• DDI (social sciences) • ISA-Tab BioCaddie (biomed) • Virtual Observatory (astro) • *+ Custom metadata blocks*
File-level Metadata	metadata inside the data file (variables, instrument details, geospatial info, etc)	• DDI (for variables), • *+ more to be determined*

Dataverse JSON Schema

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

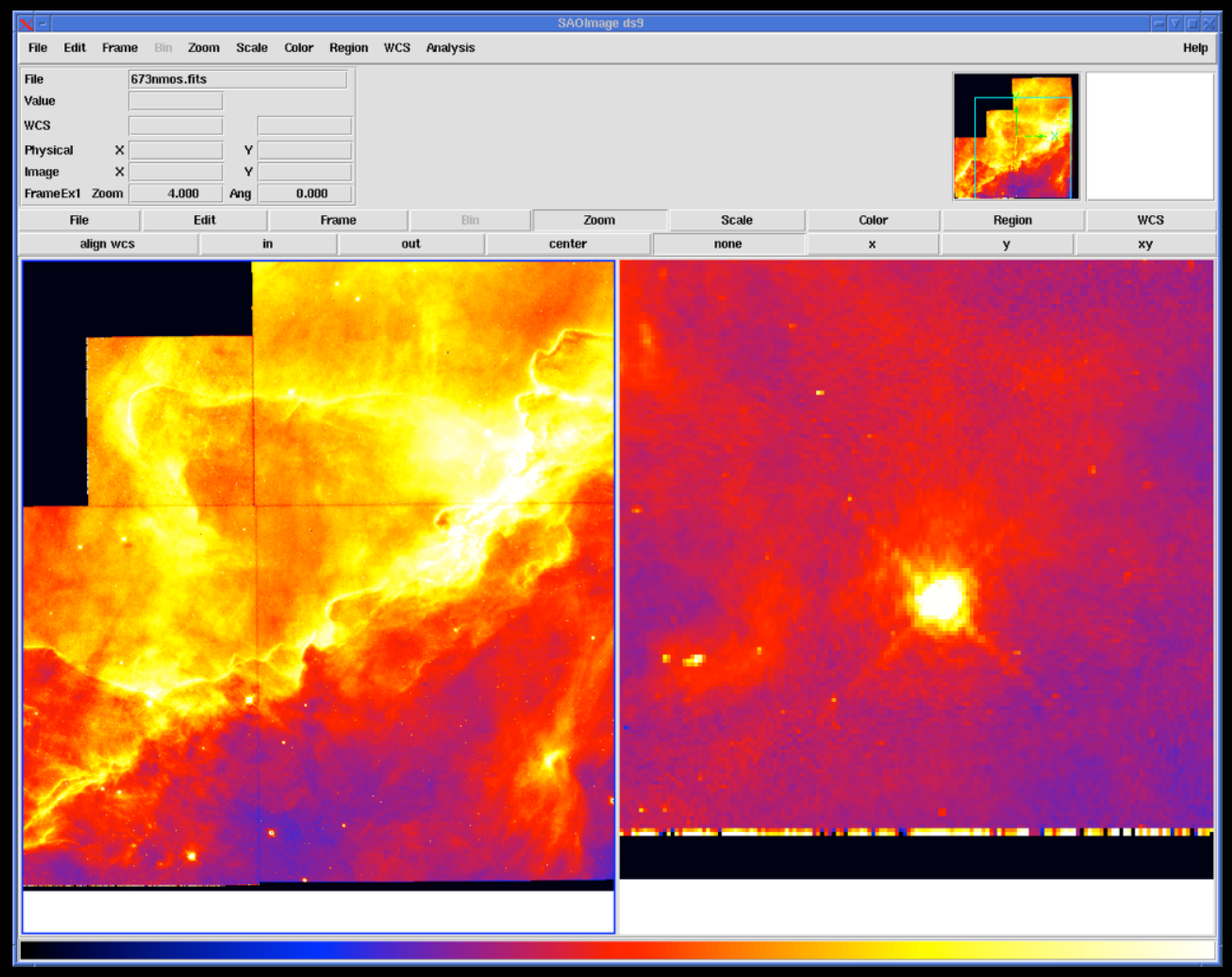
Variable Metadata:
Variable name, label, type, stats, geospatial coordinates

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 that facilitate 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

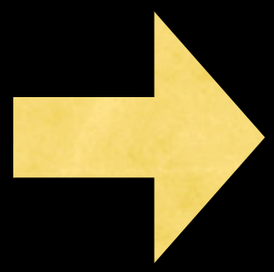
Data Publishing Workflows

Create Dataset (landing page restricted)

Review (collaborators or anonymous reviewers)

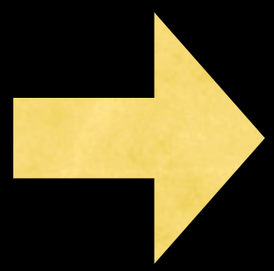
Publish v. 1

Minor change (metadata only)



Publish v. 1.1

Major change (might include new data file)



Publish v. 2

And more at dataverse.org guides ...

Search

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:

Biomedical Dataverse addresses data publication of large files: SBGridData

NATURE COMMUNICATIONS | ARTICLE OPEN

Data publication with the structural biology data grid supports live analysis

Peter A. Meyer, Stephanie Socias, Jason Key, Elizabeth Ransey, Emily C. Tjon, Alejandro Buschiazzo, 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

The Biomedical Dataverse at Harvard Medical School - also tested as a persistent repository for LINCS data

(NIH Library of Integrated Network-based Cellular Signatures)

Dataverse

About Guides Support Liz Williams

Laboratory of Systems Pharmacology

Laboratory of Systems Pharmacology Dataverse (Harvard Medical School) Unpublished

http://hits.harvard.edu/the-program/laboratory-of-systems-pharmacology/

Harvard Dataverse > **Laboratory of Systems Pharmacology Dataverse**

Dataverse

About Guides Support Liz Williams

LINCS HARVARD MEDICAL SCHOOL

HMS LINCS Center Dataverse (Harvard Medical School) Unpublished http://lincs.hms.harvard.edu

Harvard Dataverse > Laboratory of Systems Pharmacology Dataverse > **HMS LINCS Center Dataverse**

Publish Edit

This is a repository for biomedical research data generated by scientists affiliated with the Harvard Medical School LINCS Center, which is funded by NIH grant U54 HL127365 and is part of the NIH Library of Integrated Network-based Cellular Signatures (LINCS) Program. Further information about the HMS LINCS website.

Search this dataverse... Find Advanced Search Add Data

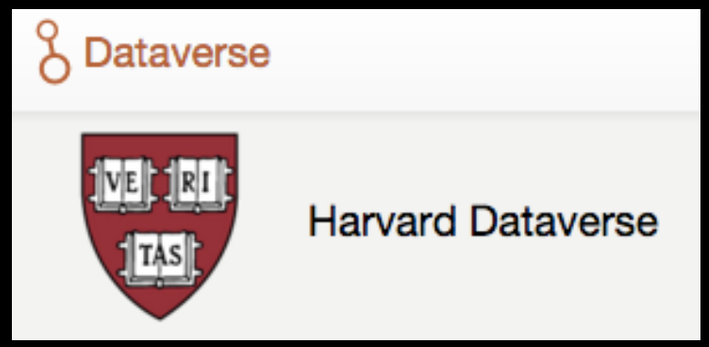
Dataverses (0)
Datasets (0)
Files (0)

This dataverse currently has no dataverses, datasets, or files. You can add to it by using the Add Data button on this page.

This is a reposit
in Therapeutic S
Search this d
Dataver
Dataset
Files (0)
Publication Sta
Unpublished (2)
Draft (1)
Dataverse Cat
Research Project
Subject
Computer and In
Mathematical Sci
Medicine, Health

Collaboration with Piotr Sliz and Caroline Shamu (HMS)

An additional challenge for data publishing: Sensitive Data



“User Uploads must be void of all identifiable information, such that re-identification of any subjects from the amalgamation of the information available from all of the materials (across datasets and dataverses) uploaded under any one author and/or user should not be possible.”

“Submitter represents and warrants that the Content does not contain any information (i) which identifies, or which can be used in conjunction with other publicly available information to personally identify, any individual;”

GenBank

"If you are submitting human sequences to GenBank, do not include any data that could reveal the personal identity of the source. It is our assumption that you have received any necessary informed consent authorizations that your organizations require prior to submitting your sequences."

How can we maximize publishing sensitive data while being mindful of privacy?

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

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

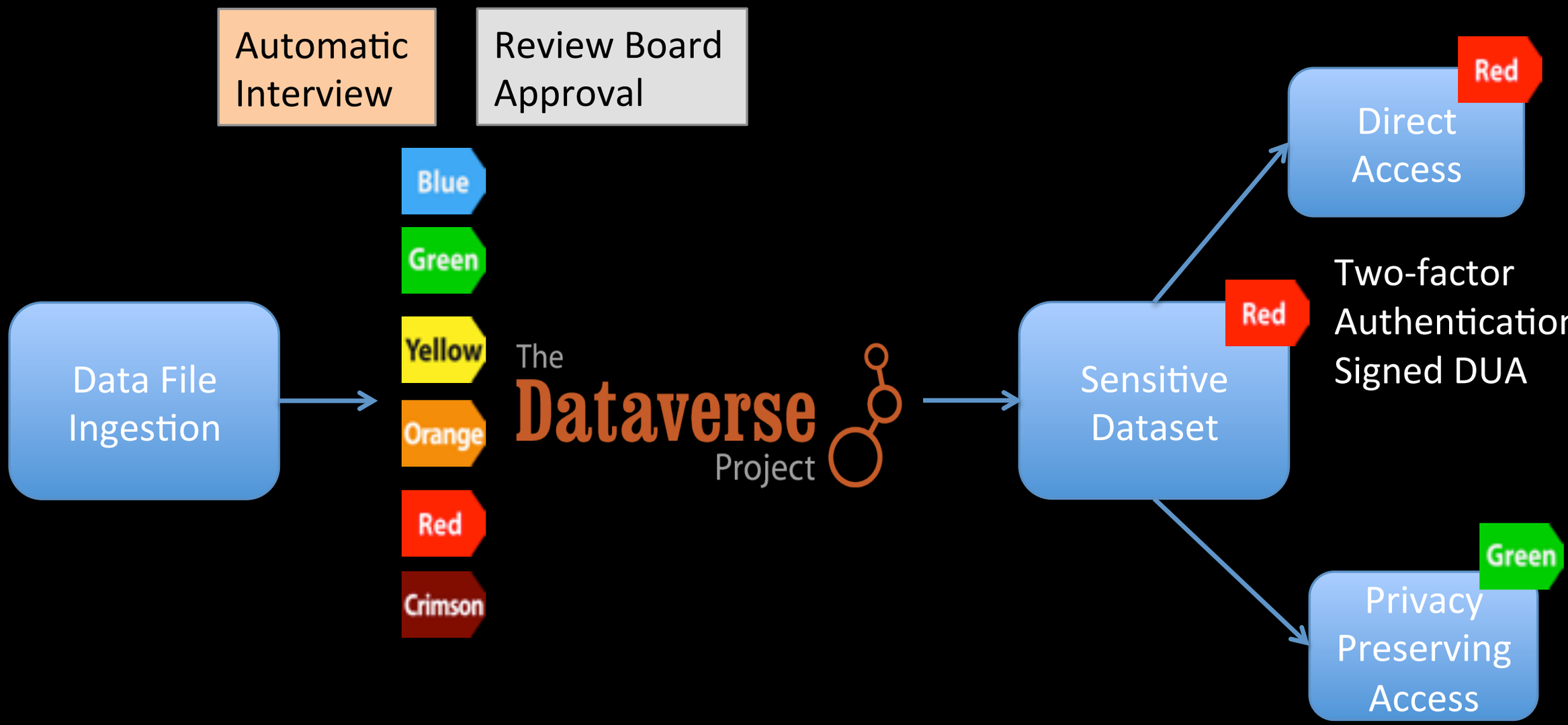
A datatag is a set of security features and access requirements for file handling

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

Datatags Levels

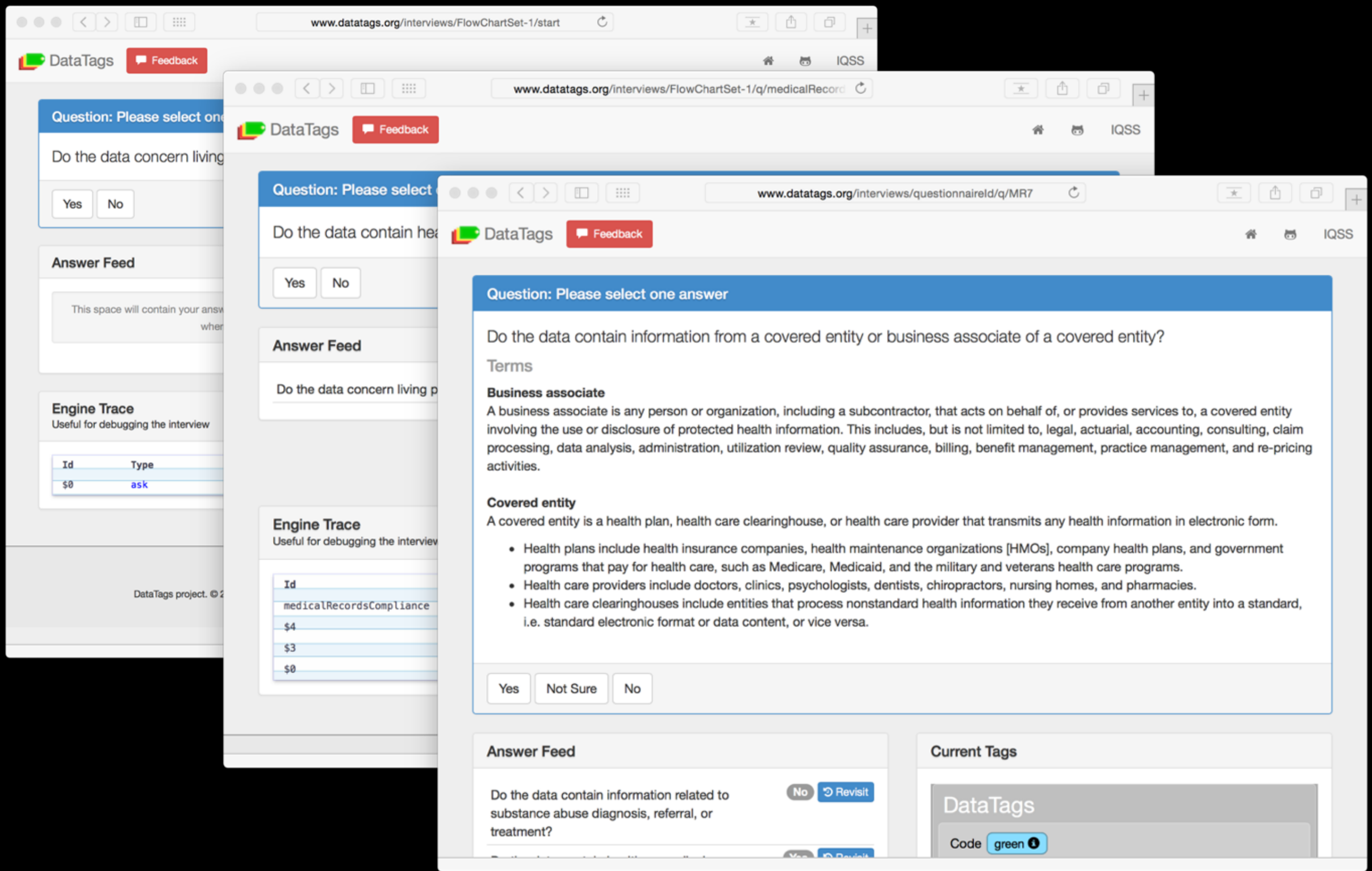
Tag Type	Description	Security Features	Access Requirements
Blue	**Public**	Clear storage Clear transmission	Open
Green	**Controlled public**	Clear storage Clear transmission	Email, OAuth verified registration
Yellow	**Accountable**	Clear storage Encrypted transmit	Password, Registered , Approval, Click 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)

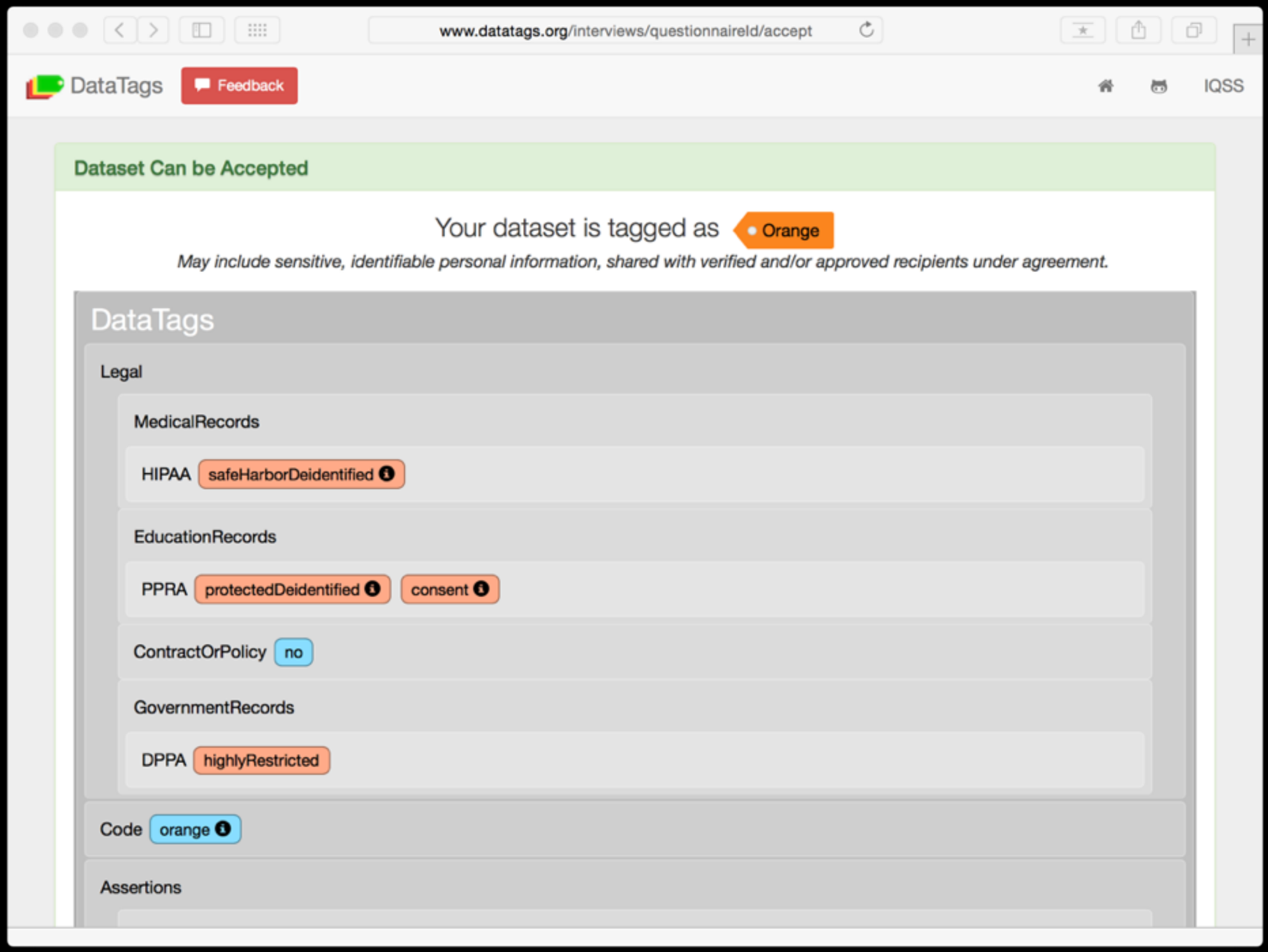


http://datatags.org
http://privacytools.seas.harvard.edu

Example of DataTags Interview



Example of DataTags Interview



Thanks!
And join us to this year's Dataverse Community Meeting

Dataverse Community Meeting 2016

July 11, 12, 13 at Harvard Medical School

The Dataverse Project

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.

References

- http://dataverse.org
- http://dataverse.harvard.edu
- http://datatags.org
- Sweeney L, Crosas M, Bar-Sinai M. 2015, *Sharing Sensitive Data with Confidence:The DataTags System*. Technology Science, http://techscience.org/a/2015101601
- Gross Harmon, Reidy, 2001, Communicating Science
- Mabe, 2003, The Growth and Number of Journals
- Friendly, 2006, A Brief History of Data Visualization