NFAIS Open Data Fostering Open Science
June 20, 2016

Dataverse and DataTags

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

@mercecrosas

"Research data publishing is the release of research data, associated metadata, accompanying documentation, and software code (in cases where the raw data have been processed or manipulated) 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."

RDA Data Publishing Workflows Working Group; 10.5281/zenodo.34542

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

Data Publishing is sharing data that are:

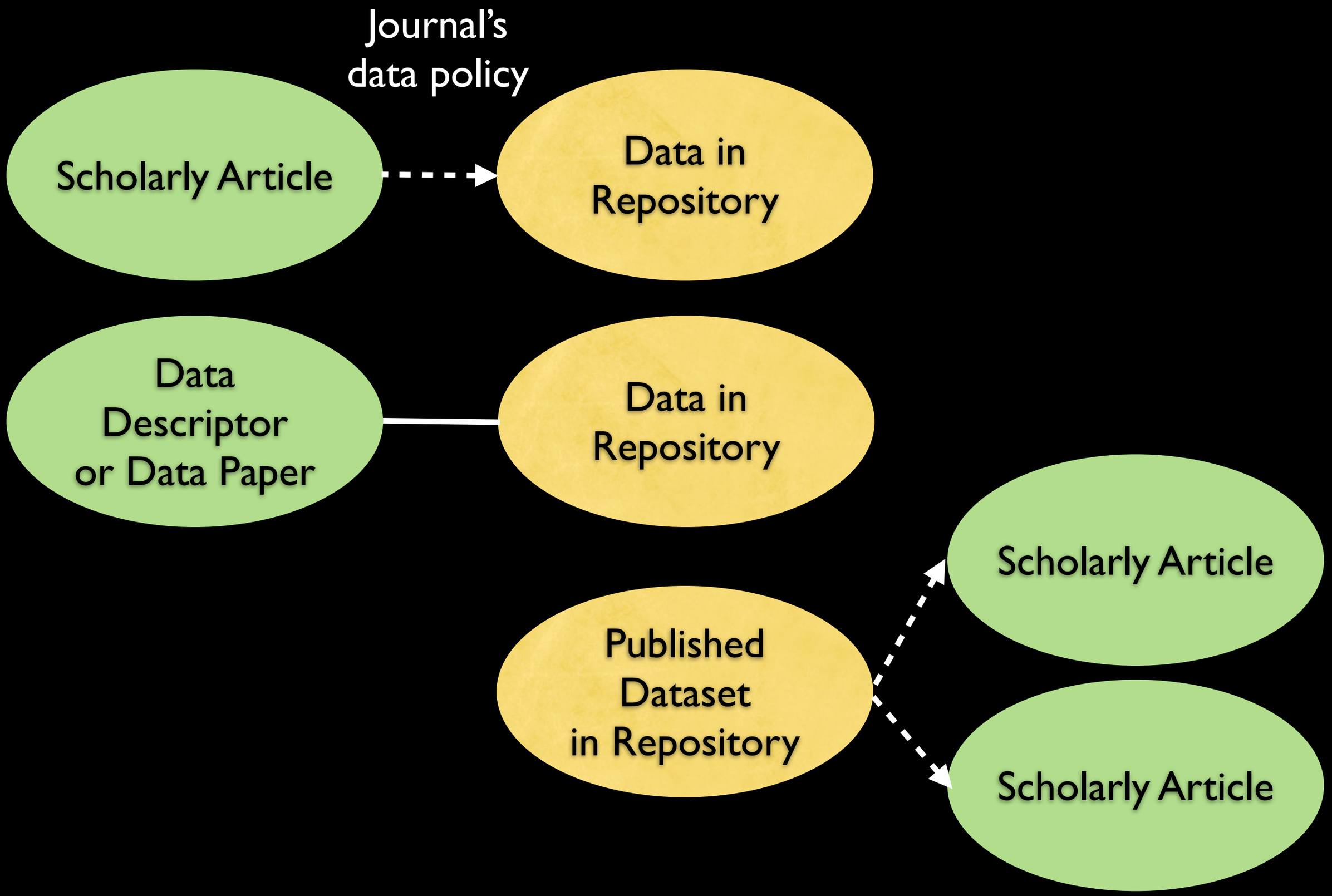
Findable **A**ccessible **I**nteroperable **R**eusable

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

Why publish data?

Researchers	→	Get credit for their data
Publishers and Journals	→	Verify published work
Federal funding agencies	→	Make public assets public
Science	→	Validate, reuse and extend previous work

Ways of Publishing Data



The Dataverse Project

A data repository system for sharing and archiving research data

A Solution for Publishing FAIR research data: Findable, Accessible, Interoperable, Reusable

The Dataverse Project

http://dataverse.org

Created and developed at Harvard's Institute for Quantitative Social Science

Dataverse Project | About | Community | Best Practices | Software | Contact

19 Installations | 1,600+ Dataverses | 61,000+ Datasets | 1,700,000+ Downloads

Dataverse Today: A growing Community

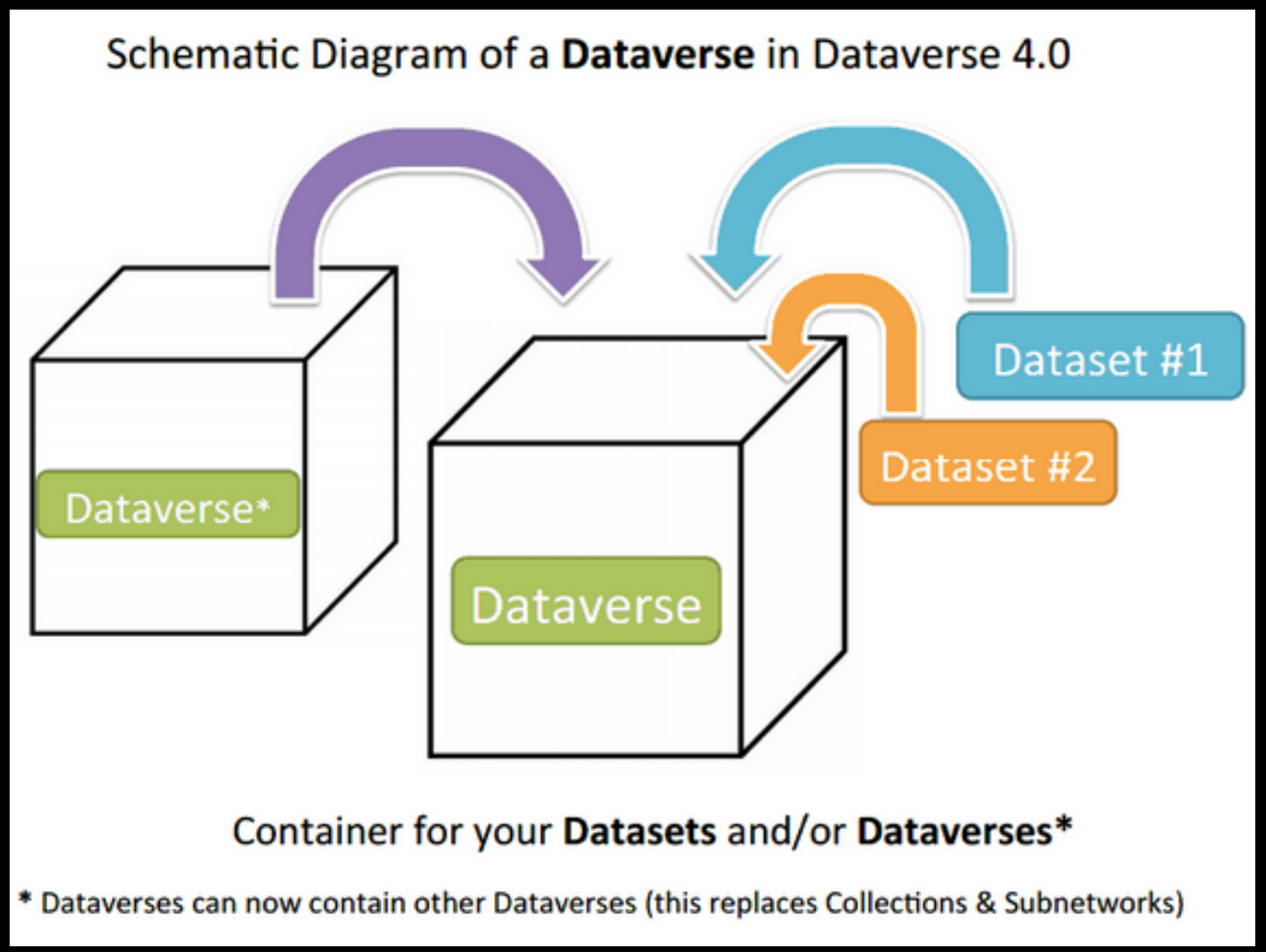
Dataverse Project:

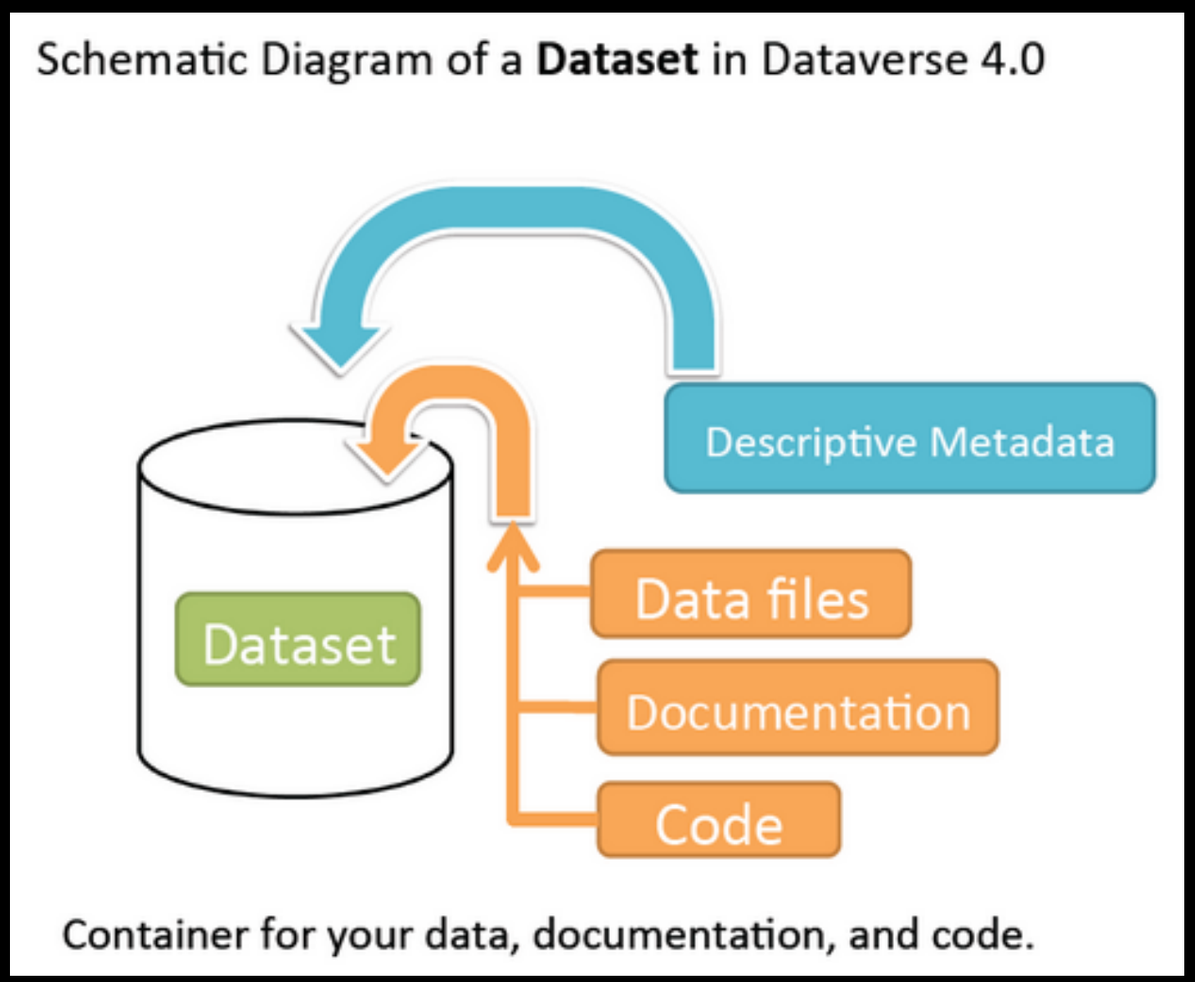
- Dataverse installations: 19; serving > 200 Universities
- User Community group: 294 members
- Open-source software: 29 contributors
- Dataverse Community Meeting (July, 2016): 107 registered, so far
- Twitter: 2940 followers

Harvard Dataverse Repository:

- Registered users: 13,795; 300 new per month
- Dataverses: 1,677; 50 new per month
- Journal Dataverses: 91
- Datasets: 61,781; 400 new per month
- Data Files: 330,462; 3,000 new per month

Dataverses contain datasets or dataverses
Datasets contain metadata and data files



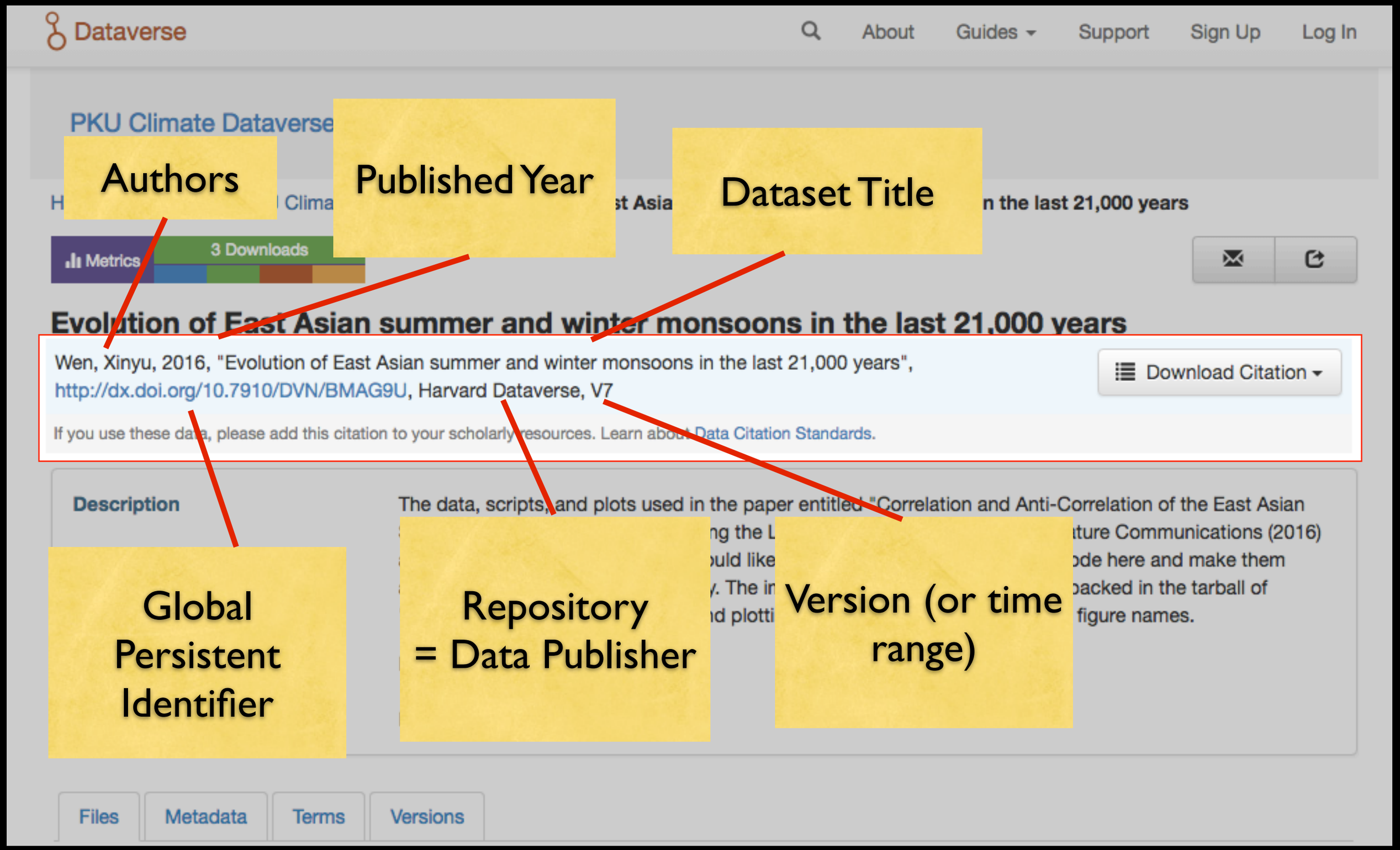


Dataverse follows best practices for FAIR Data Publishing

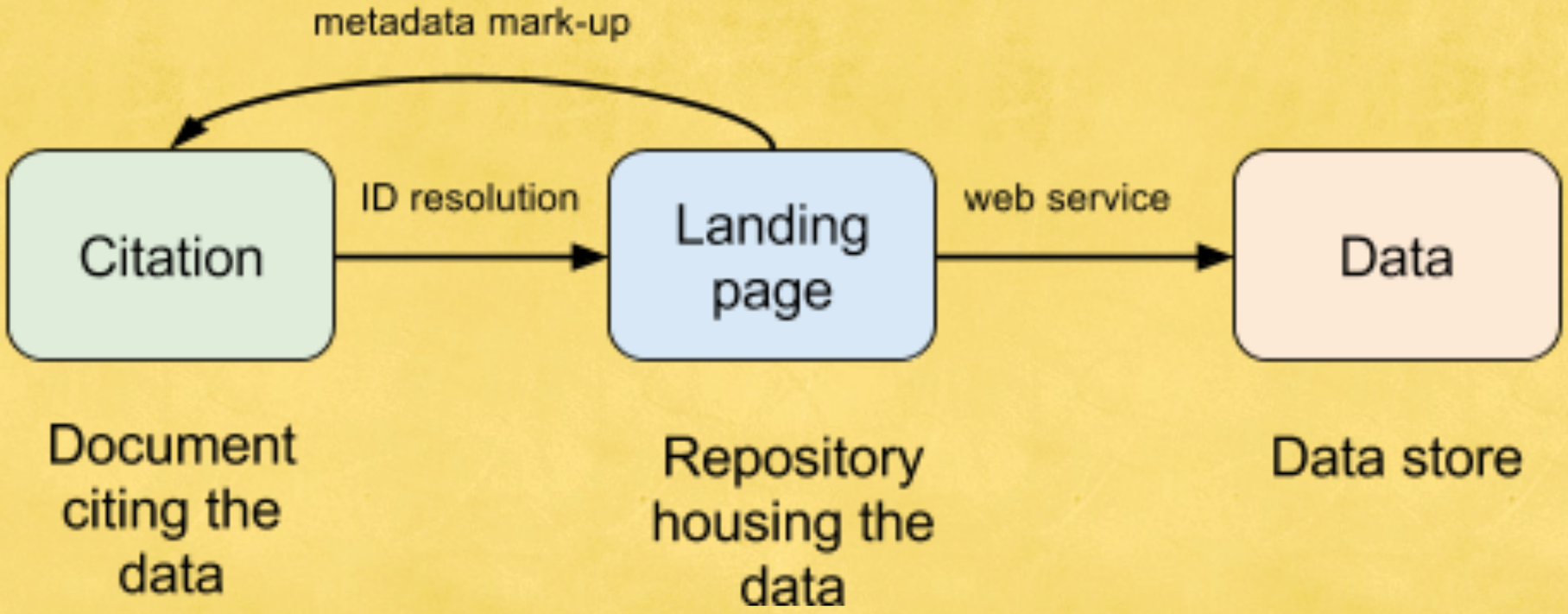
Best Practices

- Data Citation
  - Reference, locate and attribute
- Metadata
  - Discover and reuse
- Access Control and Rules
  - Access protecting privacy
- APIs and Standards
  - Interoperate

Data Citation in Dataverse



Data Citation Basics



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

Force11, Joint Declaration of Data Citation Principles, 2014; 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

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 review)

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

Dataverse Project | About | Community | Best Practices | Software | Contact

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:

- 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

Current Research Grants

Privacy tools to share sensitive data

Data provenance

Alfred P. Sloan FOUNDATION

Social Science Big Data

Journal articles connected to data

Data Privacy

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST

Biomedical large-scale data

How can we maximize data publishing of 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

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

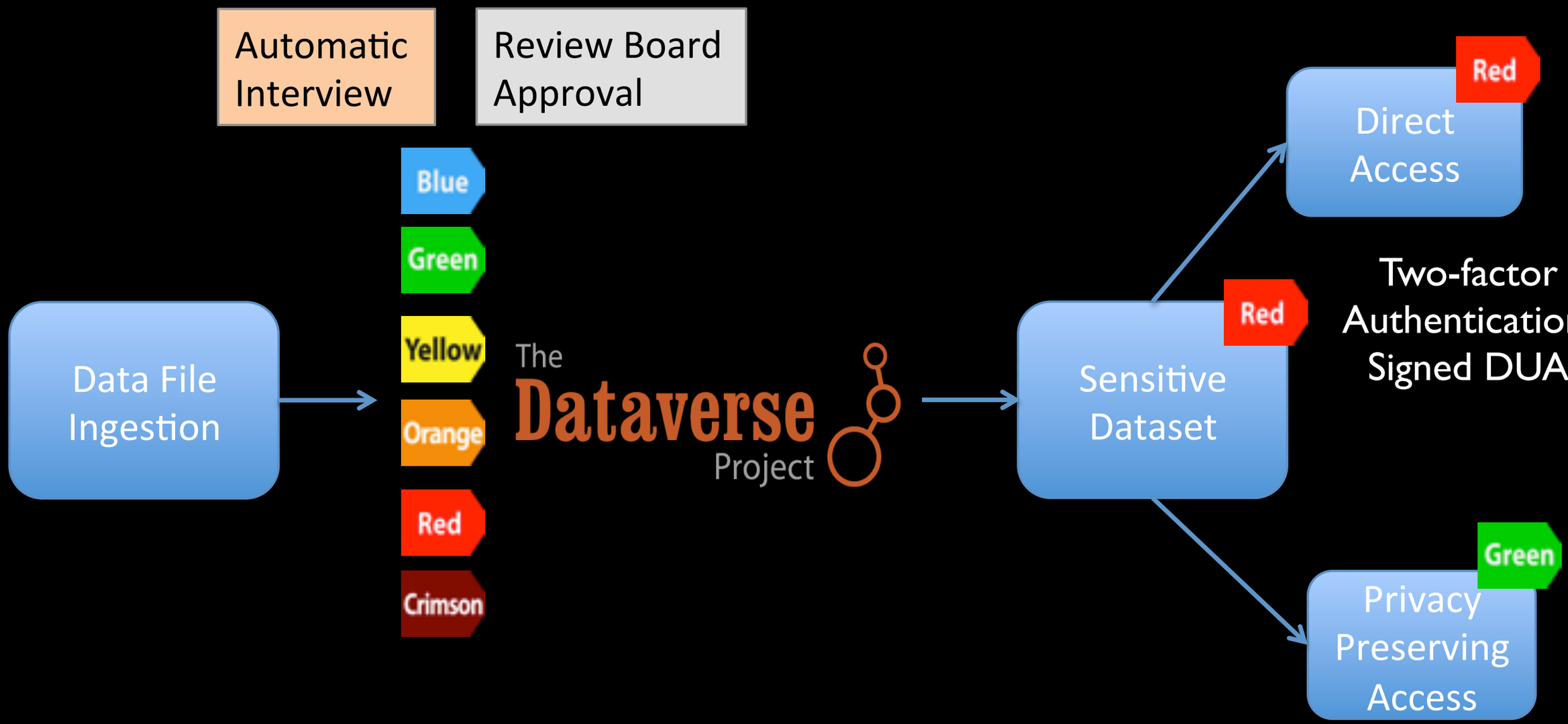
A datatag is a set of security features and access requirements for file handling.

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

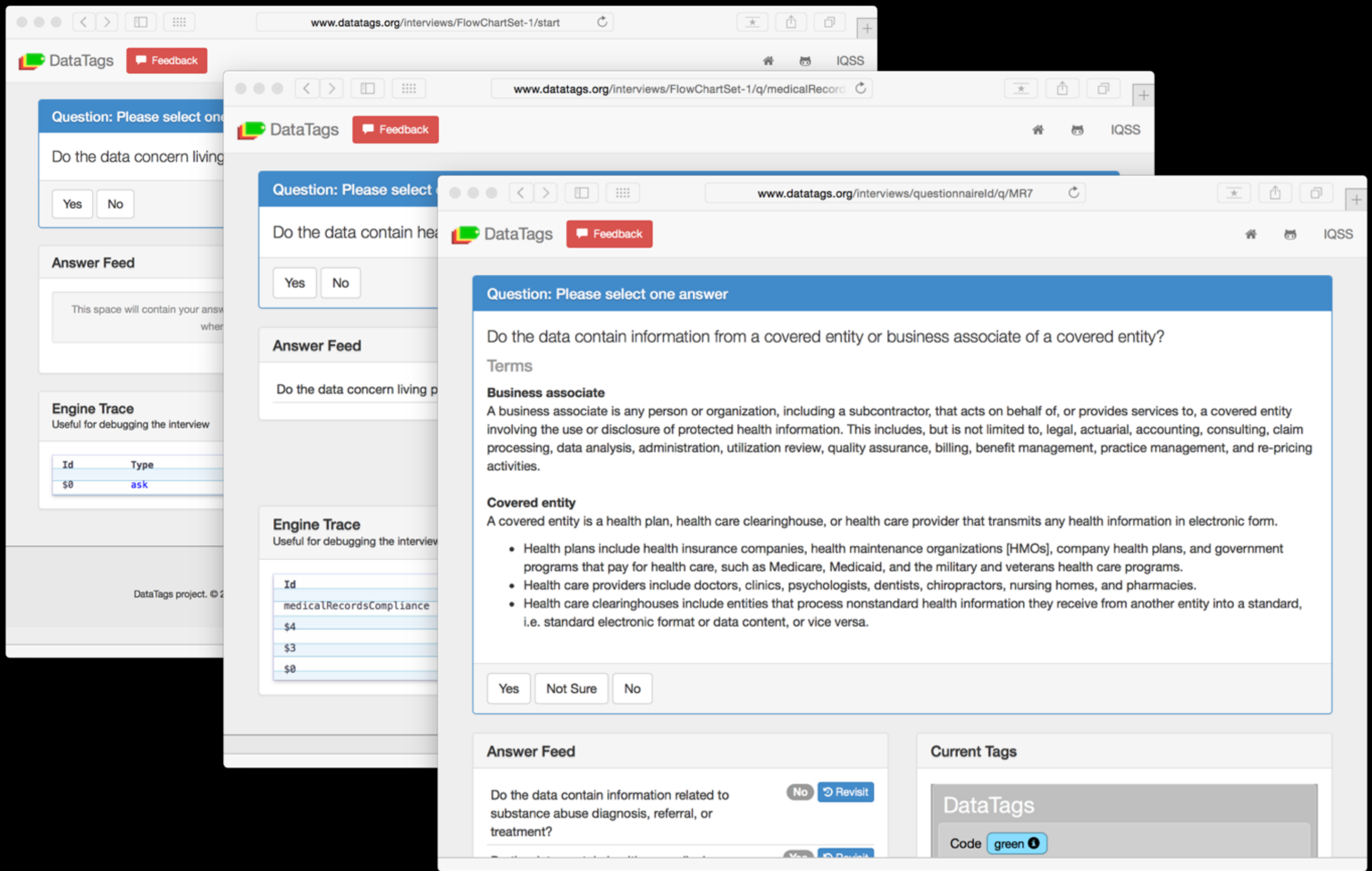


http://datatags.org
http://privacytools.seas.harvard.edu

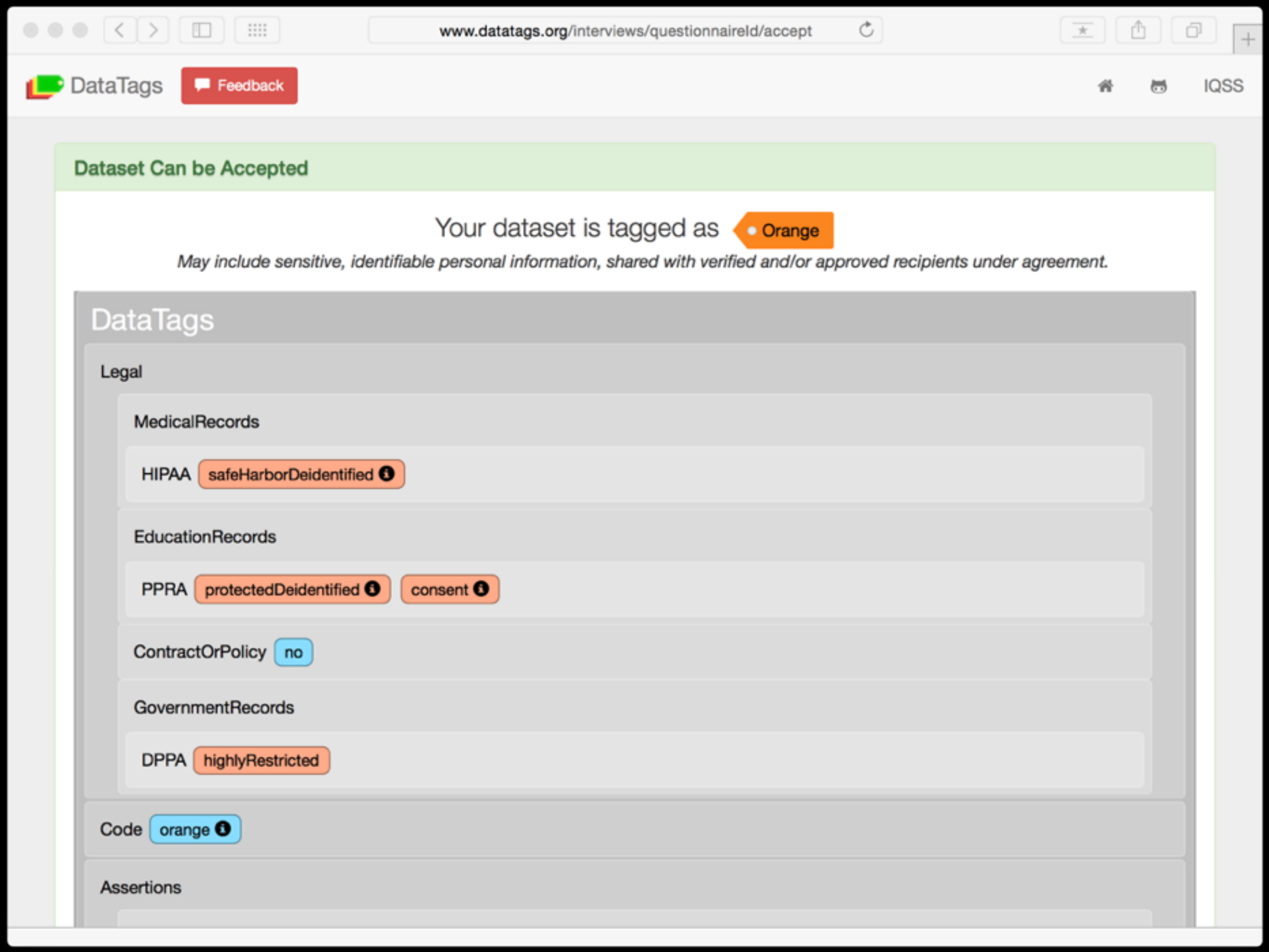
National Science Foundation

Alfred P. Sloan FOUNDATION

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



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



Summary

- Data sharing is good for researchers, journals, funding agencies, and science
- Dataverse is an open-source software for building data repositories to share research data
- 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
- DataTags generates human-readable and machine-actionable policies to support sensitive datasets in data repositories

Join us to this year's Dataverse Community Meeting

HARVARD UNIVERSITY HARVARD.EDU

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

@mercecrosas and http://scholar.harvard.edu/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