

# Data Publishing while Preserving Data Privacy

Mercè Crosas, Ph.D. [@mercecrosas](#)

Director of Data Science

Institute for Quantitative Social Science, Harvard University

<http://datascience.iq.harvard.edu>

**5R: Innovative Approaches to Promoting Transparency in Research**

IASSIST, June 5, 2014

# Introduction to Dataverse

## Dataverse Software

---

- ▣ A framework for publishing, citing and preserving research data
- ▣ Open-source, available at GitHub
- ▣ Started in 2006
- ▣ Several installations around the world, supporting all data types across multiple disciplines.

## Dataverse Repository

---

- ▣ **Harvard** hosts a Dataverse instance **free and open** to all research data
- ▣ More than 53,000 datasets, with 735,000 files.
- ▣ Dataverses can be created for researchers, journals, organizations, educators, ...
- ▣ It federates with the other Dataverse installations.

# Find and publish data at: <http://thedata.harvard.edu>



*Share, Cite, Reuse, Archive Research Data*  
Scientific data for reproducible research

POWERED BY THE **Dataverse Network** PROJECT v. 3.6.2

## Harvard Dataverse Network

[Create Account](#) [Log In](#)

[Advanced Search](#) [Tips](#)

We're redesigning Dataverse and want your feedback! Please check out our [Beta Site](#)

The Harvard Dataverse Network is open to all scientific data from all disciplines worldwide. It includes the world's largest collection of social science research data. [Learn more about the Dataverse Network.](#)

## Dataverses

**706** Dataverses

A **Dataverse** is a container for research data studies, customized and managed by its owner.

### RECENTLY RELEASED DATAVERSES

Eben N. Broadbent	Jun 2, 2014
USoc: Quantitative Methods over the Undergraduate Life Course	May 30, 2014

## Studies

**53,896** Studies, **739,606** Files, **1,015,093** Downloads

A **study** is a container for a research data set. It includes cataloging information, data files and complementary files.

### RECENTLY RELEASED STUDIES

Replication data for: Neoliberal Reform and Protest in Latin American Democracies: A Replication and Correction by Solt, Frederick; Kim, Dongkyu; Lee, Kyu Young; Willardson, Spencer; Kim, Seokdong	Jun 3, 2014
--	-------------

# Dataverse 4.0

The screenshot displays the Dataverse 4.0 Beta interface. At the top, there is a navigation bar with links for 'About', 'Software', 'Resources', 'Support', and a user profile for 'Pete Privileged'. Below this is the 'Harvard Dataverse' header. A search bar is present with the text 'Search this Dataverse...' and a 'Find' button. The main content area shows search results for '1 to 10 of 12 results'. The first result is a draft titled 'Results from the 2004 Election in Mississippi' by John Smith, 2014. The result includes a DOI link and a description of the data. The interface also features a left sidebar with filters for 'Publication Status', 'Affiliation', 'Publication Date', 'Author Name', 'Author Affiliation', 'Keyword', 'Subject', 'Contributor Type', 'Production Date', and 'Deposit Date'. A green callout box at the bottom left contains the text: 'Try Dataverse 4.0 Beta: http://dataverse-demo.iq.harvard.edu'.

## This summer:

- New UI
- New rich, faceted search
- Reformatting and metadata extraction for more data types (excel, CSV, R data, Stata, SPSS, FITS)
- Metadata standards for social sciences, astronomy, biomedical sciences.
- Integration with a new data exploration and analysis tool: **TwoRavens**



**Dataverse 4.0** will include a new interactive data exploration and analysis tool, **TwoRavens**, which integrates with **Zelig** statistical framework

TwoRavens

Estimate

Force

Reset

turnout

Variables

Subset

Summary

**educate**

Education

Mean: 12.2

Median: 12

Mode: NaN

Stand.Dev: 3.39

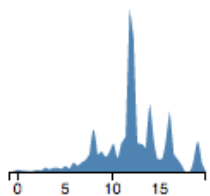
Minimum: 0

Maximum: 19

Valid: 15837

Invalid: 0

educate



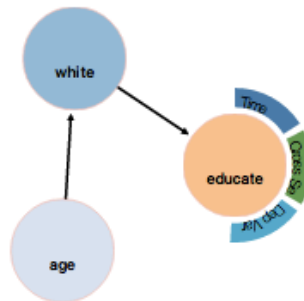
educate

log(d)

exp(d)

d^2

sqrt(d)



Results Table

Models

Set Covar.

Results

**Title \***

Replication Data for: Building a Bridge Betw

Add 'Replication Data for' to Title

**Author****Name \***

Castro, Eleni

**Affiliation**

IQSS

**Contact E-mail \***

ecastro@fas.harvard.edu

**Description \***

Research dataset for my publication on connecting journal articles and their underlying research data. Includes analysis of current data publication practices.

**Citation Metadata:**  
Compliant with DataCite, Dublin Core, DDI study description.  
Applies to all datasets.

**Keyword**

data publication

**Subject \***

- Mathematical Sciences
- Physics
- Social Sciences
- Other

**Topic Classification**

**Term**

**Vocabulary**



**URL**

**Software**

**Name**

**Version**



**Series**

**Name**

**Information**

**Time Period Covered**

**Start**

**End**



**Date of Collection**

**Start**

**End**



**Country/Nation**

**Geographic Coverage**

**Geographic Unit**

**Geographic Bounding Box**

**West Longitude**

**East Longitude**

**North Latitude**

**South Latitude**

**Social Sciences and Humanities Metadata: Compliant with DDI**

Type

- Image
- Mosaic
- EventList
- Spectrum
- Cube

Facility

Instrument

Spatial Resolution

Spectral Resolution

Time Resolution

Bandpass

Central Wavelength (m)

Wavelength Range

Minimum (m)

Maximum (m)

Dataset Date Range

Start

End

**Astronomy Metadata:  
Compliant Virtual Observatory  
(VO) schema; extract metadata  
from FITS files**



## Design Type

- Case Control
- Cross Sectional
- Not Specified
- Parallel Group Design
- Perturbation Design

## Factor Type

- Age
- Biomarkers
- Developmental Stage
- Cell Surface Markers
- Cell Type/Cell Line

## Measurement Type

- DNA Methylation Profiling (Bisulfite-Seq)
- DNA Methylation Profiling (MeDIP-Seq)
- Histone Modification (ChIP-Seq)
- Protein-RNA Binding (RIP-Seq)
- Transcription Factor Binding (ChIP-Seq)

**Bio Metadata:**  
Compliant with ISA-Tab schema,  
plus biomedical ontologies

## Organism

- Danio rerio
- Homo sapiens
- Mus musculus
- Rattus norvegicus

## Cell Type



# Data Publishing Guidelines

Three pillars to Data Publishing:

- A trusted data repository to guarantee long-term access
- A formal data citation\*
- Sufficient information to understand and reuse the data (metadata, documentation, code)

By supporting these pillars, Dataverse provides a full solution to Data Publishing.

\* Data Citation Principles: <https://www.force11.org/datacitation>

# A Rigorous Data Publishing Workflow



A Published Dataset cannot be deleted (only de-accessioned, if legally needed)

Draft dataset



Published Dataset V1



Published Dataset V1.1



Published Dataset V2

**Publish Version 1:** once published, **metadata** is always **public** (e.g., CC0), but data files might be restricted.

Authors, Title, Year, DOI Repository, UNF, V1

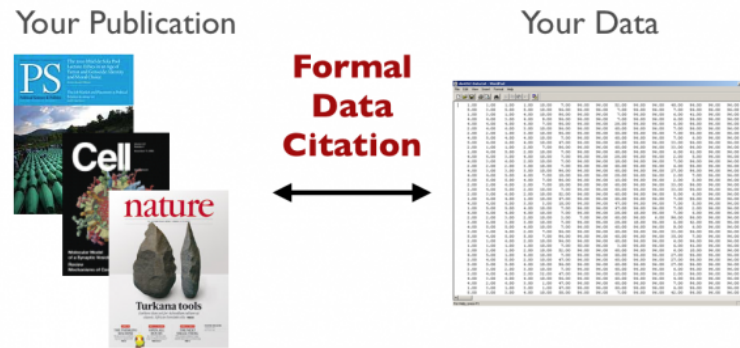
**Publish Version 1.1:** small metadata change; citation doesn't change

Authors, Title, Year, DOI Repository, UNF, V1

**Publish Version 2:** big metadata change, or file change; citation changes

Authors, Title, Year, DOI Repository, UNF, V2

# Workflows that Integrate with Journals



**Option A.** Author publishes a dataset to his/her Dataverse, then provides the Data Citation to the journal.

**Option B.** Author contributes to a journal Dataverse:

1. Add dataset to Journal Dataverse as a draft.
2. Journal Editor reviews it, and approves it for release.
3. Dataset is published with Data Citation and link from journal article to the data.

**Option C.** Seamless Integration between journal system and Dataverse.

# Example of Option C: OJS and Dataverse Integration

## OJS Journal

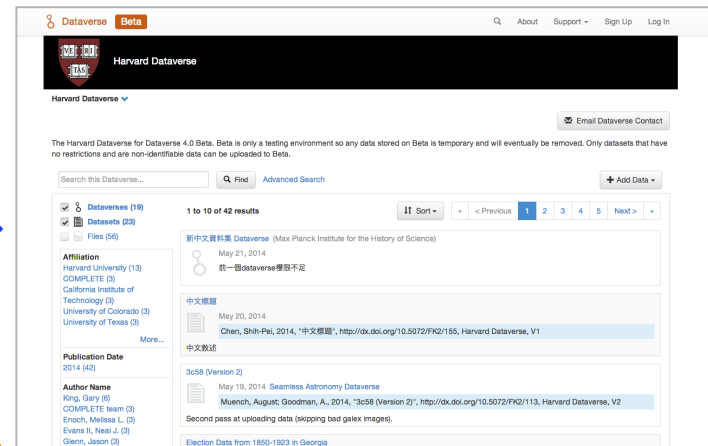


Citation  
to Data



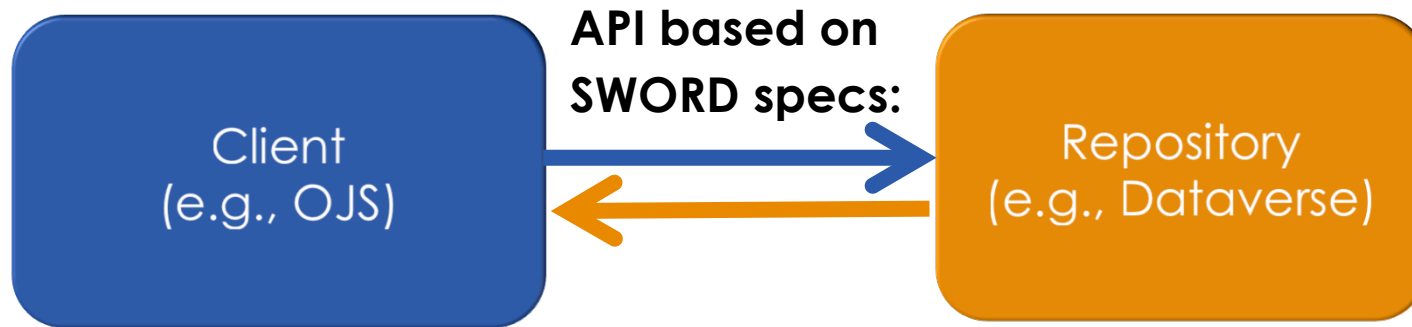
Citation  
to Article

## Journal Dataverse



- ❑ Sloan funded project to integrate PKP's Open Journal System (OJS) with the Dataverse software.
- ❑ Pilot with ~ 50 journals
- ❑ OJS Dataverse plugin now available with latest OJS release
- ❑ <http://projects.iq.harvard.edu/ojs-dvn>

# Toward a common API between journal systems and data repositories



## Client sends:

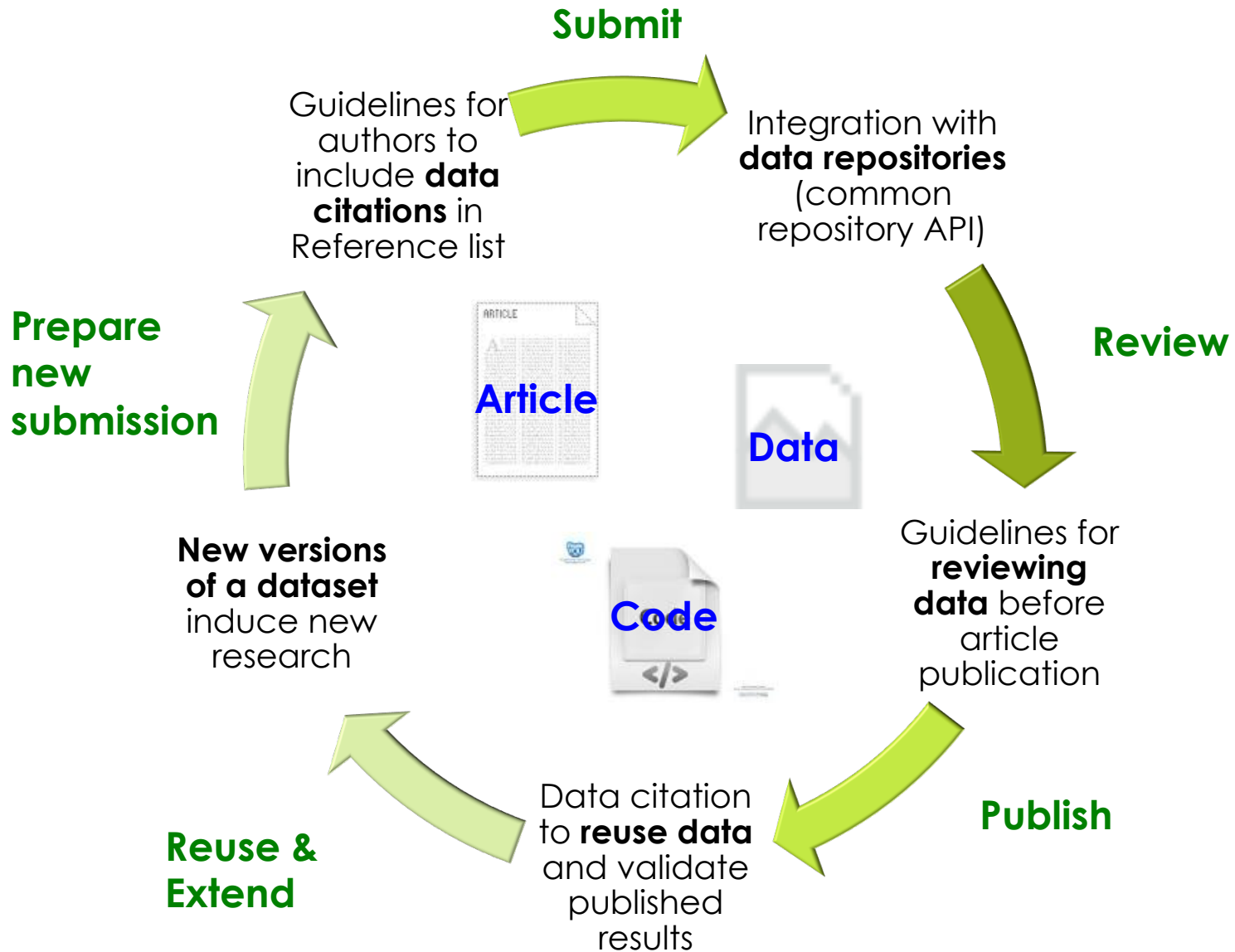
- ✓ XML file: AtomPub "entry" with Dublin Core Terms (e.g., title, creator)
- ✓ Zip file: All data files associated with that dataset.

## Repository sends:

- ✓ XML file: "Deposit Receipt"  
→ send **data citation** from repository to client

Plus updates from client to server during lifecycle:  
**In review, publish first version, new versions**

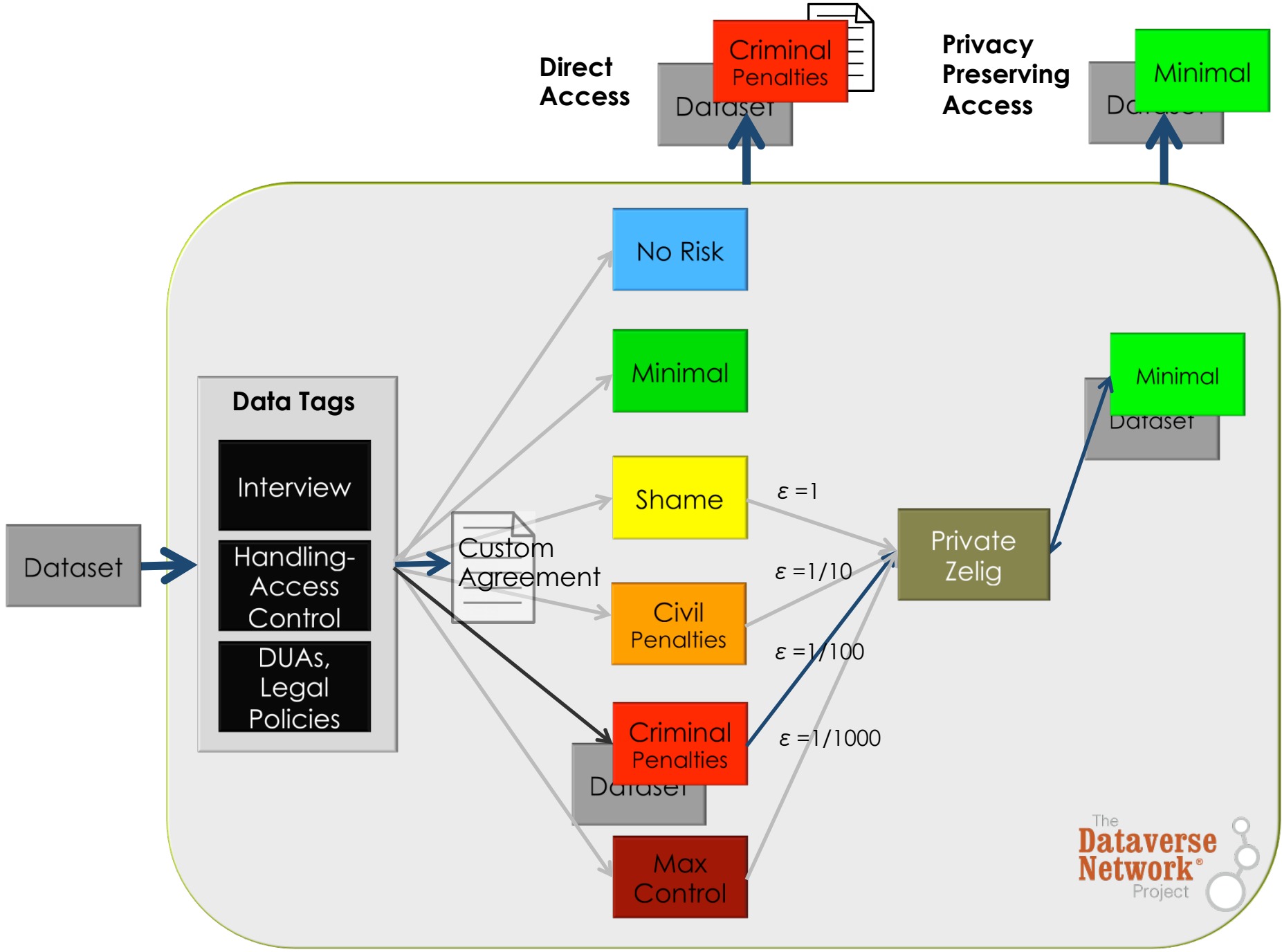
# Toward an integrated, living publishing workflow



# Publishing Sensitive Data

- ▣ Dataverse is part of a 4 years NSF funded project on **Privacy Tools for Sharing Sensitive Data** (*Harvard SEAS, Berkman Center, Data Privacy Lab, IQSS*).
- ▣ This project includes:
  - ▣ **DataTags**: A framework that provides data handling prescriptions to comply with numerous privacy regulations and data user agreements
  - ▣ **Private Zelig**: A differential privacy version of the Zelig statistical framework





Direct Access

Criminal Penalties

Privacy Preserving Access

Minimal

Data Tags

Interview

Handling-Access Control

DUAs, Legal Policies

Custom Agreement

No Risk

Minimal

Shame

Civil Penalties

Criminal Penalties

Max Control

$\epsilon = 1$

$\epsilon = 1/10$

$\epsilon = 1/100$

$\epsilon = 1/1000$

Private Zelig

Minimal

The Dataverse Network Project

Try our new Beta version: <http://datatags.org>

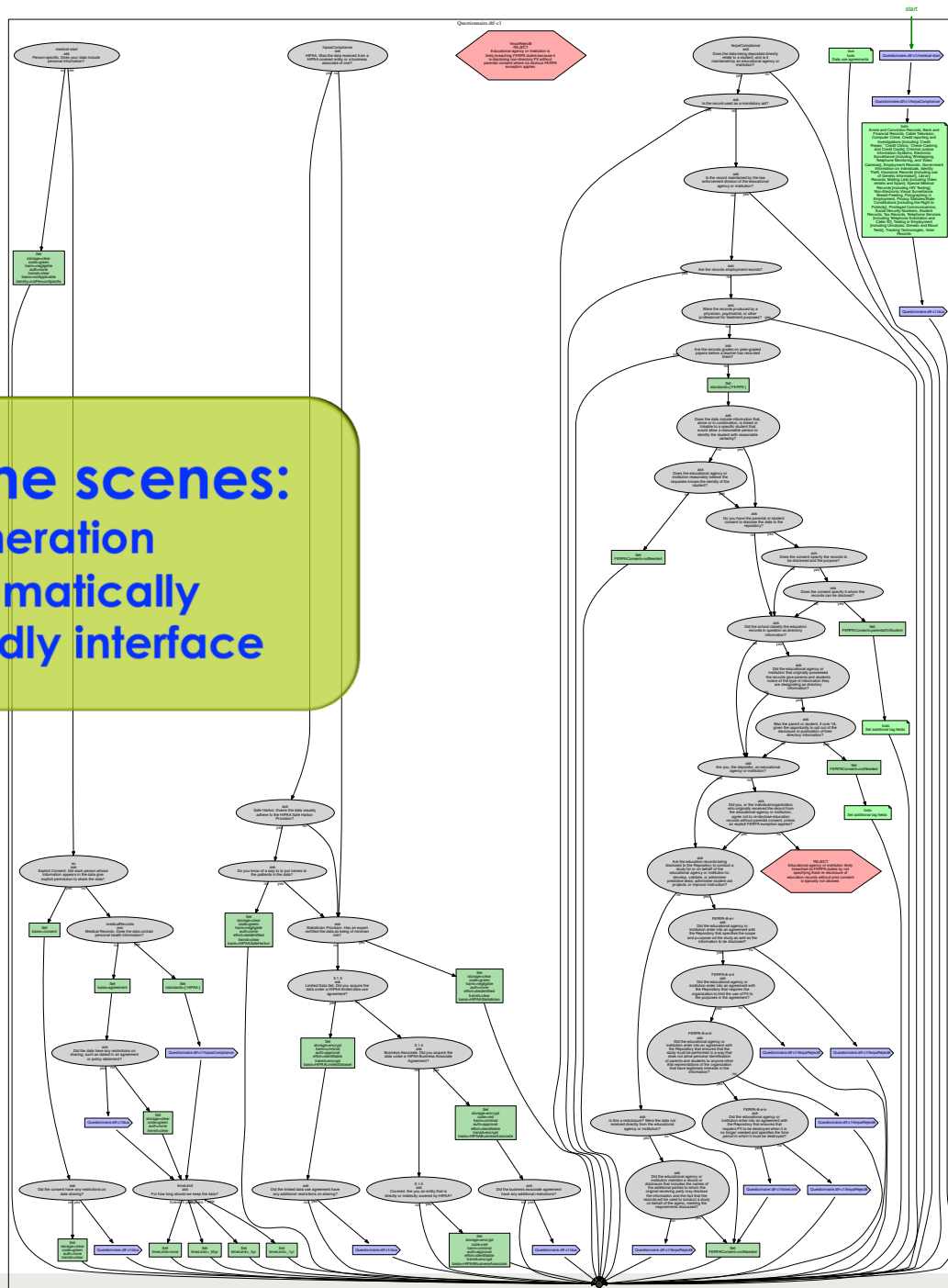
### Harm Levels and Their Appropriate Tags

The tags below denote are the minimal handling requirements, based on the harm level inherent to the data. The tags resulting from the tagging interview may be more restrictive, due to data use agreements, contracts etc. Hover/touch tags for explanation

Level	DUA Agreement Method	Authentication	Transit	Storage
<b>No Risk</b>	None	None	Clear	Clear
<b>Minimal</b>	None	Email or OAuth	Clear	Clear
<b>Shame</b>	Click Through	Password	Encrypted	Clear
<b>Civil Penalties</b>	Sign	Password	Encrypted	Encrypted
<b>Criminal Penalties</b>	Sign	Two Factor	Encrypted	Encrypted
<b>Max Control</b>	Sign	Two Factor	Double Encryption	Double Encryption

Currently supporting HIPAA and FERPA (and DUAs)

# DataTags behind the scenes: A complex interview generation framework, which is automatically converted to a user-friendly interface



# Interview Example: First question ...

**Question: Please select one answer**

Person-specific. Does your data include personal information?

**Terms**

**personal information**  
as defined in HIPAA

**data**  
0s and 1s in some structured way

# Interview Example: After several questions ...

**Question: Please select one answer**

Were the data collected by a federal agency?

**Answer Feed**

Does the data being deposited directly relate to a student,	<input checked="" type="radio"/> no	<input type="button" value="Revisit"/>
For how long should we keep the data?	<input checked="" type="radio"/> 5 years	<input type="button" value="Revisit"/>
Covered. Are you an entity that is directly or indirectly	<input checked="" type="radio"/> yes	<input type="button" value="Revisit"/>
Business Associate. Did you acquire the data under a	<input checked="" type="radio"/> no	<input type="button" value="Revisit"/>
Limited Data Set. Did you acquire the data under a HIPAA	<input checked="" type="radio"/> no	<input type="button" value="Revisit"/>
Statistician Provision. Has an expert certified the data as	<input checked="" type="radio"/> no	<input type="button" value="Revisit"/>
HIPAA. Was the data received from a HIPAA covered	<input checked="" type="radio"/> no	<input type="button" value="Revisit"/>
Medical Records. Does the data contain personal health	<input checked="" type="radio"/> yes	<input type="button" value="Revisit"/>
Explicit Consent. Did each person whose information	<input checked="" type="radio"/> no	<input type="button" value="Revisit"/>
Person-specific. Does your data include personal	<input checked="" type="radio"/> yes	<input type="button" value="Revisit"/>

# Interview Example: ... and a Final Tag

Your dataset is tagged as



*Very sensitive identifiable personal information, shared with strong verification of approved recipients under signed agreement.*

## Full Tags

DataTags	
code	red
DataType	
harm	criminal
effort	identifiable
standards	HIPAA
Handling	
storage	encrypt
auth	approval
transit	encrypt
basis	HIPAABusinessAssociate
DUA	
timeLimit	_5yr

Learn more at: <http://datascience.iq.harvard.edu>

## Data Science

*Research Frameworks for Data-Intensive Science,  
Analytical Tools and Data Stewardship*



Zelig    Dataverse    TwoRavens    DataTags    Consilience    RBuild

### About Us

Data Science at IQSS combines expertise in software engineering, statistical innovation and data curation. Meet our team.

THANKS @mercecrosas