



Dataverse 4.0 & Beyond

Eleni Castro > Institute for Quantitative Social Science (IQSS), Harvard University



DataTags Sharing data with confidence

The DataTags system helps dataset owners handle their data properly. Using a user-friendly interface, the system detects what laws, regulations and contracts apply to a given dataset, and provides the dataset owner with a set of "DataTags", which explain what is the harm level the dataset can cause, and what is the proper way of handling it, both legally and ethically.

The DataTags project is in Beta. Don't use the tags as a legal recommendation...yet

[Start Tagging](#)

Harm Levels and Their Appropriate Tags

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

DUA Agreement Method	Authentication	Transit	Storage
None	None	Clear	Clear
None	Email or OAuth	Clear	Clear
Click Through	Password	Encrypted	Clear
Sign	Two Factor	Encrypted	Encrypted
Sign	Two Factor	Double Encryption	Double Encryption

Dataverse Beta About Support Feedback Sign Up Log In

Harvard Dataverse

Email Dataverse Contact

* Beta is only a testing environment so any data stored on Beta is temporary and will eventually be removed. To upload real data and receive a formal data citation, please use thedata.harvard.edu ** Only datasets that have no restrictions and are non-identifiable data can be uploaded to Beta.

Featured Datasets

Center for Astrophysics Database | Stem Cell Research Database | Ubiquity Press Database | Election Data Database

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

Datasets (9) Datasets (18) Files (35)

1 to 10 of 27 results

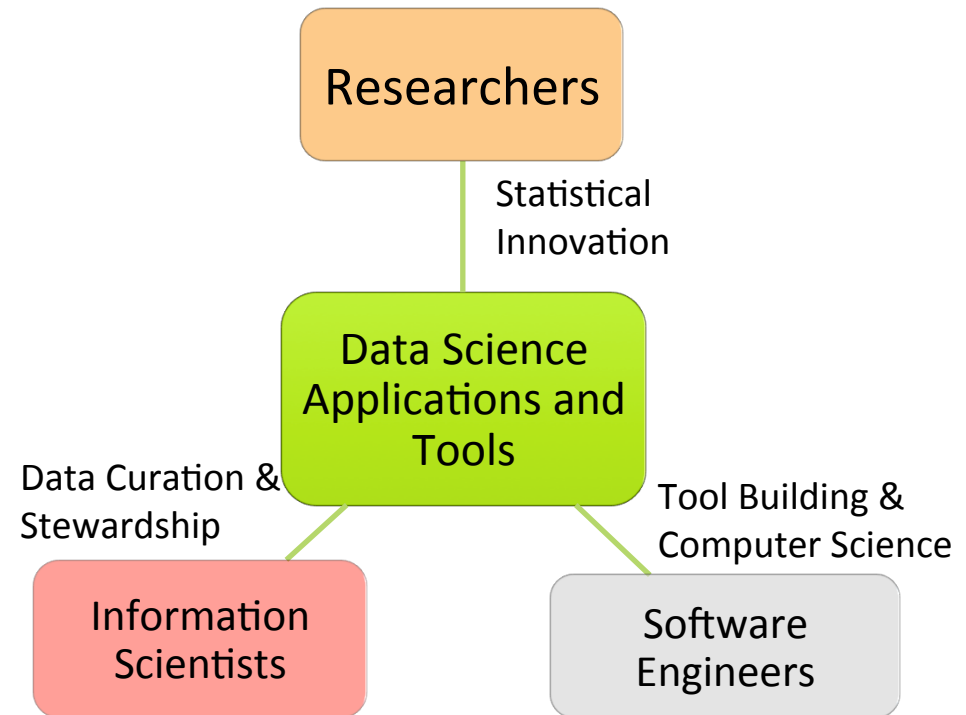
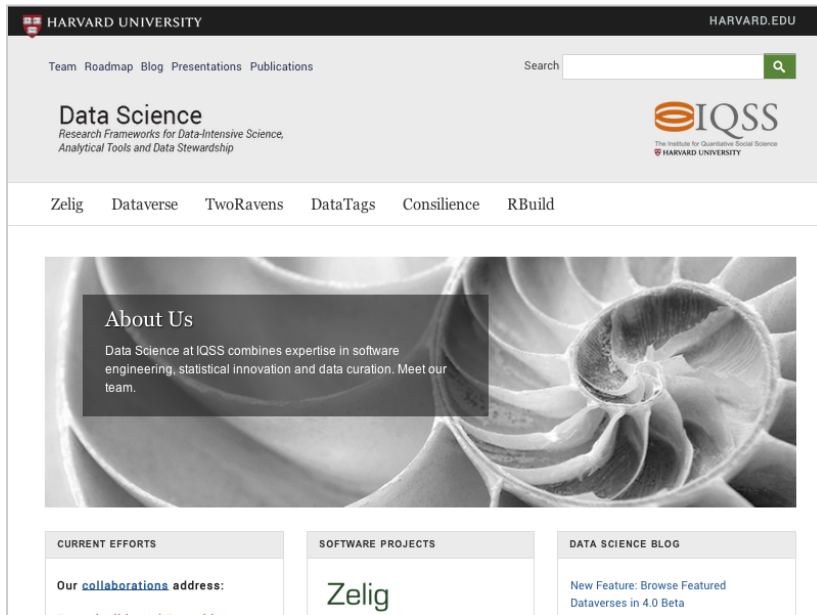
10 Million International Dyadic Events
Aug 15, 2014 Gary King Dataverse
King, Gary; Lowe, Will, 2014, "10 Million International Dyadic Events", <http://dx.doi.org/10.5072/FK2/11>, Harvard Dataverse, V2

When the Palestinians launch a mortar attack into Israel, the Israeli army does not wait until the end of the calendar year to react. Yet, most modern data collections are aggregated to the month or y...

Cause of Death Data



Data Science Team



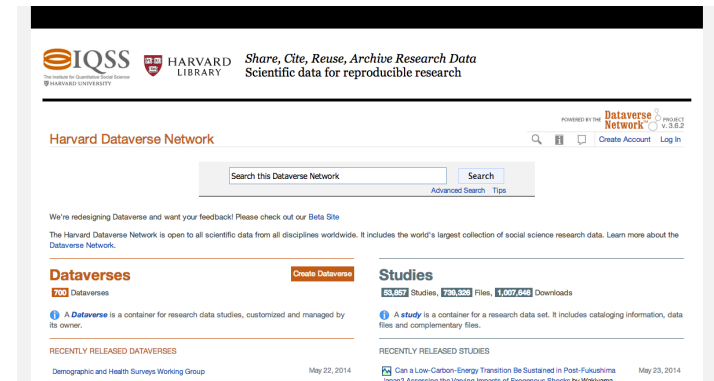
Find out more: <http://datascience.iq.harvard.edu>

What is Dataverse?

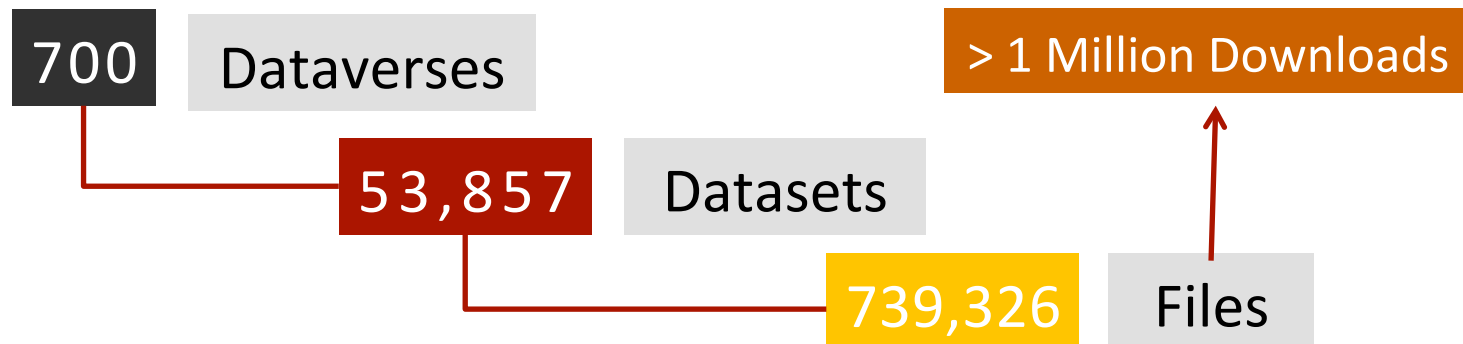
Software framework for publishing, citing and preserving research data
(open source on [github](#) for others to install)

Provides incentives for researchers to share:

- Recognition & credit via data citations
- Control over data & branding
- Fulfill Data Management Plan requirements

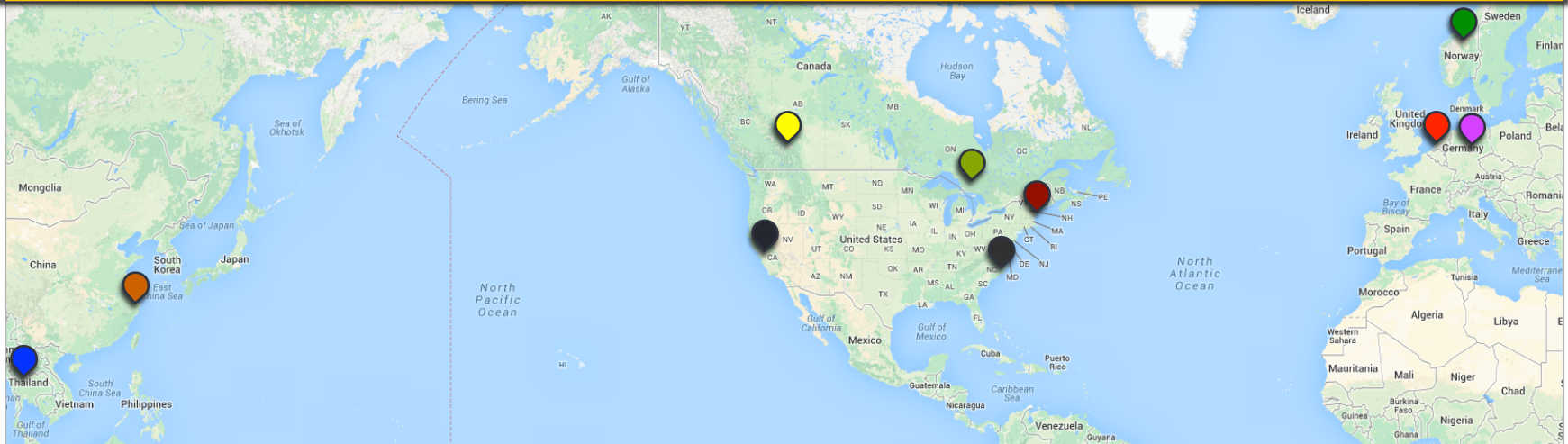


Harvard Dataverse (open to all, repository instance at Harvard) currently has:



Who is using Dataverse?

Worldwide Dataverse Installations

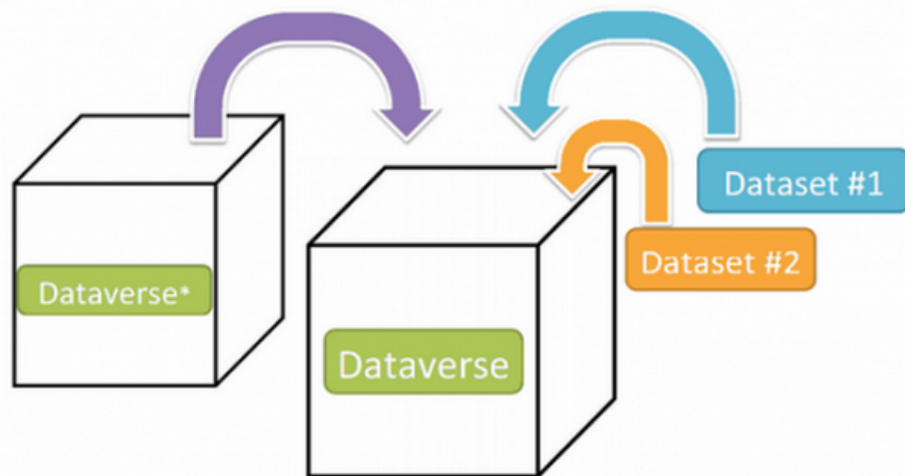


Institutions can setup/host their own Dataverse installation (OCUL, UoA, etc) and within them can have dataverses for a variety of users (across all research domains): Researchers, Projects, Journals (OJS – Dataverse integration), etc.

Streamlined Workflows

Based on extensive continuous usability testing: improved account creation process, dataverse setup (incl. customizations), and dataset (prev. study) creation.

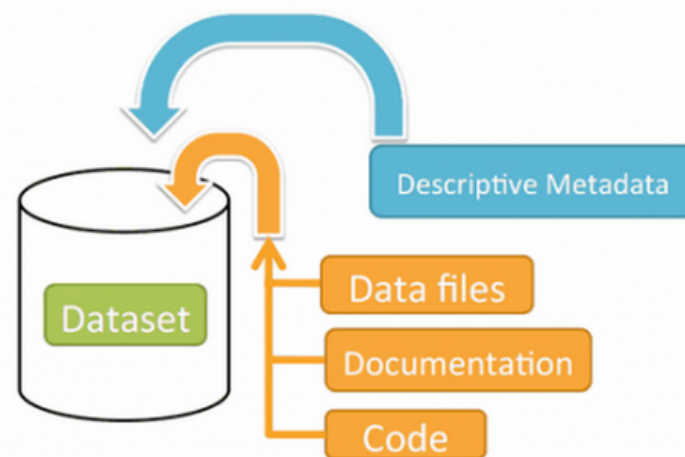
Schematic Diagram of a **Dataverse** in Dataverse 4.0



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.

Featured Dataverses

 Dataverse **Beta**

 [About](#) [Support](#) [Feedback](#) [Sign Up](#) [Log In](#)



Harvard Dataverse

Harvard Dataverse

 [Email Dataverse Contact](#)



* Beta is only a testing environment so any data stored on Beta is temporary and will eventually be removed. To upload real data and receive a formal data citation, please use thedata.harvard.edu ** Only datasets that have no restrictions and are non-identifiable data can be uploaded to Beta.

Featured Dataverses



Search this Dataverse...

 Find [Advanced Search](#)



 Add Data 

 [Dataverses \(9\)](#)

 [Datasets \(18\)](#)

 [Files \(35\)](#)

1 to 10 of 27 results

 Sort 


« < Previous **1** 2 3 Next > »

[10 Million International Dyadic Events](#)


Aug 15, 2014 [Garv King Dataverse](#)

Improved File Upload & Handling

Select multiple files, Drag-n-Drop, Dropbox, File Previews, and extra handling for csv, tsv and excel files (no control card needed).

 **Dataverse** Beta

[Q](#)
[About](#)
[Support](#)
[Feedback](#)
[Pete Privileged](#) 13

Privileged, Pete, 2014, "Test", <http://dx.doi.org/10.5072/FK2/90>, Root Dataverse, DRAFT VERSION 





look

Keyword world



Subject Business and Management

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

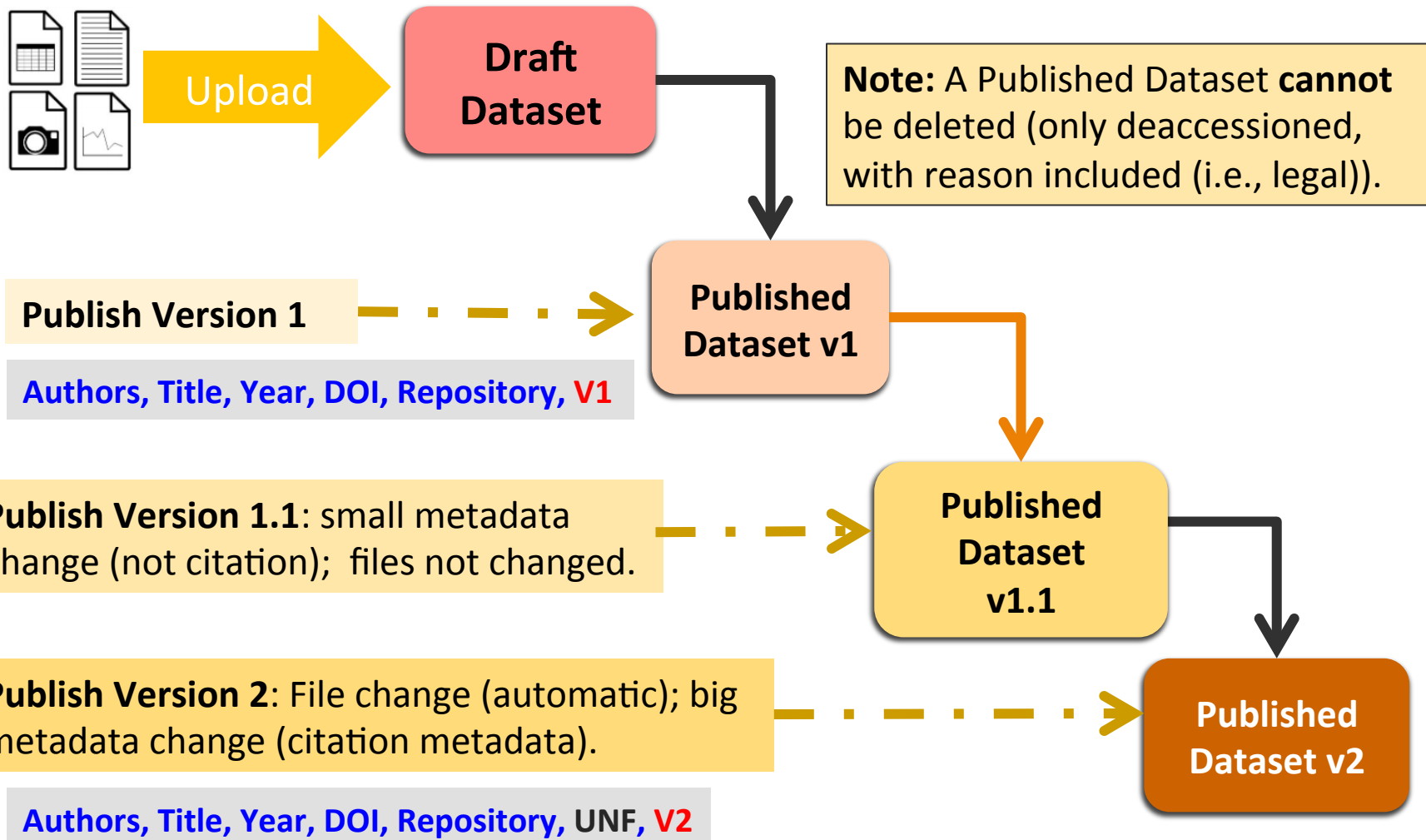
[+ Upload + Edit Files](#)

	<p>Brasil-Dataverse.png PNG Image, MD5: be49926d437bfa1a501ef5c56f5e8c5d</p>	Download
	<p>Example PDF version of Codebook.pdf Adobe PDF, MD5: b7467cb38befd33f3fa9c0e849fac1ed</p>	Download
	<p>OJSjournals_with_10_articles_2013.csv Comma Separated Values, MD5: 13bee8a47b055501c655dcb4cb2303df</p>	 Download

© Copyright 1997-2014, President & Fellows Harvard University.

Powered by  **Dataverse**
Project  v. 4.0

Rigorous Data Publishing Workflows



Expanding Metadata Support

Metadata Schema	Version 3.6	Version 4.0
DDI (General & Social Science)*	X (v2.1)	X (v.2.5)
Simple Dublin Core	X	X
Dublin Core Terms		X
DataCite 3.0		X
Virtual Observatory (Astronomy)**		X
ISA-Tab (Biomedical)***		X

* Including variable level metadata found in [tabular data files](#).

** Automatically extracts relevant metadata from the header [FITS files](#).

*** Controlled vocabulary maps to ontologies/taxonomies (OBI, NCBI,...).

Type

- Image
- Mosaic
- EventList
- Spectrum
- Cube

Facility

Instrument

Spatial Resolution

Spectral Resolution

Time Resolution

Bandpass

Central Wavelength (m)

Wavelength Range

Minimum (m)	Maximum (m)	<input type="button" value="+"/>
<input type="text"/>	<input type="text"/>	

Dataset Date Range

Start	End	<input type="button" value="+"/>
<input type="text" value="YYYY-MM-DD"/>	<input type="text" value="YYYY-MM-DD"/>	

Astronomy Metadata:
 Certain values (e.g., Type, Facility, Instrument, etc) automatically extracted from FITS file header.



Design Type

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

Biomedical Metadata

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)

Organism

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

Cell Type



Enhanced Faceted Search




 [Email Dataverse Contact](#)

Search this Dataverse...

 Find

[Advanced Search](#)

 Add Data

 **Dataverses (0)**

 **Datasets (1)**

 **Files (0)**

Affiliation

Harvard University (1) X
[Stanford University \(1\)](#)
[University of California \(1\)](#)

Publication Date


2014 (1)

Author Name

King, Gary (1) X
Tomz, Michael (1) X
[Wittenberg, Jason \(1\)](#)

Affiliation: Harvard University X **Author Name: King, Gary X** **Author Name: Tomz, Michael X**

1 to 1 of 1 result

 Sort

[Replication Data for: Making the Most of Statistical Analyses: Improving Interpretation and Presentation](#)




Aug 13, 2014 [Gary King Dataverse](#)

King, Gary; Tomz, Michael; Wittenberg, Jason, 2014, "Replication Data for: Making the Most of Statistical Analyses: Improving Interpretation and Presentation", <http://dx.doi.org/10.5072/FK2/15>, Harvard Dataverse, V1

Social Scientists rarely take full advantage of the information available in their statistical results. As a consequence, they miss opportunities to present quantities that are of greatest substantive...

Expanded Advanced Search

 **Dataverse** Beta Q About Support - Feedback Sign Up Log In

Dataverses

Name

Affiliation

Description

Datasets: Citation Metadata

Datasets: Geospatial Metadata

Datasets: Social Science and Humanities Metadata

Datasets: Astronomy and Astrophysics Metadata

Datasets: Biomedical Metadata

Files

Name

Description

File Type

Variable Name

Variable Label

Ability to search on specific dataverses, dataset metadata fields across various domains, and files (variables).

Visualize & Analyze Data: TwoRavens

- Integrated with Dataverse & Zelig (statistical software)
- From beginners up to advanced stats users
- Explore data, view descriptive statistics, and estimate statistical models for files in datasets



keyal

R call: func(var)



Estimate

Data Selection

Variables

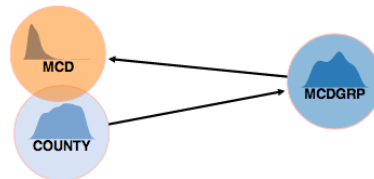
Subset

MCDGRP

COUNTY

MCD

- ● +



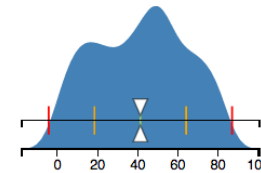
Model Selection

Models

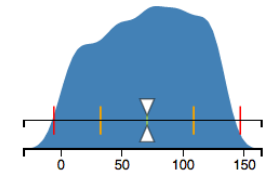
Set Covar.

Results

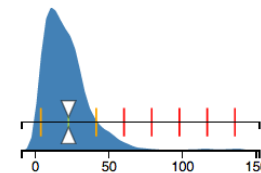
MCDGRP

x: 41.35
x1: 41.35

COUNTY

x: 70.55
x1: 70.55

MCD

x: 22.54
x1: 22.54

WorldMap Integration

1. Upload a file containing geographic data into Dataverse
2. Easily visualize the data on the WorldMap system.
3. WorldMap layer embedded into dataset in Dataverse

Dataverse Beta About Support Feedback Pete Privileged 5

Transportation to Work, ACS 2008-2012 estimates Draft Unpublished

O'Brian, 2014, "Transportation to Work, ACS 2008-2012 estimates", <http://dx.doi.org/10.5072/FK2/275>, Root Dataverse, DRAFT VERSION

Transportation method and travel time to work in Massachusetts census tracts. Data collected from the American Community Survey, 2008-2012 estimates. Note: Includes multiple visualizations. Right click and go to 'Styles' to access.

Keyword transportation; Massachusetts; commute
Subject Social Sciences

Files Metadata Versions **Map Layer appears with Shapefile**

Access the WorldMap layer Upload + Edit Files

transportation_to_work_v24.zip
Shapefile as ZIP Archive, MD5: 8bc85dc411c2341fa4b2267f62b1d07

View on WorldMap

Ability to "Re-Map" It

Re-Map It Download



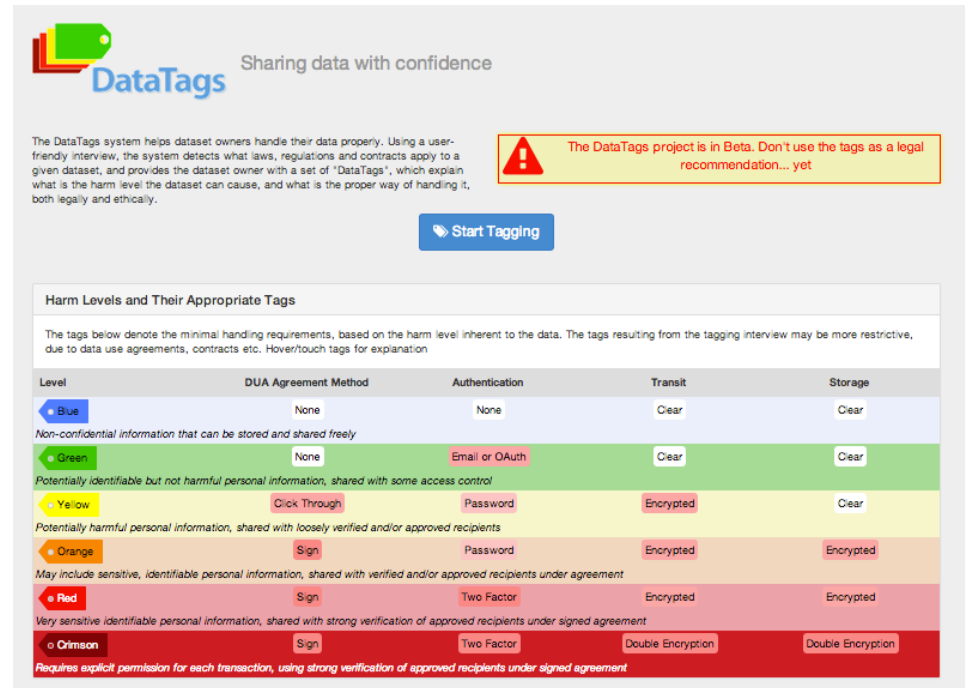
Read more on: [Data Science Blog](#).

After 4.0

- Sharing Privacy Sensitive Data
- Secure Dataverse
- **DataTags** (questionnaires based on privacy laws)

- ORCID Integration (API)

ORCID



DataTags Sharing data with confidence

The DataTags system helps dataset owners handle their data properly. Using a user-friendly interview, the system detects what laws, regulations and contracts apply to a given dataset, and provides the dataset owner with a set of 'DataTags', which explain what is the harm level the dataset can cause, and what is the proper way of handling it, both legally and ethically.

Warning: The DataTags project is in Beta. Don't use the tags as a legal recommendation... yet

[Start Tagging](#)

Harm Levels and Their Appropriate Tags

The tags below denote 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
Blue	None	None	Clear	Clear
<i>Non-confidential information that can be stored and shared freely</i>				
Green	None	Email or OAuth	Clear	Clear
<i>Potentially identifiable but not harmful personal information, shared with some access control</i>				
Yellow	Click Through	Password	Encrypted	Clear
<i>Potentially harmful personal information, shared with loosely verified and/or approved recipients</i>				
Orange	Sign	Password	Encrypted	Encrypted
<i>May include sensitive, identifiable personal information, shared with verified and/or approved recipients under agreement</i>				
Red	Sign	Two Factor	Encrypted	Encrypted
<i>Very sensitive identifiable personal information, shared with strong verification of approved recipients under signed agreement</i>				
Crimson	Sign	Two Factor	Double Encryption	Double Encryption
<i>Requires explicit permission for each transaction, using strong verification of approved recipients under signed agreement</i>				

Longer-Term

- Large-scale datasets (efficient storage)
- Ensuring long-term preservation for more file formats (e.g., Archivematica)

Get Involved: Dataverse Community

- Let us know your thoughts on [Dataverse 4.0 Beta](#) in the [Dataverse Google Group](#).
- Sign up to participate in usability testing of Dataverse 4.0 Beta by filling out this [form](#).
- Contribute to our code or scripts: [GitHub Pull Requests](#).
- Read our [Data Science Blog](#) for any upcoming updates and notifications.



Thank You!

Eleni Castro, Research Coordinator
IQSS, Harvard University
ecastro@fas.harvard.edu

Dataverse 4.0 Demo:
<http://dataverse-demo.iq.harvard.edu/>
Dataverse Twitter: @thedataorg

