

# Data Publishing with Dataverse



<https://flic.kr/p/7QeoWp>

Eleni Castro, Research Coordinator  
The Institute for Quantitative Social Science (IQSS), Harvard University  
@thedataorg



# Data Science Team

## Researchers



Statistical  
Innovation

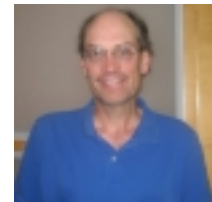
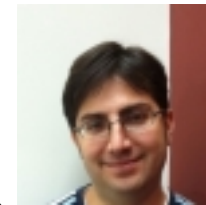
## Data Science Applications and Tools

Data Curation &  
Stewardship

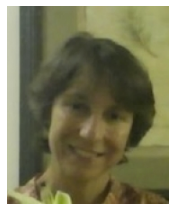
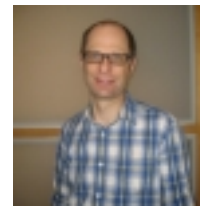


## Information Scientists

Tool Building &  
Computer Science



## Software Engineers



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



# What is the Dataverse Project?

Software framework for Research Data (since 2006):

- Publishing
- Citing
- Analyzing
- Preserving

Open source on [github](https://github.com).

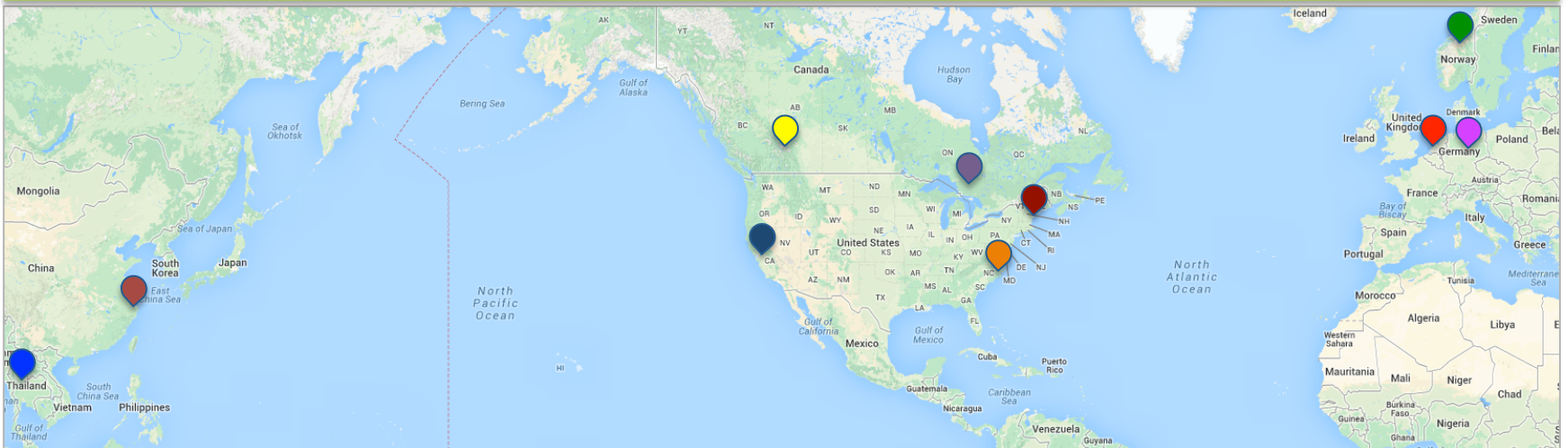


**Provides incentives for researchers to share:**

- Recognition & credit via **data citations**
- Control over data & branding
- Fulfill journal data availability and funder requirements.

# Who Uses Dataverse ?

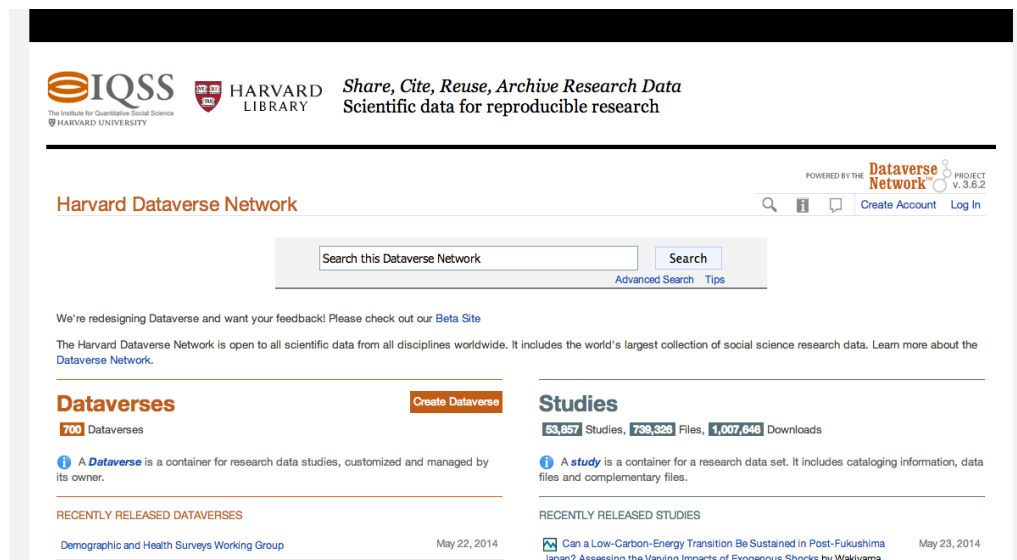
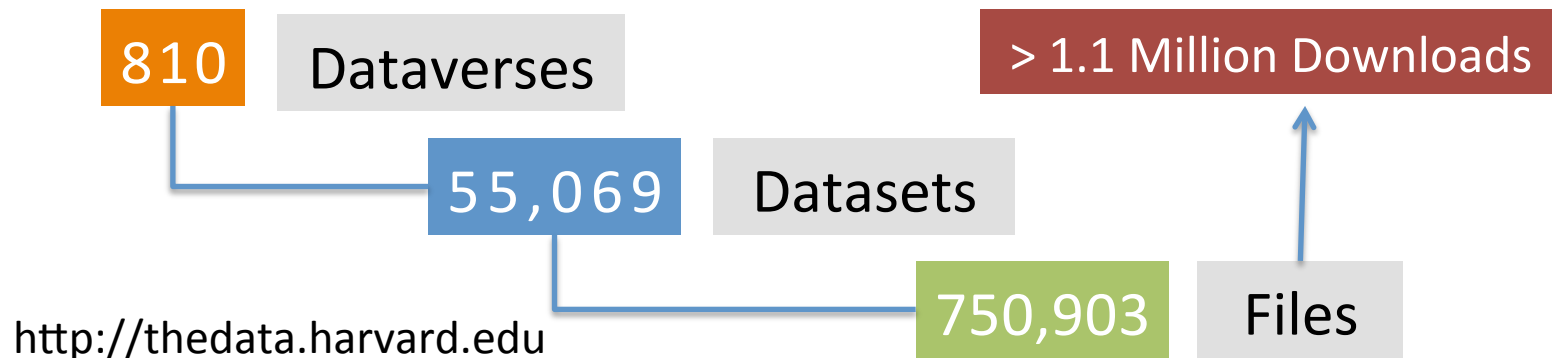
## Worldwide Dataverse Installations



**Institutions** can setup/host their own Dataverse installation (UNC ODUM, Fudan Univ, Scholars Portal, DANS, etc) and within them can have dataverses for a variety of users (across all research domains): Researchers, Projects, Journals, etc.

# Harvard Dataverse

Open to all researchers; general community repository instance:





Harvard Dataverse Network >

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

## National Digital Stewardship Alliance Dataverse

[Search](#) [i](#) [Q](#) [Create Account](#) [Log In](#)

Data archive for data collected by the national digital stewardship alliance.

**National Digital Stewardship Alliance**

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

Sort By:


Studies: **3** | Downloads: **20**


<a href="#">NDSA Web Archiving Surveys</a> by Content Working Group Description: The NDSA conducted two surveys of organizations in the United States that are actively involved in, or planning to start, programs to archive content from the web. The goal of the survey was to better understand the landscape of web archiv... <a href="#">Continue</a> [+] Distribution Date: October 14, 2014	doi:10.7910/DVN/27593 8 downloads Last Released: Oct 23, 2014
<a href="#">Replication data for: Staffing for Effective Digital Preservation: An NDSA Report</a> by National Digital Stewardship Alliance (NDSA). Standards and Practices Working Group. Description: Businesses, cultural memory institutions, repositories, and government bodies seeking to preserve digital assets responsibly face significant staffing challenges. How many staff and what types of positions are required? What skills, educat... <a href="#">Continue</a> [+] Distribution Date: October 01, 2013	doi:10.7910/DVN/27562 Last Released: Oct 9, 2014
<a href="#">Data for NDSA Storage Report 2011</a> by Altman, Micah; Bailey, Jefferson; Cariani, Karen; Gallinger, Michelle; Owens, Trevor Description: The NDSA storage survey aims to gather information on preservation storage systems. The respondents represent a diverse cross section of organizations working with preservation storage systems including university libraries, consortia, ins... <a href="#">Continue</a> [+] Distribution Date: 2012 Related Publications: Micah Altman, Jefferson Bailey, Karen Cariani, Michelle Gallinger, Jane Mandelbaum, Trevor Owens, and the NDSA Infrastructure Working Group, 2013, Reflections on National Digital Stewardship Alliance Member Approaches to Preservation Storage Technologies: NDSA Storage Report, D-Lib Magazine. Forthcoming.	hdl:1902.1/19768 12 downloads Last Released: Feb 10, 2013



## Harvard Dataverse

Harvard Dataverse

 [Email Dataverse Contact](#)

 [Edit Dataverse](#)

\* 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](http://thedata.harvard.edu) \*\* Only datasets that have no restrictions and are non-identifiable data can be uploaded to Beta.

### Featured Datasets



Astrophysics Department  
Dataverse



Stem Cell Research Dataverse



Ubiquity Press Dataverse



Election Data Dataverse




# Dataverse 4.0 (December 2014)

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

☒  **Dataverses (10)**

11 to 20 of 30 results

 Sort

[<<](#) [< Previous](#) [1](#) [2](#) [3](#) [Next >](#) [>>](#)

☒  **Datasets (20)**

☐  **Files (37)**

#### Publication Status

[Published \(27\)](#)

[Draft \(2\)](#)

[Unpublished \(2\)](#)

#### Affiliation

[Harvard University \(9\)](#)

[COMPLETE \(3\)](#)

[California Institute of  
Technology \(3\)](#)

[University of Colorado \(3\)](#)

[University of Texas \(3\)](#)

[More...](#)

#### GBT Ophiuchus HI Datacube



Aug 13, 2014 [COMPLETE Dataverse](#)

COMPLETE team, 2014, "GBT Ophiuchus HI Datacube", <http://dx.doi.org/10.5072/FK2/19>, Harvard Dataverse, V1

21 cm HI maps obtained at the 100 m NRAO Green Bank Telescope. The line profiles of HI in Ophiuchus reveal a strong and extensive HI Narrow SelfAbsorption (HINSA; Li & Goldsmith 2003) component, which...

#### GBT Perseus HI Datacube



Aug 13, 2014 [COMPLETE Dataverse](#)

COMPLETE team, 2014, "GBT Perseus HI Datacube", <http://dx.doi.org/10.5072/FK2/20>, Harvard Dataverse, V1

21 cm HI maps obtained at the 100 m NRAO Green Bank Telescope. The main component of HI emission toward the line of sight of Perseus is centered around 4 to 8 km s<sup>-1</sup>, with the velocity of peak emissio...

# Dataverse Best Practices (1)

- Standard Metadata Schemas
  - DDI, FGDC & OAI DC
  - Coming in 4.0: DataCite 3.1, ISA-Tab (biomedical), and VO Resource (astronomy) → **Export to JSON & XML**
- Formal Data Citation (Altman & King, 2007)
  - Endorse + comply w/ Joint Declaration of Data Citation Principles (FORCE11)
  - Versioning and File Fixity
- Persistent IDs: Handles & DOI (DataCite/EZID)



# Dataverse Best Practices (2)

- Preservation format conversion for tabular data (extract column/variable metadata)
- File Fixity:
  - UNF (Altman, 2008) for tabular data
  - MD5 checksums for other files

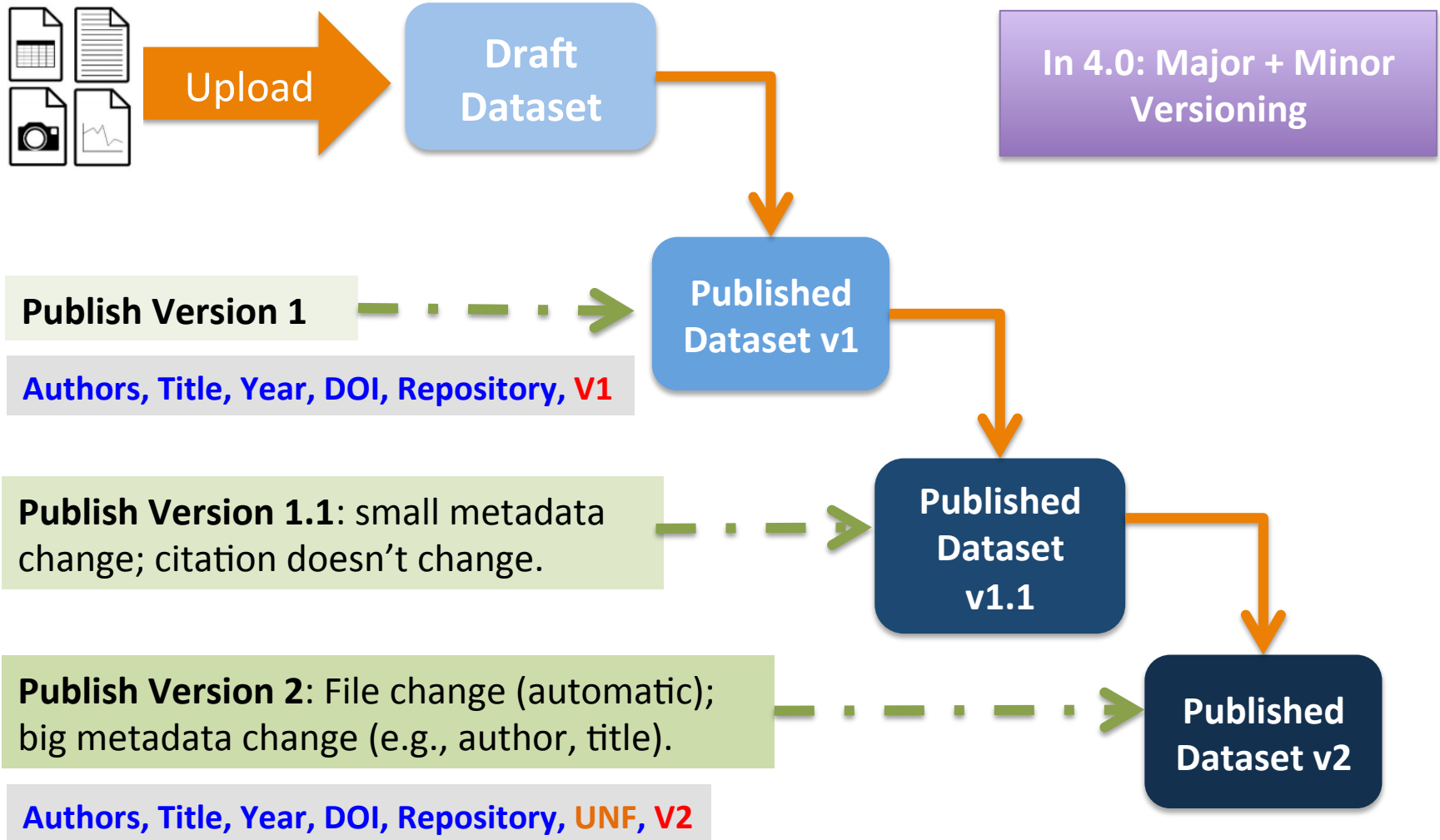
# Dataverse Best Practices (3)

- LOCKSS (replication of files)
  - Data-PASS: (ICPSR, ODUM, NARA, ROPER,...)



- OAI-PMH: Harvesting metadata (DC, DDI)
  - From other Dataverse installations
  - From other OAI-DC compliant repositories
- If necessary: Deaccession a Dataset

# Rigorous Data Publishing Workflows



# Dataset Versioning (1)

The screenshot shows the Dataverse Beta interface. At the top, there's a navigation bar with the Dataverse logo, a search icon, and links for About, Support, Feedback, and a user profile (admin Privileged 20). Below this is a header for 'Gary King Dataverse (Harvard University)' with a link to 'Gary King's Personal Site'. The main content area displays the dataset '10 Million International Dyadic Events' by King, Gary; Lowe, Will, 2014. A modal dialog titled 'Publish Dataset' is open, asking 'Are you sure you want to republish your dataset?'. It offers two options: 'Major Release (3.0)' (selected) and 'Minor Release (2.1)'. Below the options are 'Continue' and 'Cancel' buttons. The background dataset page includes buttons for 'Email Dataset Contact', 'View Published Version', 'Publish Dataset', and 'Edit Dataset'. The dataset description mentions that events are aggregated to the month or year. The 'Related Publication' section lists a paper by King, Gary; Lowe, Will, 2003, 'An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design'.

Dataverse Beta

Search About Support Feedback admin Privileged 20

Gary King Dataverse (Harvard University) Gary King's Personal Site

Harvard Dataverse > Gary King Dataverse > 10 Million International Dyadic Events

Email Dataset Contact View Published Version Publish Dataset Edit Dataset

### 10 Million International Dyadic Events

King, Gary; Lowe, Will, 2014, "10 Million International Dyadic Events"

When the Palestinians launch a mortar attack into Israel, the data is aggregated to the month or year. The data available here in the dataset is summarized in the data as "Actor A does something to Actor B". The data is coded in an ontology of about 200 types of actions. The data is available in a draft version.

**Keyword** events; palestine

**Subject** Social Sciences

**Related Publication** King, Gary; Lowe, Will, 2003, "An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design" International Organization, Vol. 57, No. 03, pp. 617-642:

Files Metadata Versions

### Publish Dataset

⚠ Are you sure you want to republish your dataset?


Select version number:

☒ Major Release (3.0) ☐ Minor Release (2.1)

Continue Cancel



# Dataset Versioning (2)

 Dataverse Beta

Q About Support ▾ Feedback admin Privileged 26 ▾

## 10 Million International Dyadic Events

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 year. The data available here include almost 10 million individual events, each coded to the exact day they occur or become known. Each event is summarized in the data as "Actor A does something to Actor B", with Actors A and B recording about 450 countries and other (within-country) actors and "does something to" coded in an ontology of about 200 types of actions. The data are coded by computer from millions of Reuters news reports.

Keyword

events; palestine

Subject

Social Sciences

Related Publication

King, Gary; Lowe, Will, 2003, "An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design" International Organization, Vol. 57, No. 03, pp. 617-642:

Files

Metadata

Versions

Show Differences

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

Powered by  The Dataverse Project v. 4.0

# Deaccession Data in 4.0

Before a Dataset is published the DOI is private (reserved). Only when published is it made public & searchable.

## In accordance w/ **Data Citation Principle #6**

Persistence: A Published Dataset **cannot** be deleted; only deaccessioned, with a reason.

You can Deaccession (in 4.0):

1. **version(s)** of a Dataset, or
2. an **entire** Dataset.

# Deaccession A Version (1)

**Dataverse** Beta

Search About Support Feedback admin Privileged

Sample Dataverse (IQSS)

Harvard Dataverse > Sample Dataverse > Sample Dataset For Demo Purposes Only

Email Dataset Contact Edit Dataset

**Sample Dataset For Demo Purposes Only**

King, Gary, 2014, "Sample Dataset For Demo Purposes Only", <http://dx.doi.org/10.5072/FK2/577>, Harvard Dataverse, V2

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam accumsan velit lorem, sit amet dapibus erat condimentum vel dapibus metus, ac malesuada erat. Morbi tortor metus, bibendum ut luctus vitae, rhoncus a odio. Ut sed molestie ligula. Donec adipiscing mi eu lacus vulputate elementum. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas.

**Keyword** sample; demo

**Subject** Social Sciences

Files Metadata Versions

+ Upload + Edit Files

**garyking.jpg**  
JPEG Image, MD5: eb947c4c2474492be147c07155c7c556  
Download

**harvard.png**  
PNG Image, MD5: 1203a0d9f910a2a48e726c44738518bb  
Download

Ex. This file was added in v2 and has identifiable information.

# Deaccession A Version (2)

**Dataverse Beta**

Sample Dataverse (IQSS)

Harvard Dataverse > Sample Dataset

King, Gary, 2014, "Sample Dataset"

Keyword  
Subject

Files Metadata

garyking.jpeg  
harvard.png

Deaccession Dataset

⚠ Once you deaccession your dataset it will not be viewable by the public.

Which version(s) do you want to deaccession?

<input type="checkbox"/>	Version 1.0, July 17, 2014
<input checked="" type="checkbox"/>	Version 2.0, July 17, 2014
<input type="checkbox"/>	Version 1.0, July 17, 2014

What is the reason for deaccession?


Select one...

- Select one...
- There is identifiable data in one or more files
- The research article has been retracted
- The dataset has been transferred to another repository
- IRB request
- Legal issue or Data Usage Agreement
- Not a valid dataset
- Other (Please type reason in space provided below)

Deaccession Cancel



# Deaccession A Version (3)

 **Dataverse** Beta

Q About Support Feedback admin Privileged 4

Sample Dataverse (IQSS)

Harvard Dataverse > Sample Dataverse > Sample Dataset For Demo Purposes Only

Deaccession Landing Page  
Data Citation Principle #6  
Persistence

Email Dataset Contact View Published Version

**Sample Dataset For Demo Purposes Only** Deaccessioned


King, Gary, 2014, "Sample Dataset For Demo Purposes Only", <http://dx.doi.org/10.5072/FK2/577>, Harvard Dataverse, V2, DEACCESSIONED VERSION ?

**Deaccession Reason:** There is identifiable data in one or more files. Gary King can be identified by the photo, and file name, added in this version.

Versions

2.0	Deaccessioned Reason: There is identifiable data in one or more files. Gary King can be identified by the photo, and file name, added in this version.	admin Privileged	July 17, 2014
1.0	This is the first published version.	admin Privileged	July 17, 2014

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

Powered by  **Dataverse**  
Project v. 4.0

# Data Publishing After 4.0 (2015)

## Publishing Privacy Sensitive Data

- Secure Dataverse
- [DataTags \(demo\)](#) (based on Privacy Laws and DUAs)

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.

**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

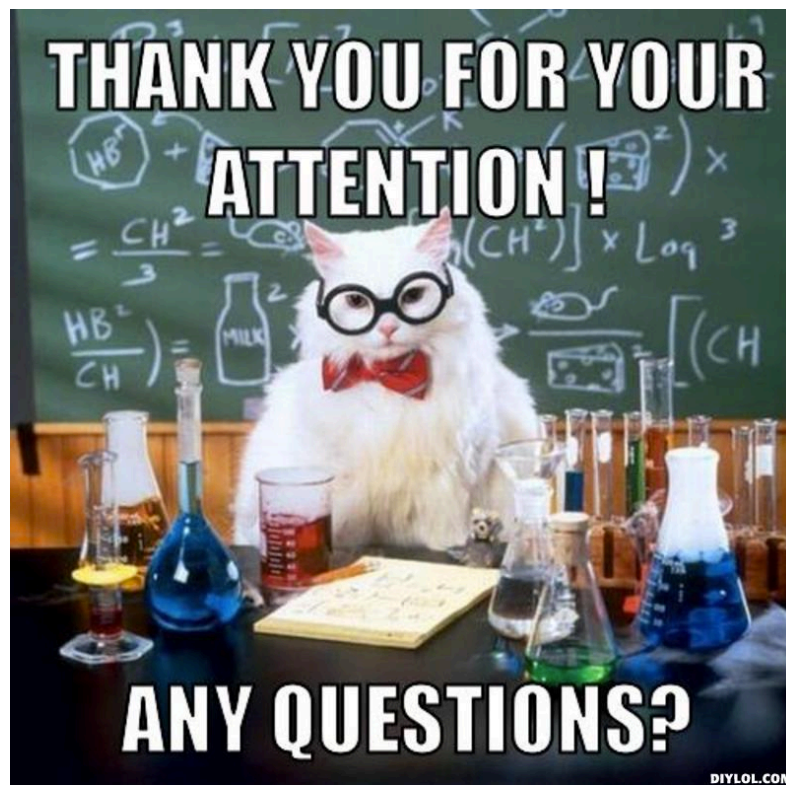
Level	DUA Agreement Method	Authentication	Transit	Storage
<b>Blue</b>	None	None	Clear	Clear
Non-confidential information that can be stored and shared freely				
<b>Green</b>	None	Email or OAuth	Clear	Clear
Potentially identifiable but not harmful personal information, shared with some access control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

Logos: NSF, Harvard University, Berkman Center for Internet & Society at Harvard University, DATA PRIVACY LAB

- Data Citation Provenance Registry (w/ SEAS funded by NSF)
- Author Disambiguation: ORCID Integration (API)
- Long-term preservation for more file formats (beyond tabular)
  - UDFR, Archivematica,...

# Thank you!

Contact: [ecastro@fas.harvard.edu](mailto:ecastro@fas.harvard.edu)



## References

- Altman M. A Fingerprint Method for Verification of Scientific Data. In A Fingerprint Method for Verification of Scientific Data. Springer-Verlag; 2008.
- Altman M, King G. A Proposed Standard for the Scholarly Citation of Quantitative Data. D-Lib Magazine [Internet]. 2007;13(3/4).