

Update on Data Publishing With Dataverse

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

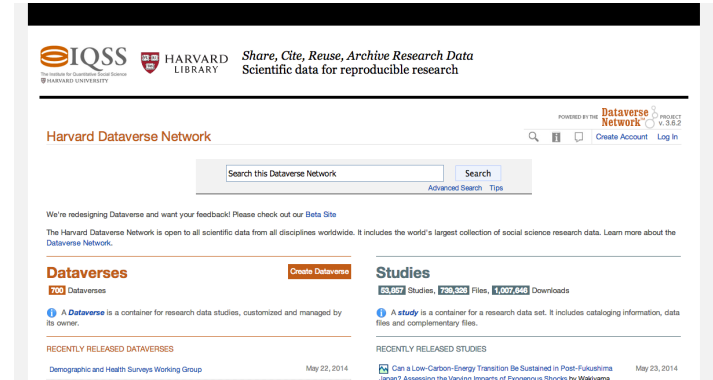
DataCite Annual 2014
Nancy, France
August 25, 2014

Introduction to Dataverse

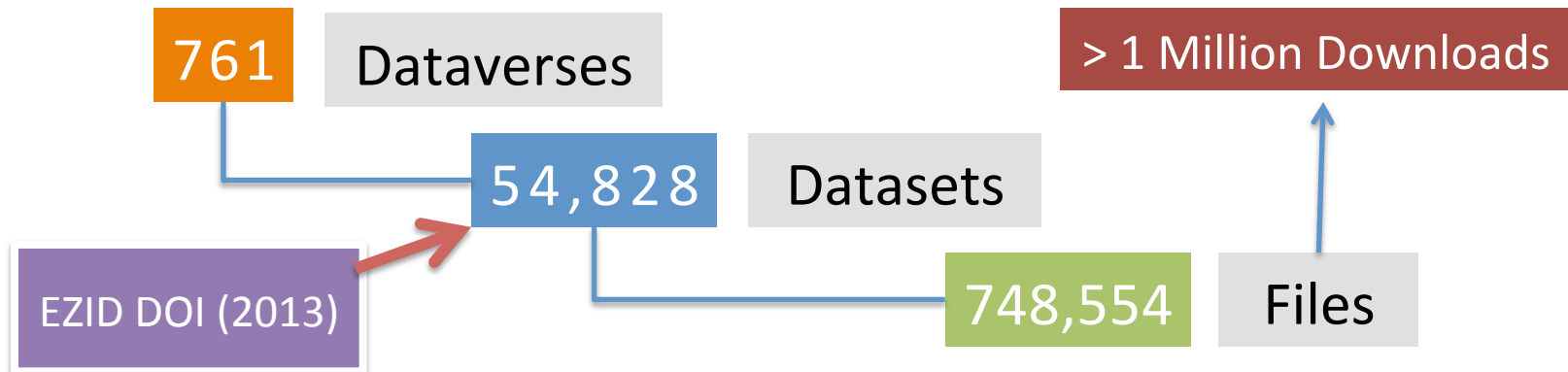
Software framework for publishing, citing and preserving research data (open source on [github](https://github.com) for others to install)

Provides incentives for researchers to share:

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

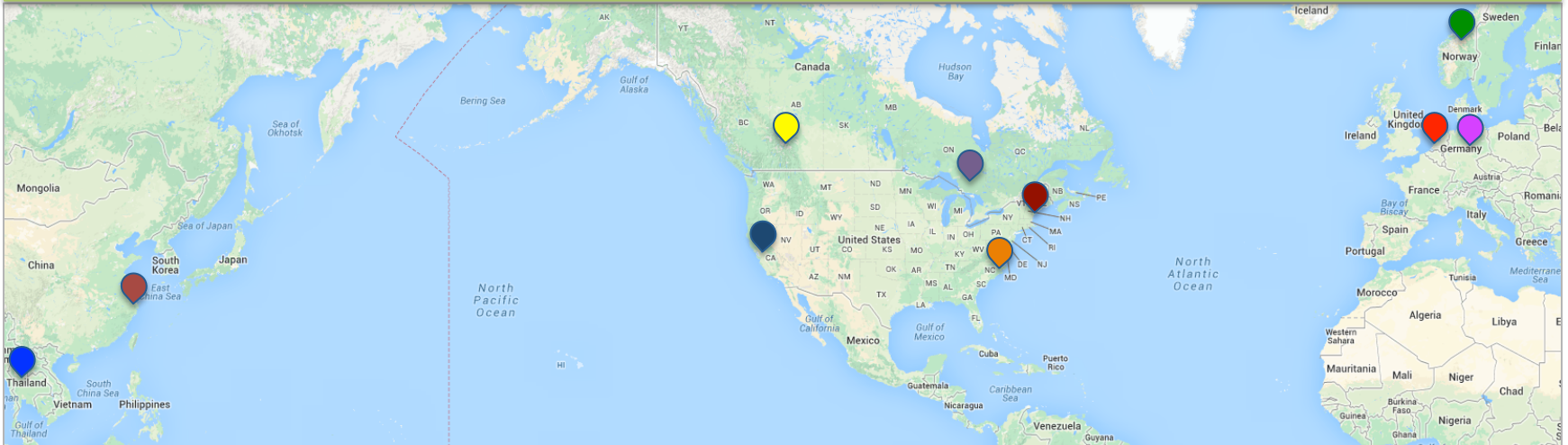


Harvard Dataverse (open to all; repository instance at Harvard) currently has:



Who's Using Dataverse?

Worldwide Dataverse Installations



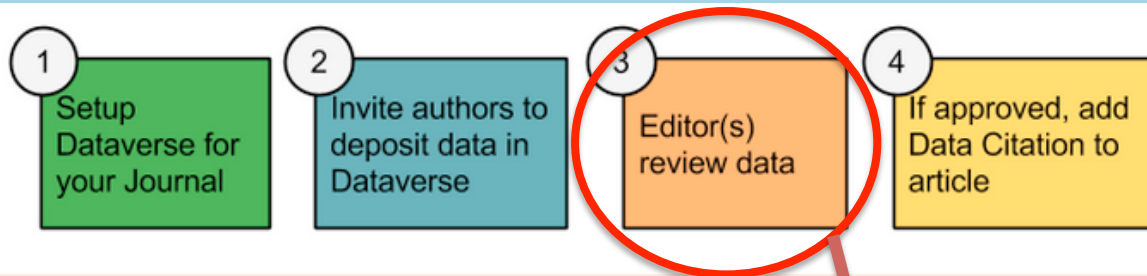
Institutions can setup & host their own Dataverse installation (e.g., Odum, OCUL, DANS, Fudan, etc) and within them can support datasets from a variety of users (across all research domains): Researchers, Projects, Departments, Journals, etc.

Journals Publishing Data w/ Dataverse

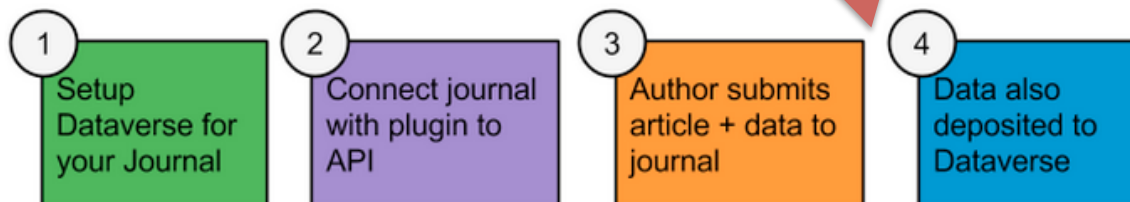
Option A. Journals include Dataverse as a Recommended Repository



Option B. Authors Contribute Directly to a Journal Dataverse



Option C. Seamless Integration btw Journal + Dataverse (e.g., OJS)

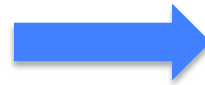


OJS Journal



The screenshot shows the Open Health Data journal homepage. It features a navigation menu with links for Home, About, Editorial Board, Contact, Repositories, FAQ, Articles, Archive, and Social Collections. A central text block describes the journal as a peer-reviewed platform for open health research. Below this, there are sections for 'Most recent articles' and 'Most popular articles', each displaying a list of recent publications with their titles and authors.

Citation
to Data



Citation
to Article

Journal Dataverse



The screenshot shows the Open Health Data Dataverse interface. It includes a search bar, a 'Search Studies' button, and a list of available datasets. The first dataset listed is 'Data from "Social determinants of unmet hospitalisation need amongst the poor in Andhra Pradesh, India: A cross-sectional study."' by Nagulapalli, Srikanth. The interface also displays the number of studies (1) and downloads (46) for the selected dataset.

Details/Updates: 2 Year Project 2012-2014

- Integrating w/ PKP's Open Journal Systems (Data Deposit API).
- Pilot with ~ 50 journals + expanding outreach (100s) .
- OJS' Dataverse plugin now available with latest OJS release.
- **Future:** Embed Dataverse widgets into journal article.

OJS Plugin: Journal Data Policies Boilerplate Templates

Home > User > Journal Management > **Dataverse Plugin**

Dataverse Plugin

CONNECT SELECT DATAVERSE **SETTINGS**

Data Policies

Configure data policies.

Data Availability Policy

General data availability policy for the journal. This will appear in About the Journal.

Super Plugin Testing Happy Hour requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive. Super Plugin Testing Happy Hour recommends the [Harvard Dataverse](#), which is free and open to all researchers worldwide to share, cite, reuse and archive research data. Data are important products of the scientific enterprise, and they should be preserved and usable for decades in the future. Authors may elect to have the data publicly available at time of publication. Exceptions may be granted at the discretion of the editor, especially for sensitive information such as human subject data or the location of endangered species. Any exceptions should be documented in a statement in the public article. (Adapted from [Joint Data Archiving Policy \(JDAP\)](#)).



Including Guidelines for:

- 1) Authors (w/ data citation)
- 2) Reviewers

Read full Data Policies / Guidelines Template: <http://bit.ly/1xkLjoZ>

OJS Plugin: Author Manuscript + Data Submission

Dataverse Plugin: Publish Data

Research data can be submitted for publication by depositing supplementary files in the journal's Dataverse or by supplying a citation to data already in a repository.

- Treat uploaded file as supplementary file. File **will not** be deposited in Dataverse.
- Deposit file in Dataverse. Author accepts and agrees to Dataverse [terms of use](#).

External Data Citation

If the data to be published with this submission is already in a different repository, please provide a data citation.

Supplementary File

File Name [6-17-1-SP.pdf](#)
Original file name Analytics All Web Site Data Pages 20121227-20130417.pdf
File Size 117KB
Date uploaded 2014-07-27 10:43 AM

- Present file to reviewers (without metadata), as it will not compromise blind review.

Replace file

No file chosen

Option to: (A) deposit into Dataverse AND/OR; (B) if data is already in a repository can include the data citation (w/ persistent URL/identifier).



OJS Plugin: Editor Reviews Article + Data

Home > User > Editor > Submissions > #6 > Review

#6 Review

SUMMARY **REVIEW** EDITING HISTORY REFERENCES

Submission

Authors Eleni Castro, Alex Garnett 
Title Testing for Force11 & RDA
Section Articles
Editor axfelix 

Review Version [6-18-1-RV.PDF](#) 2014-07-27

Upload a revised Review Version

Choose File

No file chosen

Upload

Supp. files [6-17-1-SP.PDF](#) 2014-07-27

Present file to reviewers

Record

PeerReview

Round 1

[SELECT REVIEWER](#)

[VIEW REGRETS, CANCELS, PREVIOUS ROUNDS](#)

Editor Decision

Select decision


- ✓ Choose One
- Accept Submission
- Revisions Required
- Resubmit for Review
- Decline Submission

Record Decision

Decision

Notify Author

Review Version

Will Record  No Comments

27

Data Published in Dataverse w/ OJS Plugin

In OJS:

Super Plugin Testing Happy Hour

HOME ABOUT USER HOME SEARCH CURRENT ARCHIVES

Home > User > Editor > Submissions > #6 > Review

#6 Review

Notification

Submission data has been published in Dataverse:

Castro, Eleni; Garnett, Alex, "Testing for Force11 & RDA", <http://dx.doi.org/10.5072/FK2/ZN5ZT> > <http://dx.doi.org/10.5072/FK2/ZN5ZT> V1 [Version]

You are logged in as...

In Dataverse:

Demo Dataverse Network >

Journal of Plugin Testing Dataverse

TESTING FOR FORCE11 & RDA

doi:10.5072/FK2/ZN5ZT

Version: 1 - Released: Sun Jul 27 14:07:12 EDT 2014

POWERED BY THE **Dataverse Network** PROJECT v. 3.6.1

Eleni Castro Log Out

- Edit Cataloging Information
- Edit/Delete File + Information
- Add File(s)
- Permissions
- Create Study Template
- Deaccession

CATALOGING INFORMATION

Data & Analysis Comments (0) Versions

i If you use these data, please add the following citation to your scholarly work

Data Citation

Castro, Eleni; Garnett, Alex, "Testing for Force11 & RDA", <http://dx.doi.org/10.5072/FK2/ZN5ZT> V1 [Version]

Citation Format Print

Publications

Castro, E., & Garnett, A. (in press). Testing for Force11 & RDA. *Super Plugin Testing Happy Hour*.

Data Citation Details

Title Testing for Force11 & RDA

Study Global ID doi:10.5072/FK2/ZN5ZT

2 Options in OJS:

- 1) Dataset Published (with DOI) at **Article Approval**.
- 2) Dataset Published when **Journal Issue is Released**.

OJS Plugin: Article Published w/ Data Citation

Journal of Plugin Testing

[HOME](#) [ABOUT](#) [LOGIN](#) [REGISTER](#) [SEARCH](#) [CURRENT](#) [ARCHIVES](#)

Home > Vol 1, No 1 (1) > **Garnett**

Alex uploads a file

Alex Garnett

Abstract

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Full Text:

[PDF](#)

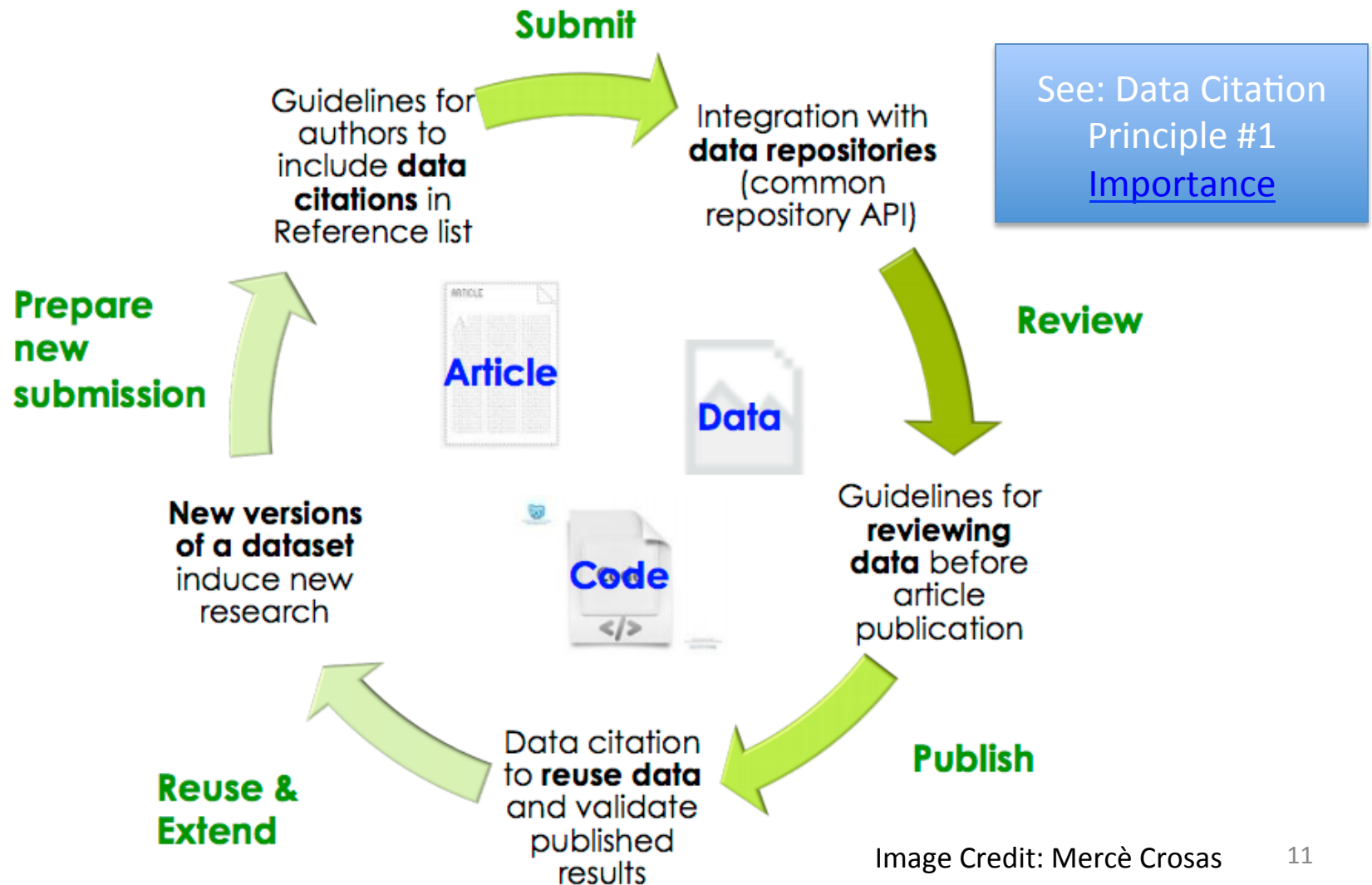
Data citation

Garnett, Alex, 2014, "Replication data for: Alex uploads a file", <http://dx.doi.org/10.5072/FK2/MYRXI>, Harvard Dataverse [Repository], V1 [Version]

Refbacks

There are currently no refbacks.

Towards An Integrated Publishing Lifecycle





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 ** Only datasets that have no restrictions and are non-identifiable data can be uploaded to Beta.

Featured Dataverses

CfA Astrophysics Department Dataverse

HSCI Harvard Stem Cell Institute Stem Cell Research Dataverse

Ju[ubiquity press open data Ubiquity Press Dataverse

Election Data Dataverse

Publishing in 4.0 (Late Fall 2014)

Search this Dataverse... Add Data

- Dataverses (10)
- Datasets (20)
- Files (37)

11 to 20 of 30 results

Sort < Previous 1 2 3 Next >

- 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

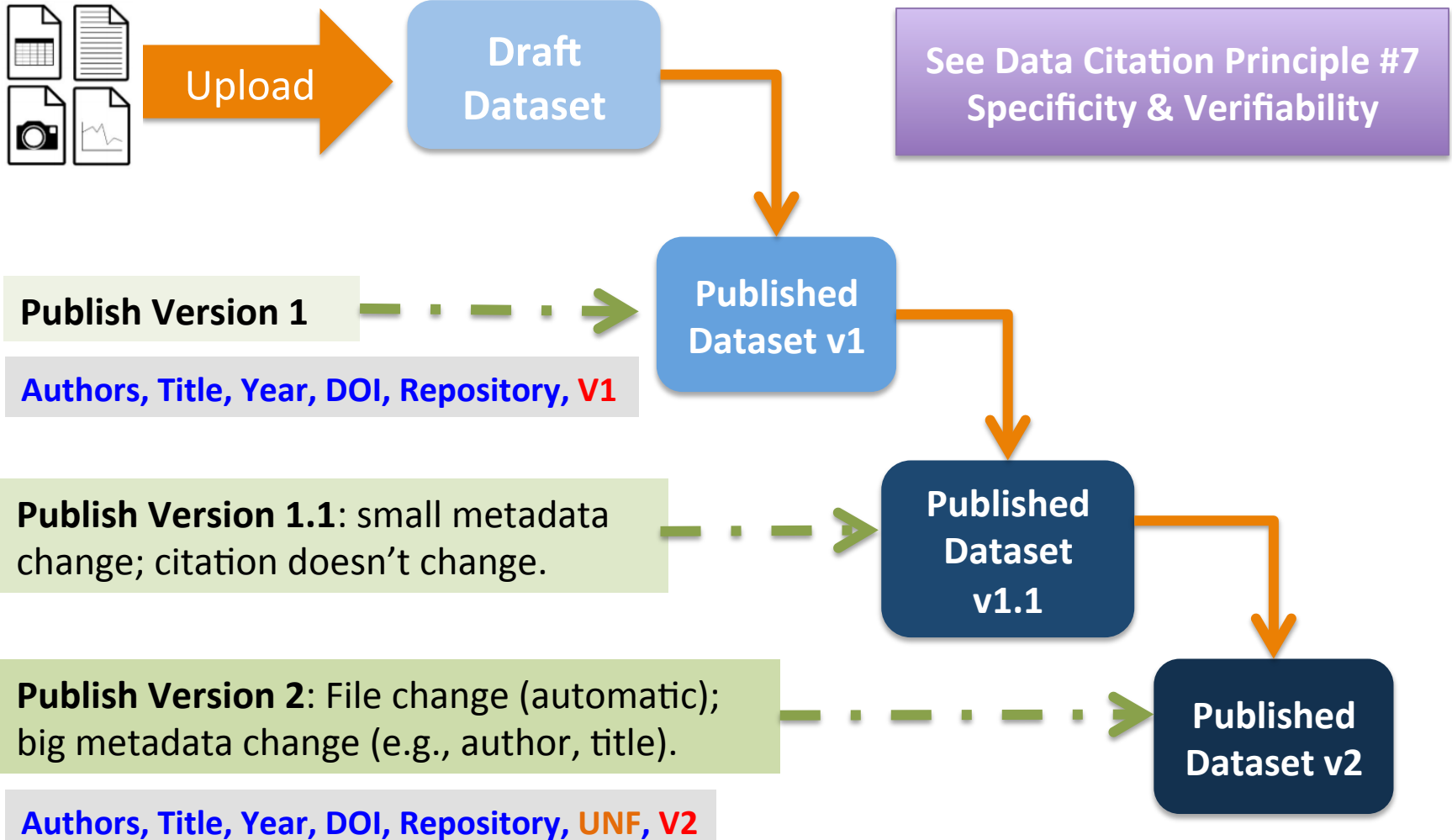


Aug 13, 2014 COMPLETE Dataverse

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-1, with the velocity of peak emissio...

Rigorous Data Publishing Workflows



Dataset Versioning (1)

The screenshot shows the Dataverse interface for a dataset titled "10 Million International Dyadic Events". The user is logged in as "admin Privileged" with 29 notifications. The dataset is currently in a "DRAFT VERSION" state. A modal dialog box titled "Publish Dataset" is open, asking the user to confirm if they want to republish the dataset. The dialog also allows the user to select a version number, with "Major Release (3.0)" selected and "Minor Release (2.1)" as an option. The background shows the dataset's metadata, including keywords like "events; palestine" and "Social Sciences", and a related publication by King, Gary; Lowe, Will, 2003.

Dataverse **Beta** Search About Support Feedback admin Privileged 29

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" Dataverse, DRAFT VERSION

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 the following 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](#)

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


<input type="checkbox"/>	2.1	Additional Citation Metadata: (2 Added, 1 Changed); Show Details	admin Privileged	August 22, 2014
<input type="checkbox"/>	2.0	Files (Added: 3) Show Details	admin Privileged	August 15, 2014
<input type="checkbox"/>	1.0	This is the first published version.	admin Privileged	August 13, 2014

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

Powered by  v. 4.0

Dataset Versioning (3)

Dataverse Beta Q About Support Feedback admin Privileged 28

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

Ex. Added files to a Dataset so it bumped up to a major version change.

Version Differences Details


Viewing: 10 Million International Dyadic Events

	Version: 1.0	Version: 2.0
	Last Updated: Wed Aug 13 17:48:16 EDT 2014	Last Updated: Fri Aug 15 10:37:47 EDT 2014

Files		
File ID: 71		File Name: 1990-1995 IDEA Event Forms.tab File Type: Tab-Delimited Category: Description: IDEA Event Forms Data File
File ID: 72		File Name: 1990-1995 Country Dyads.tab File Type: Tab-Delimited Category: Description: Country Dyads Data File
		File Name: 1990-1994 Data

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

<input type="checkbox"/>	2.1	Additional Citation Metadata: (2 Added, 1 Changed); Show Details	admin Privileged	August 22, 2014
<input type="checkbox"/>	2.0	Files (Added: 3) Show Details	admin Privileged	August 15, 2014



Dataset Versioning (4)



Ex. Added small metadata change to a Dataset so it bumped up to a minor version change.

Version Differences Details

Viewing: 10 Million International Dyadic Events

Version: 2.0 Last Updated: Fri Aug 15 10:37:47 EDT 2014	Version: 2.1 Last Updated: Fri Aug 22 23:54:52 EDT 2014
------------------------------------------------------------	------------------------------------------------------------

Citation Metadata		
Keyword		events; palestine
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: http://gking.harvard.edu/files/abs/infoex-abs.shtml	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: ; url; http://gking.harvard.edu/files/abs/infoex-abs.shtml

Files Metadata **Versions**



Show Differences			
<input type="checkbox"/>	2.1	Additional Citation Metadata: (2 Added, 1 Changed); Show Details	admin Privileged August 22, 2014
<input type="checkbox"/>	2.0	Files (Added: 3) Show Details	admin Privileged August 15, 2014

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

Deaccession Workflow (Step 1)

Dataverse Beta

Search About Support Feedback admin Privileged 4

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 Workflow (Step 2)

Dataverse Beta

Search About Support Feedback admin Privileged 4

Sample Dataset (IQSS)

Harvard Dataset > S

Sample Dataset

King, Gary, 2014, "Sam

Lorem ipsum dolor sit a
dapibus metus, ac male
adipiscing mi eu lacus v

Keyword

Subject

Files Metadata

Upload + Edit Files

Download

Download

garyking
JPEG Im

harvard.png
PNG Image, MD5: 1203a0d9f910a2a48e726c44738518bb

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"/>	
<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 Workflow (Step 3)

The screenshot shows the Dataverse interface with a modal dialog box titled "Deaccession Dataset". The dialog contains the following elements:

- Warning:** A red triangle icon followed by the text: "Once you deaccession your dataset it will not be viewable by the public."
- Question:** "Which version(s) do you want to deaccession?"
- Version Selection:** A list of two versions:
 - (unselected)
 - Version 2.0, July 17, 2014 (highlighted in blue)
 - Version 1.0, July 17, 2014
- Reason Selection:** "What is the reason for deaccession?" with a dropdown menu showing "There is identifiable data in one or more files".
- Additional Information:** "Please enter additional information about the reason for deaccession." with a text input field containing "Gary King can be identified by the photo, and file name, added in this version."
- URL Field:** "If applicable, please leave a URL where this dataset can be accessed after deaccessioning." with an empty text input field.
- Buttons:** "Deaccession" (green) and "Cancel" (grey).

The background shows the Dataverse interface for a "Sample Dataset" with a "Deaccession Dataset" button highlighted.

Deaccession Workflow (Step 4)

Deaccession Landing Page Data Citation Principle #6 Persistence

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

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 **The 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)

DataTags

Question: Please select one answer

Do the data concern living persons?

No Yes

Full interview

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.

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



Integration with ORCID (API): create ORCID account, connect all Dataverse datasets to ORCID account. (Note: 4.0 will already allow for authors to enter ID.)



Thank you!

Contact: ecastro@fas.harvard.edu

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

Twitter: @thedataorg

