

The **Dataverse** Project

A Practical Example of Data Archiving

Eleni Castro > IQSS Harvard
SHARE 2014 Fall Meeting
October 14, 2014

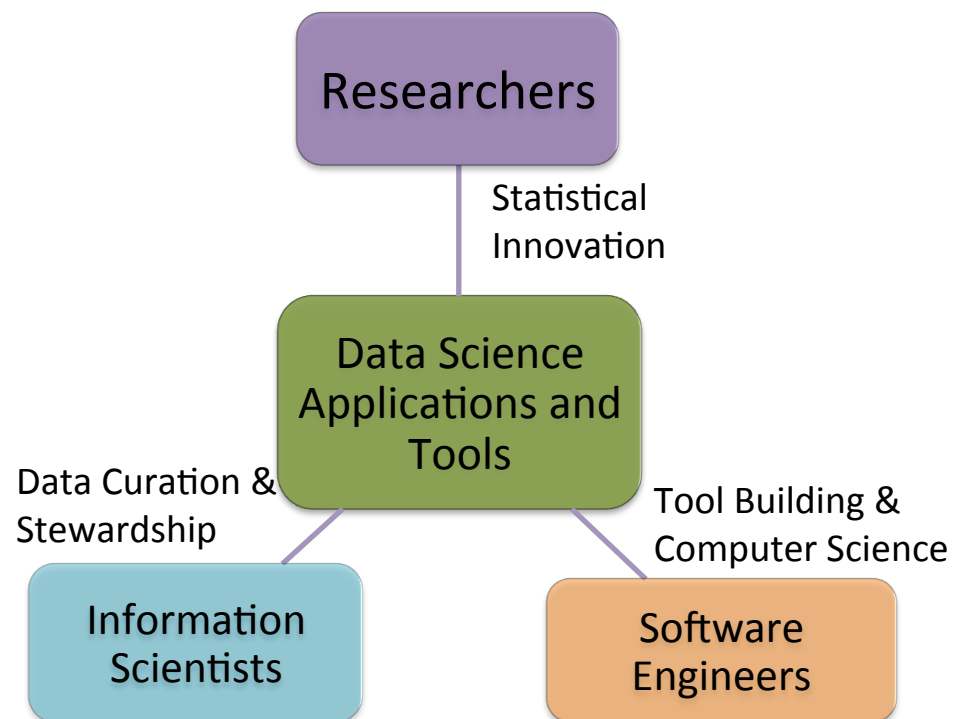
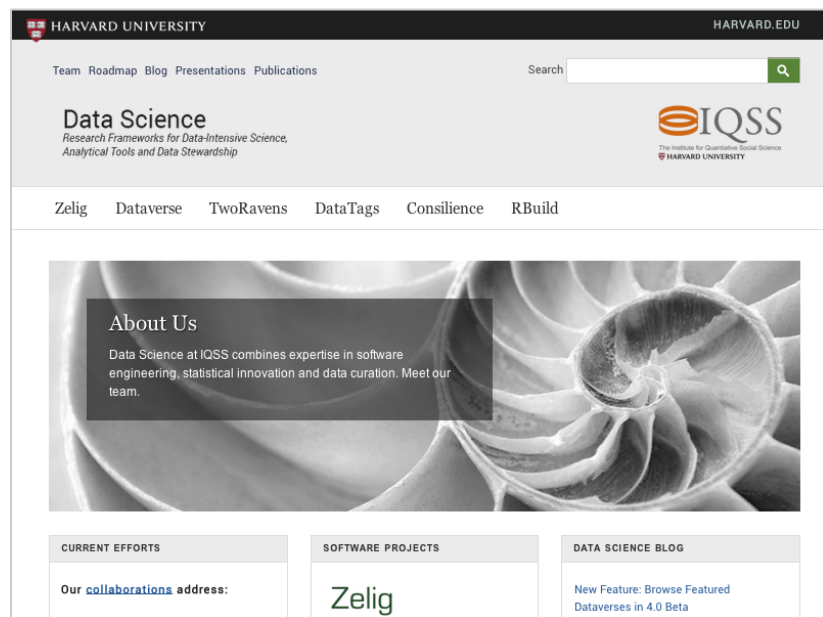


Image: <https://flic.kr/p/7vu434s>



The Institute for Quantitative Social Science

Data Science Team



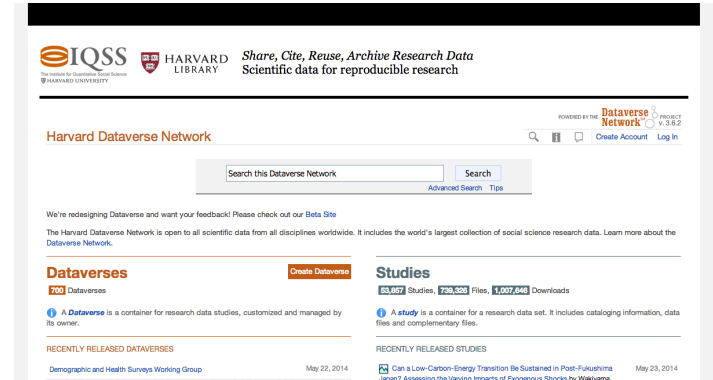
Find out more: <http://datascience.iq.harvard.edu>

Introduction to Dataverse

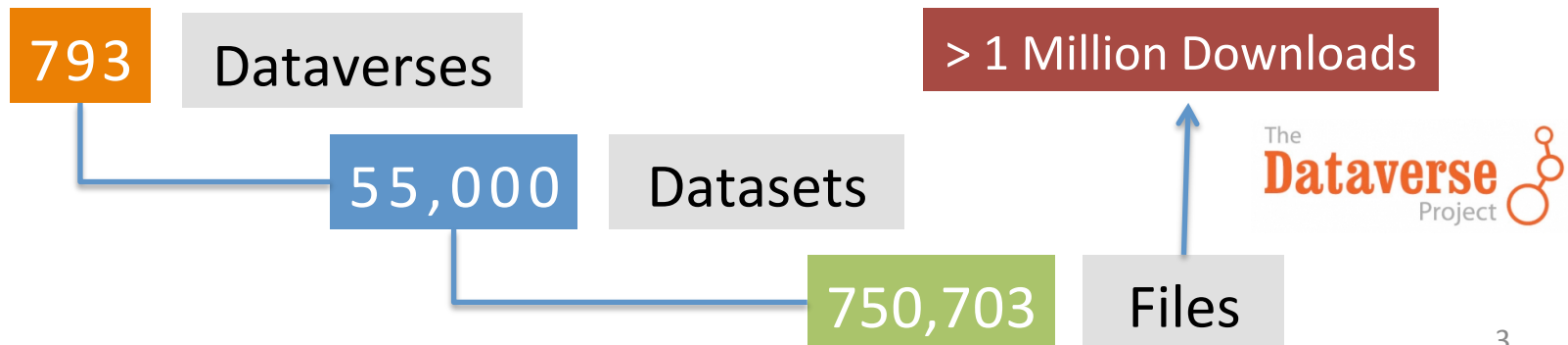
Software framework for publishing, citing and preserving research data
(open source on [github](https://github.com) for others to install)

Provides incentives for researchers to share:

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



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



Why did we launch Dataverse?

- 1 Replication Standard (King, 1995)
- 2 Virtual Data Center (1999-2006)



HARVARD
LIBRARY



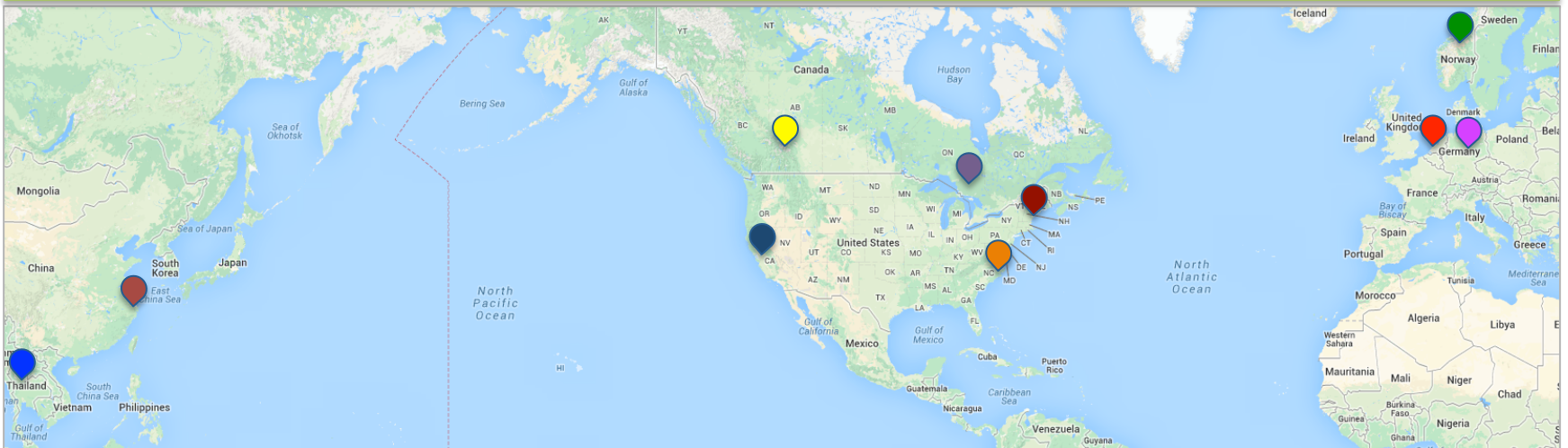
The Institute for Quantitative Social Science

3
The
**Dataverse
Network**[®]
Project
Since 2006




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.

Example of a Scholar's Dataverse

 HARVARD UNIVERSITY

DEPARTMENT OF ECONOMICS | FACULTY OF ARTS AND SCIENCES | HARVARD.EDU

John Y. Campbell

Morton L. and Carole S. Olshan Professor of Economics

(email)

[HOME /](#)

[Harvard Dataverse Network >](#)

John Y. Campbell Dataverse

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

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

[Biography & CV](#)

[Outside Activities](#)

[Data Sets](#)

[Talks and Columns](#)

[Papers](#)

[Courses](#)

[International Household Finance](#)

John Campbell

[Search](#)

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

Sort By: Global ID


Studies: **6** | Downloads: **74353**

[Replication data for: A Multivariate Model of Strategic Asset Allocation](#)

by John Y. Campbell; Yeung L. Chan; and Luis Viceira

Description: We develop an approximate solution method for the optimal consumption and portfolio choice problem of an infinitely long-lived investor with Epstein-Zin utility who faces a set of asset returns described by a vector autoregression in retur...

hdl:1902.1/QBXRSLBQJ

 20089 downloads + analyses

Last Released: Oct 3, 2013

Research Center Dataverse



HARVARD-SMITHSONIAN
CENTER FOR ASTROPHYSICS

EXPLORING THE UNIVERSE

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

Harvard Dataverse Network > CfA Dataverses



[Create Account](#)

[Log In](#)

Search

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

This is the Astronomy data repository at Harvard. It is currently open to all scientific data from astronomical institutions worldwide. Administration and support is provided by the [Harvard-Smithsonian Center for Astrophysics \(CfA\)](#) in collaboration with [Harvard Library \(HL\)](#) and the [Institute for Quantitative Social Science \(IQSS\)](#). Infrastructure is provided by [Harvard University Information Technology Services](#).

The Astronomy Dataverse Network plays an important role in fulfilling your Data Management Plan requirements (e.g. as mandated by NSF), and for providing data re-use and citation opportunities. Find out more about our team by exploring the [Seamless Astronomy](#) and [Wolbach Library](#) teams at the CfA. We *...more >>*

Dataverses

Create Dataverse

30 Dataverses

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

RECENTLY RELEASED DATAVERSES

Studies

105 Studies, **1,737** Files, **55,992** Downloads

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

RECENTLY RELEASED STUDIES

Example of an Institute Dataverse



INTERNATIONAL FOOD POLICY
RESEARCH INSTITUTE
sustainable solutions for ending hunger and poverty

[Staff](#) [Pressroom](#) [Careers](#) [Contact Us](#) [RSS](#)



[OUR WORK](#)

[OUR PRODUCTS](#)

[RESOURCES](#)

[COUNTRIES](#)

[NEWS & EVENTS](#)

Datasets

[Harvard Dataverse Network](#) >

International Food Policy Research Institute (IFPRI) Dataverse

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



[Create Account](#)

[Log In](#)

In collaboration with institutions throughout the world, IFPRI is often involved in the collection of primary data and the compilation and processing of secondary data. The resulting datasets provide a wealth of information at the local (household and community), national, and global levels. IFPRI freely distributes as many of these datasets as possible and encourages their use in research and policy analysis. Please note that the datasets require proper citation and citation information is included with the accompanying documentation to each dataset. Please contact IFPRI-Data@cgiar.org or IFPRI-Library@cgiar.org for questions about IFPRI datasets.

IFPRI Data by Type and Region

- Country Level
- Geospatial
- Household Surveys
- Institution-level Surveys
- Regional
- Social Accounting Matrix (SAM)

Search Studies

Search

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

IFPRI Data by Type and Region

Sort By:

Global ID

Studies: **110** | Downloads:

30681

[1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#)

[Total and Partial Factor Productivity in Developing Countries](#)
by International Food Policy Research Institute

hdl:1902.1/20518

93 downloads

Example of a Journal Dataverse

OXFORD JOURNALS

CONTACT US

MY BASKET

MY ACCOUNT

POLITICAL ANALYSIS

ABOUT THIS JOURNAL

CONTACT THIS JOURNAL

SUBSCRIPTIONS

CURRENT ISSUE

ARCHIVE

SEARCH

[Oxford Journals](#) > [Social Sciences](#) > Political Analysis

Harvard Dataverse Network >

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

Political Analysis Dataverse

The Society for Political Methodology, the Political Methodology Section of the American Political Science Association, and the central web site for the political methodology community. The primary purpose of this site is to serve as the gateway to the Working Paper archive. There is also information on our Conference page covering our annual summer meetings held continuously for over twenty years. This site also serves as the gateway to our publications, including The Political Methodologist, our newsletter, and Political Analysis, the official journal of the section. Additionally, we house a collection of Syllabi for undergraduate and graduate courses in political methodology. Please direct questions to politicalanalysis@hss.caltech.edu.

OPEN DATaverse
Create an account to add your own study to this dataverse. Already have an account? [Log in](#).

Political Analysis

Forthcoming

Volume 08, 1999 - 2000

Volume 09, 2001

Volume 10, 2002

Volume 11, 2003

Volume 12, 2004

Volume 13, 2005

Volume 14, 2006

Volume 15, 2007

Volume 16, 2008

Volume 17, 2009

Volume 18, 2010

Volume 19, 2011

Volume 20, 2012

Search Studies

Search

Advanced Search

Tips

Studies: 221

Sort By: Global ID

Downloads: 14182

1 2 3 4 > >> >>>

Replication data for: Relating latent class assignments to external variables: standard errors for correct inference by Bak, Z

doi:10.7910/DVN/24497
45 downloads + analyses
Last Released: Oct 10, 2014

Replication data for: Multiple Imputation for Continuous and Categorical Data: Comparing Joint Multivariate Normal and Conditional Approaches by Kropko, Jonathan; Goodrich, Ben; Gelman, Andrew; Hill, Jennifer

doi:10.7910/DVN/24672
72 downloads + analyses
Last Released: Oct 10, 2014

Description: We consider the relative performance of two common approaches to multiple imputation (MI): joint multivariate

Dataverse for Teaching Replication

Harvard Dataverse Network >

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

Project TIER: Teaching Integrity in Empirical Research Dataverse

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

This dataverse supports a protocol for teaching undergraduates to document the statistical analysis they do for empirical research projects in such a way that their results are completely reproducible and verifiable. The protocol is guided by the principle that the documentation prepared to accompany an empirical research project should be sufficient to allow an independent researcher to replicate easily and exactly every step of the data management and analysis that generated the results reported in the study. You will find in this dataverse examples of the protocol as applied in senior thesis and introductory statistics projects.

We hope that requiring students to follow this protocol will not only teach them how to document their research appropriately, but also instill in them the belief that it is an important professional responsibility to do so. For more information, visit the [Project TIER website](#).

Project TIER

Introductory Statistics

Senior Theses

Search Studies


Search

Advanced Search Tips


Sort By: Global ID

Studies: 10 | Downloads: 176


A Ticket to the Olympics: An Assessment of the "Olympic Effect" on Tourism
by Costanzo, Laura
Description: Previous studies on the topic of the "Olympic Effect" and its impact on tourism reveal both positive and negative returns for host countries as well as unsuccessful bid host countries. The returns experienced, as explained by subject s...[Continue](#) [\[+\]](#)

doi:10.7910/DVN/24231
 44 downloads
Last Released: Jan 13, 2014

Impact of Governmental Characteristics on Economic Prosperity
by Seitz, Colin; Adams, Gaines; Mutt, Nina; Okun, Harry
Description: We are looking into the topic of political regime characteristics and economic standing. Such a topic could lead into exciting discoveries about which type of government and specifically what legal rights each country employs in order to m...[Continue](#) [\[+\]](#)

doi:10.7910/DVN/24252
 18 downloads
Last Released: Jan 14, 2014

The Effect of Sports Participation on Academic Performance
by Brennan, Claire; Finn, Grant; Grunden, Rachel; Lee, Kayoung
Description: For our project we will be comparing the academic performance of athletes and non-athletes, as well as how gender and race can influence academic performance, both for athletes and non-athletes. We have found previous research that discuss...[Continue](#)

doi:10.7910/DVN/24255
 6 downloads
Last Released: Jan 14, 2014

Dataverse Best Practices (1)

- Standard Metadata Schemas
 - DDI (great for social science data) & DC
 - Coming in 4.0: DataCite 3.0, ISA- Tab (biomedical), and VO Resource (astronomy)
- Formal Data Citation (Altman & King, 2007)
 - Endorse + comply w/ Joint Declaration of Data Citation Principles (incl. Crosas)
- Persistent IDs: Handles & DOI (DataCite/EZID)

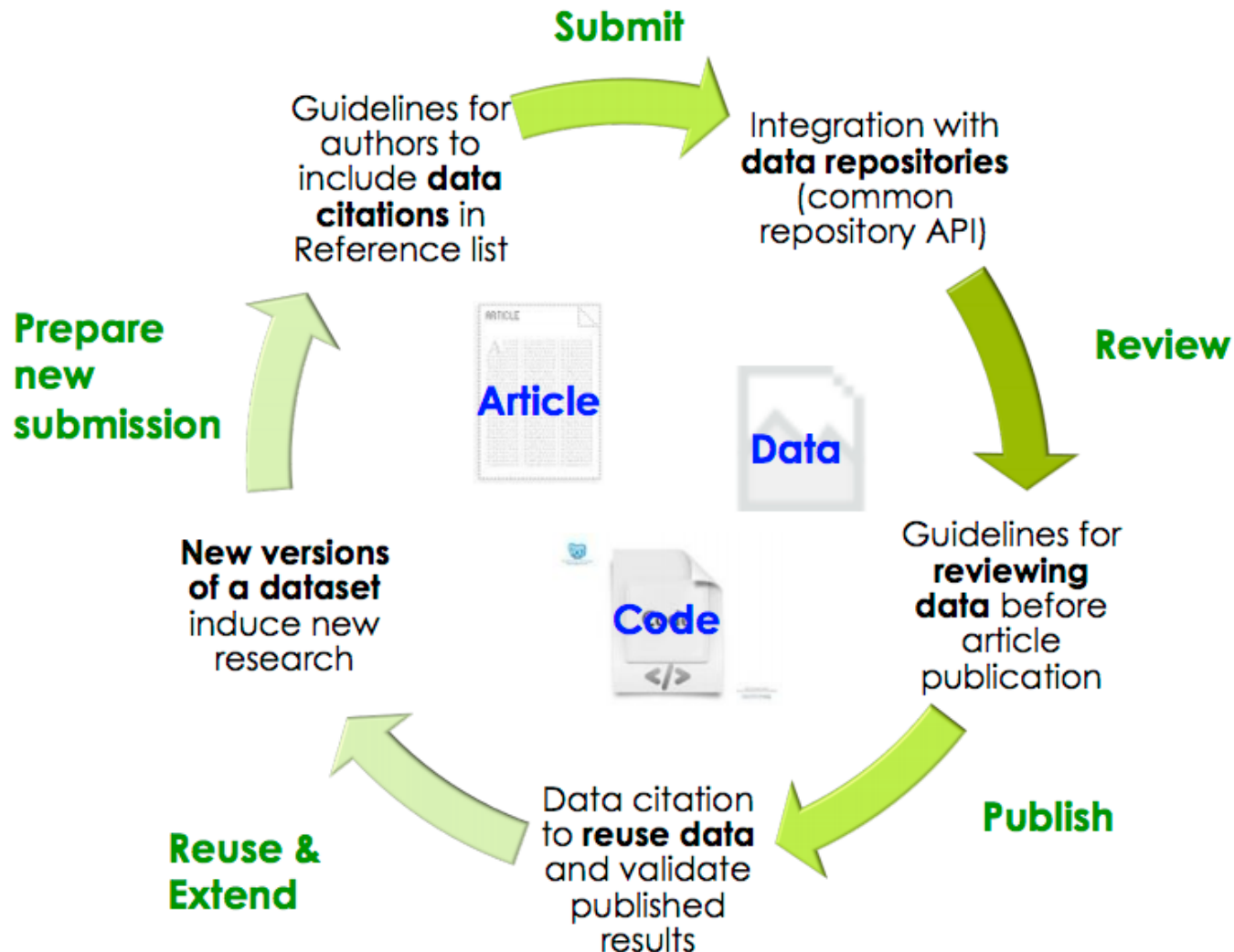


Dataverse Best Practices (2)

- Fixity:
 - UNF (King & Altman) for tabular data
 - MD5 checksums for other files
- Open Data (+ metadata) Licenses (CC0) *or* custom
- **OAI-PMH**: harvesting metadata (DC, DDI,...)
- LOCKSS (replication of files) → Data-PASS



Towards An Integrated Publishing Workflow



API Integration with Dataverse

Data Deposit API (metadata + data w/ SWORDv2)

For depositing datasets into Dataverse via API
See: OJS-Dataverse Journal Integration Project

<http://projects.iq.harvard.edu/ojs-dvn/home>



PKP
PUBLIC
KNOWLEDGE
PROJECT

Also: dvn R Package, **OSF** Dataverse Add-on, etc



Data Sharing API

For searching/downloading Dataverse datasets
(metadata + data) via API.

See: Thomas Leeper's dvn R package

Future of Dataverse?

- Dataverse 4.0 (try [beta](#))
 - Based on usability testing
- WorldMap Integration (geospatial viz w/ GeoConnect)
- Sharing Privacy Sensitive Data
 - Secure Dataverse
 - [DataTags](#) (questionnaires based on privacy laws)

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.

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
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				

Logos: NSF, University of Michigan, Berkman

Longer-Term

- Provenance Registry (data citation & provenance w/ SEAS (NSF))
- ORCID Integration (API)
- Large-scale datasets (efficient storage) → iRods w/ ODUM
- Ensuring long-term preservation for more file formats (e.g., Archivematica)
- Integrate with more Publishing Systems

Thank you!

Contact: ecastro@fas.harvard.edu

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

Twitter: @thedataorg

