



# Sharing With Dataverse

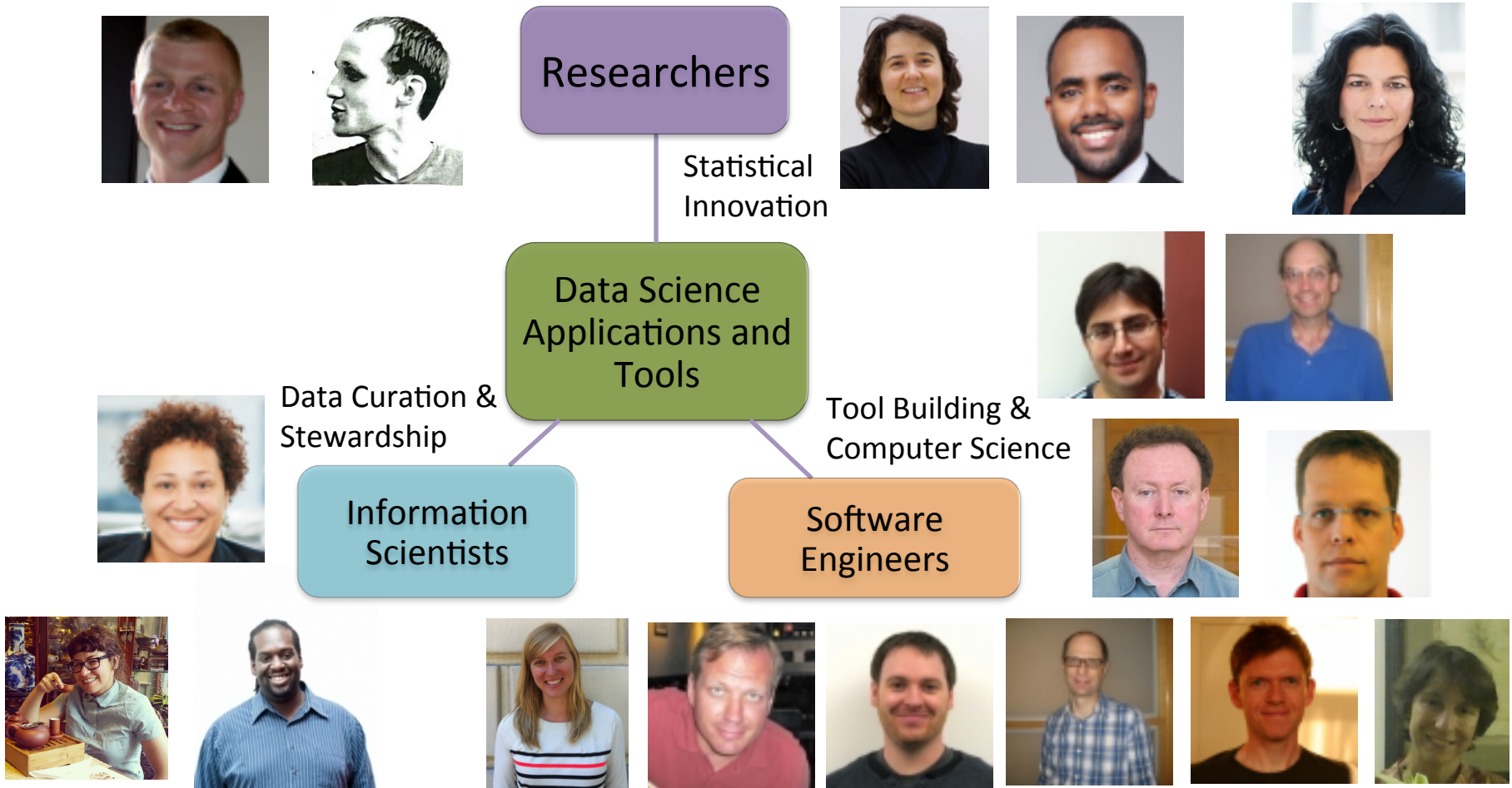
Eleni Castro  
IQSS Harvard  
NDS2 October 24, 2014

The  
**Dataverse**  
Project 



The Institute for Quantitative Social Science

# Data Science Team



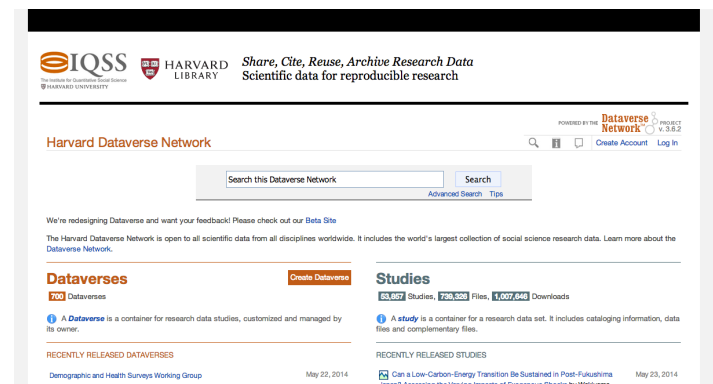
Find out more: <http://datascience.iq.harvard.edu>

# Introduction to Dataverse

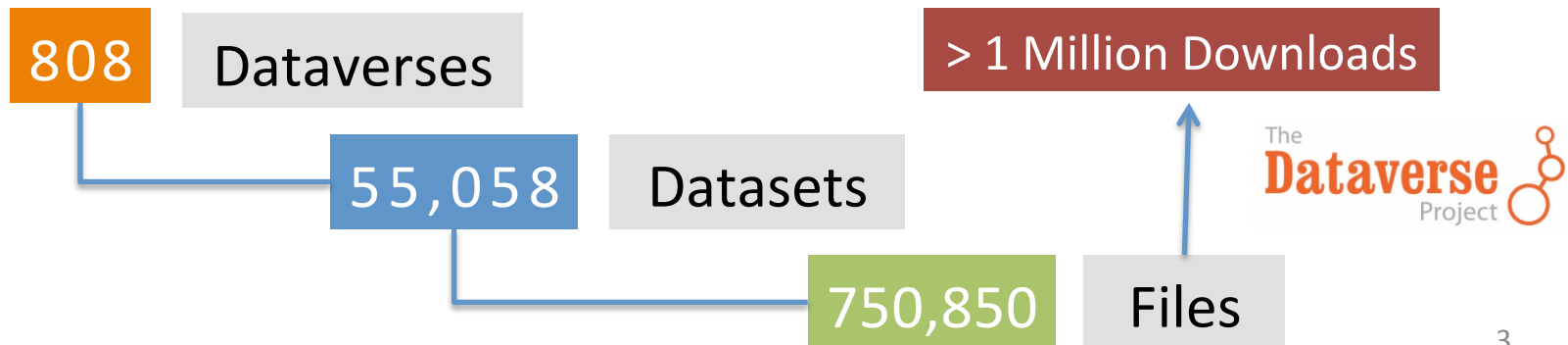
Software framework for publishing, citing and preserving research data (open source on [github](https://github.com) for others to install)

**Provides incentives for researchers to share:**

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

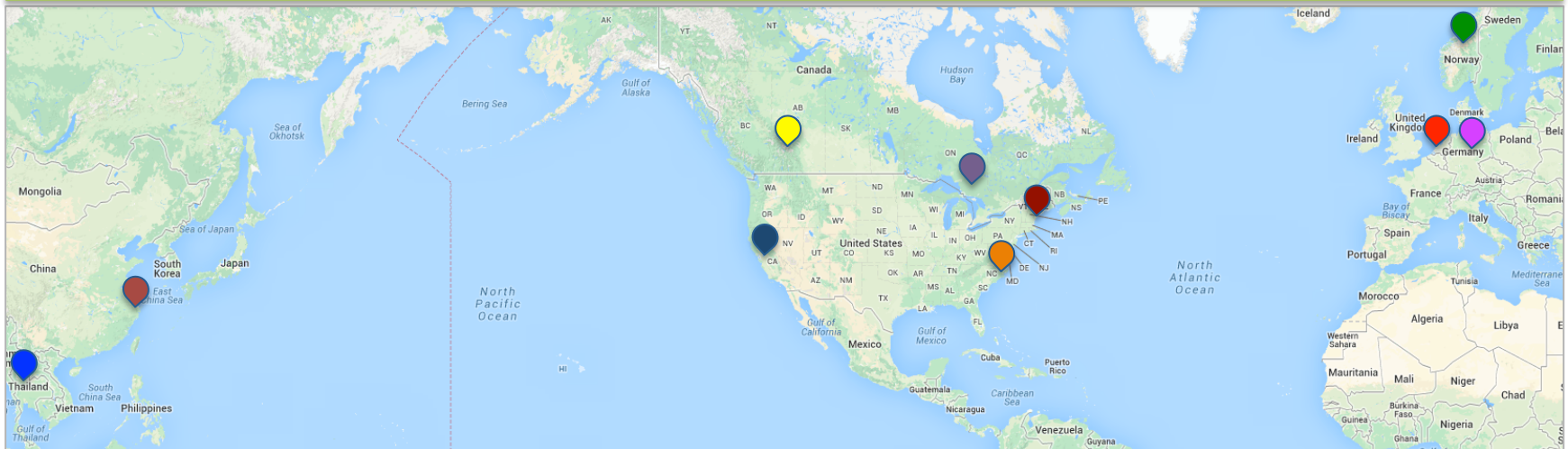


**Harvard Dataverse** (open to all; general community repository instance):



# 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.

# 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)
- Formal Data Citation (Altman & King, 2007)
  - Endorsed + comply w/ Joint Declaration of Data Citation Principles (FORCE11)
- Persistent IDs: Handles & DOI (DataCite/EZID)

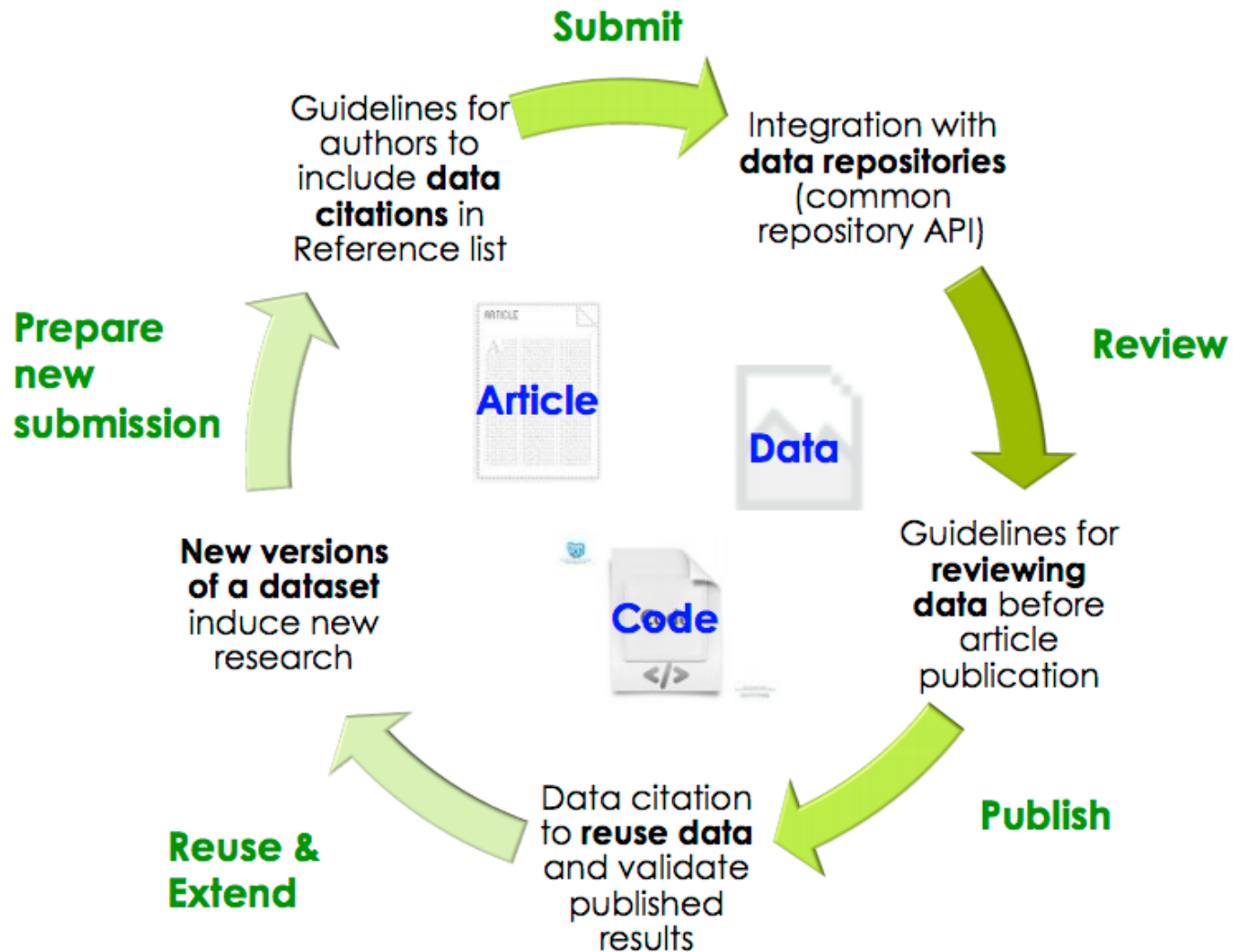


# Dataverse Best Practices (2)

- File Fixity:
  - UNF for tabular data
  - MD5 checksums for other files
- Open Data (+ metadata) Licenses (CC0)
- **OAI-PMH**: harvesting metadata (DC, ...)
- LOCKSS (replication of files)
  - Data-PASS: (ICPSR, ODUM, NARA, ROPER,...)



# 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>



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



# Visualize & Analyze Data: TwoRavens

- With Dataverse 4.0: for Tabular Data (i.e., R, csv, SPSS, ...)
- 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

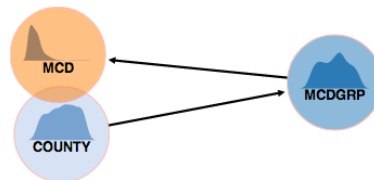
## Data Selection

Variables Subset

MCDGRP

COUNTY

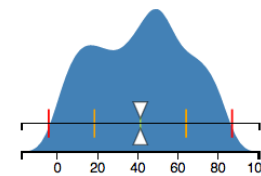
MCD



## Model Selection

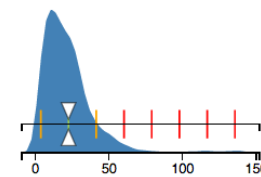
Models Set Covar. Results

MCDGRP



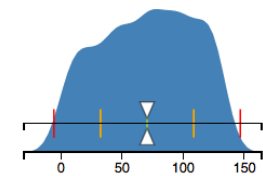
x: 41.35  
x1: 41.35

MCD



x: 22.54  
x1: 22.54

COUNTY



x: 70.55  
x1: 70.55

# WorldMap Integration

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

**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: 8bc85dc41c2341fa4b2267f62b1d07

[View on WorldMap](#)

Notes -

Ability to "Re-Map" It

Re-Map It Download

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



CENTER FOR GEOGRAPHIC ANALYSIS  
HARVARD UNIVERSITY



Boston  
Area  
Research  
Initiative

# Future Dataverse Collaborations

## 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           |
|--|----------------------|----------------|-------------------|-------------------|
| <b>Blue</b>  | None                 | None           | Clear             | Clear             |
| <i>Non-confidential information that can be stored and shared freely</i>   |                      |                |                   |                   |
| <b>Green</b>   | None                 | Email or OAuth | Clear             | Clear             |
| <i>Potentially identifiable but not harmful personal information, shared with some access control</i>                                  |                      |                |                   |                   |
| <b>Yellow</b>  | Click Through        | Password       | Encrypted         | Clear             |
| <i>Potentially harmful personal information, shared with loosely verified and/or approved recipients</i>                               |                      |                |                   |                   |
| <b>Orange</b>  | Sign                 | Password       | Encrypted         | Encrypted         |
| <i>May include sensitive, identifiable personal information, shared with verified and/or approved recipients under agreement</i>       |                      |                |                   |                   |
| <b>Red</b>   | Sign                 | Two Factor     | Encrypted         | Encrypted         |
| <i>Very sensitive identifiable personal information, shared with strong verification of approved recipients under signed agreement</i> |                      |                |                   |                   |
| <b>Crimson</b>   | 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

- Data Citation Provenance Registry (NSF funder w/ SEAS)
- Large-scale datasets (efficient storage) → iRods
- Integrate w/ more Publishing Systems (new Sloan grant!)
- Author Disambiguation: ORCID Integration (API)
- Long-term preservation for more file formats (Archivematica)

Thank you!

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

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

Twitter: @thedataorg

