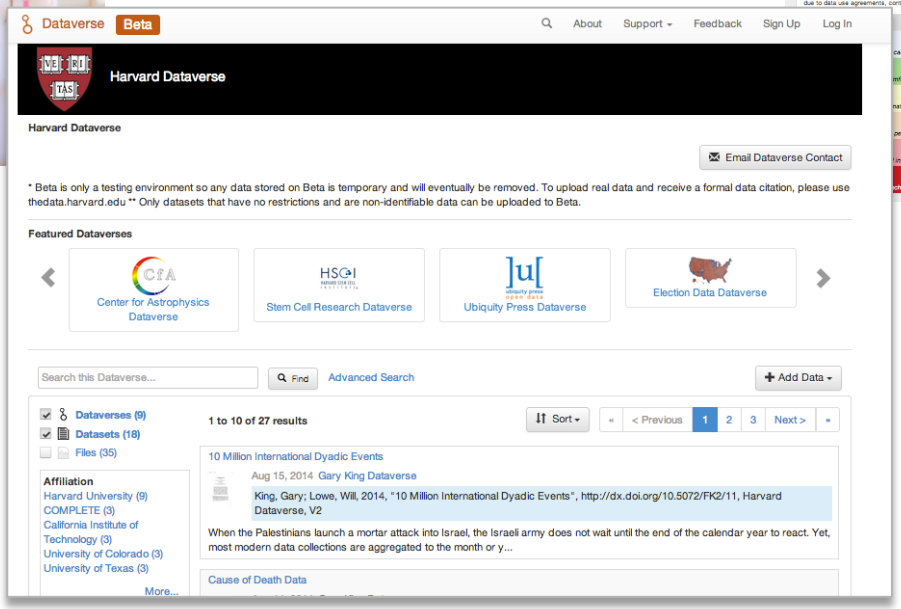
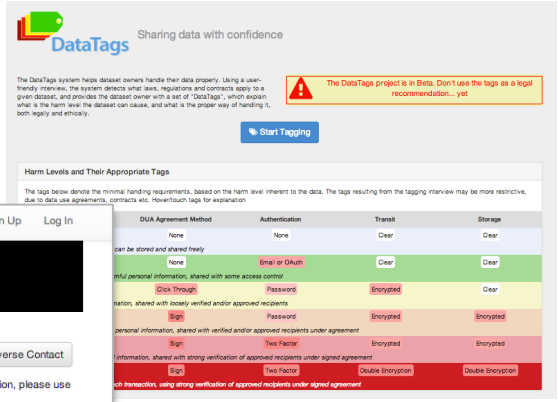
Dataverse 4.0 & Beyond

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

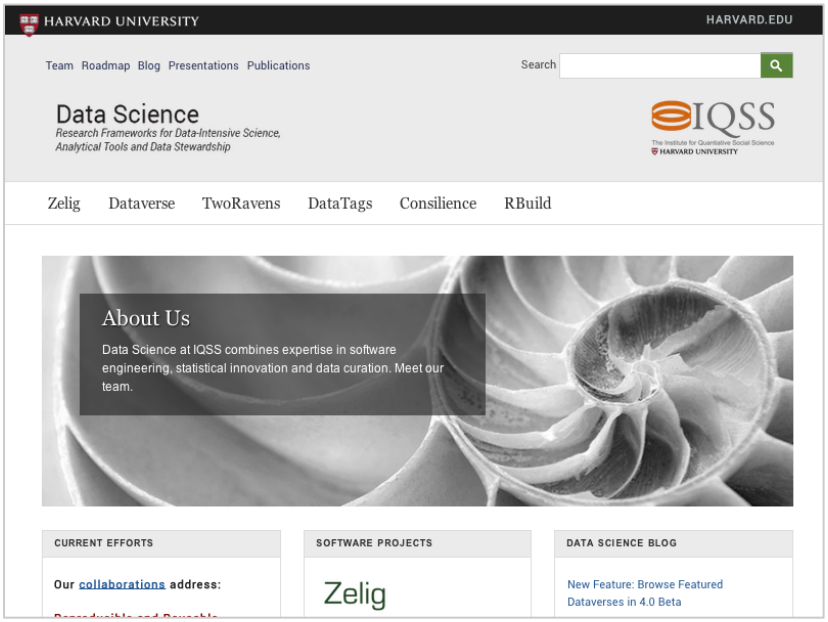


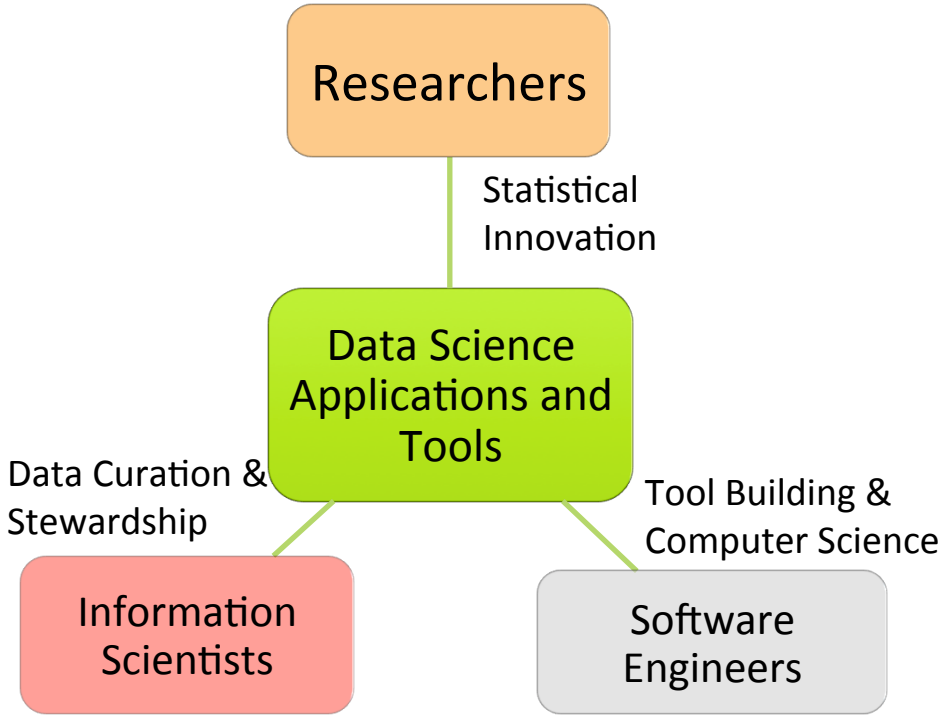


IQSS
The Institute for Quantitative Social Science

The Dataverse Project

Data Science Team





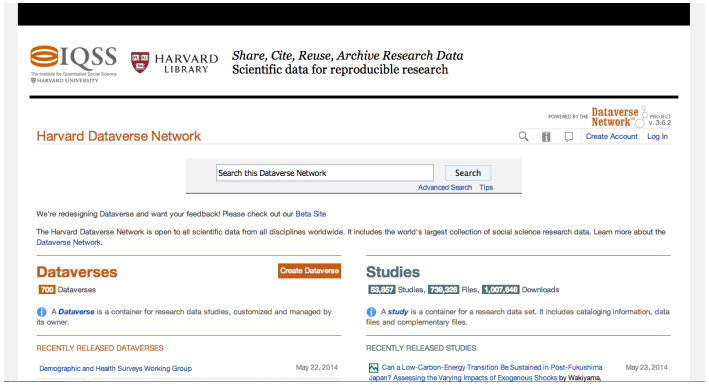
Find out more: **http://datascience.iq.harvard.edu**

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:

700 Dataverses

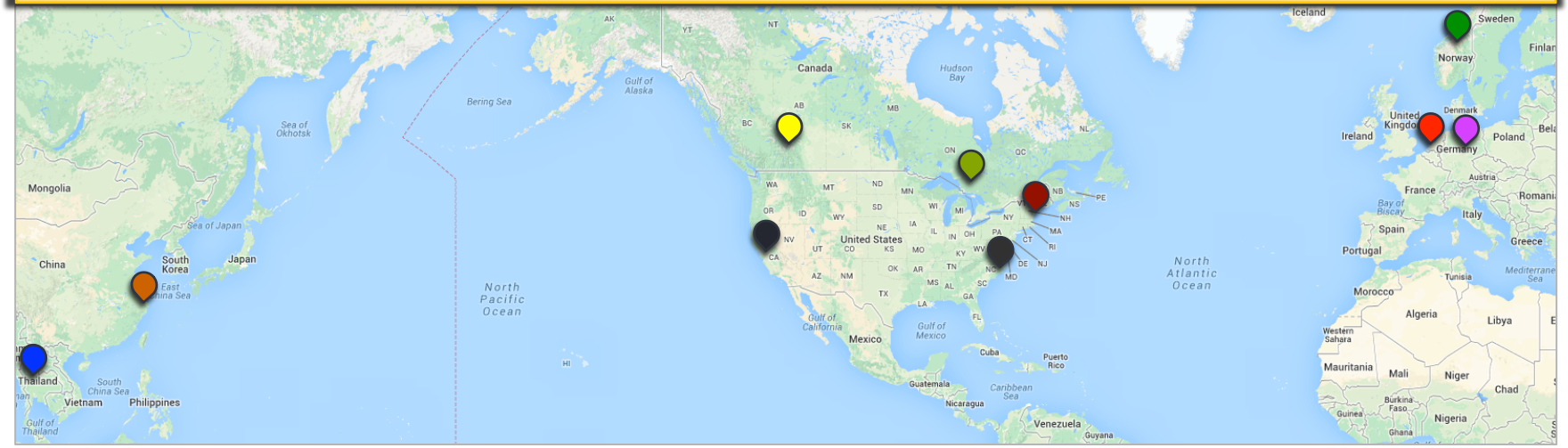
53,857 Datasets

739,326 Files

> 1 Million Downloads

Who is using Dataverse?

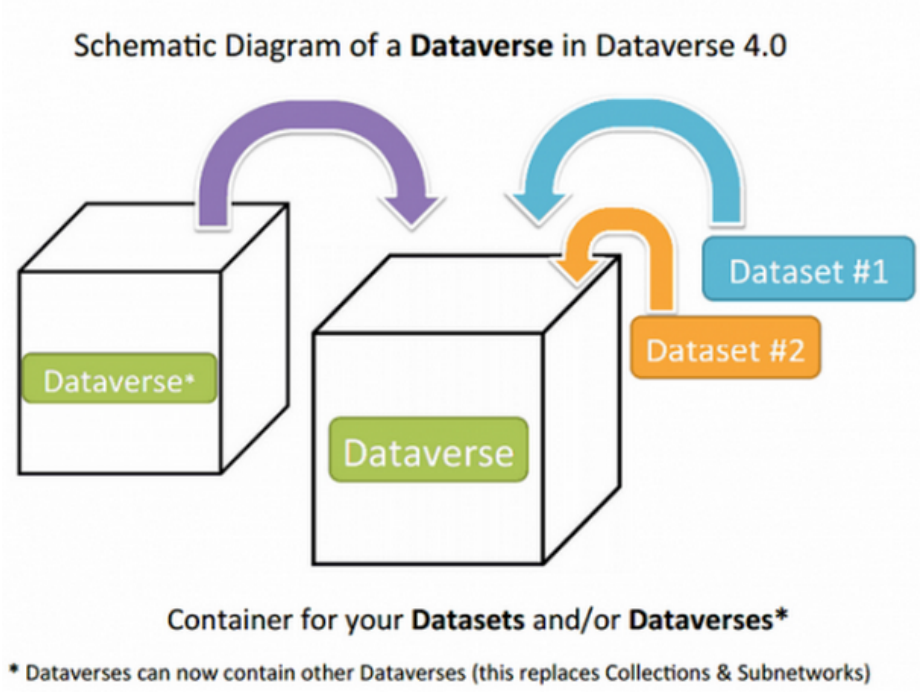
Worldwide Dataverse Installations

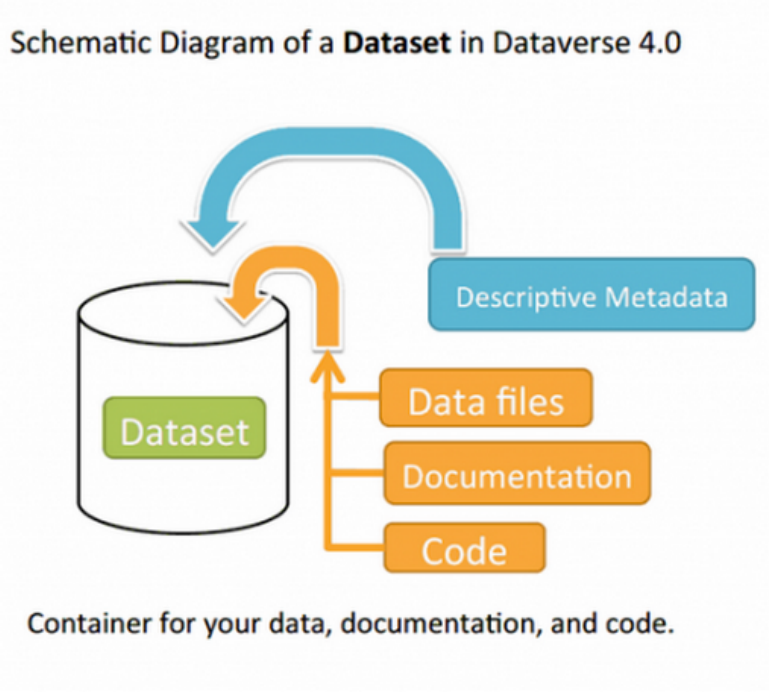


Institutions can setup/host their own Dataverse installation (OCUL, UoA, etc) and within them can have dataverses for a variety of users (across all research domains): Researchers, Projects, Journals (OJS – Dataverse integration), etc.

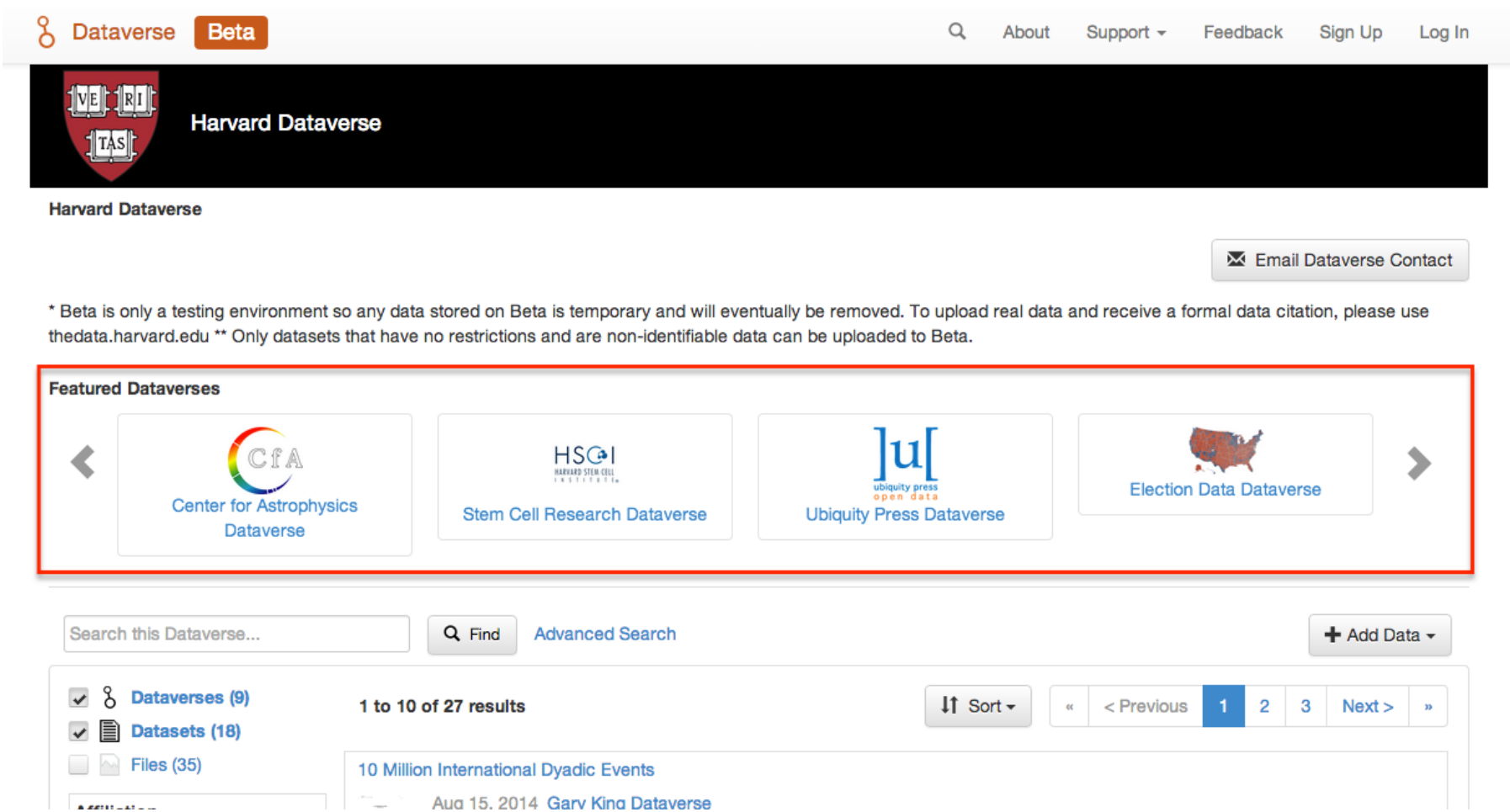
Streamlined Workflows

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



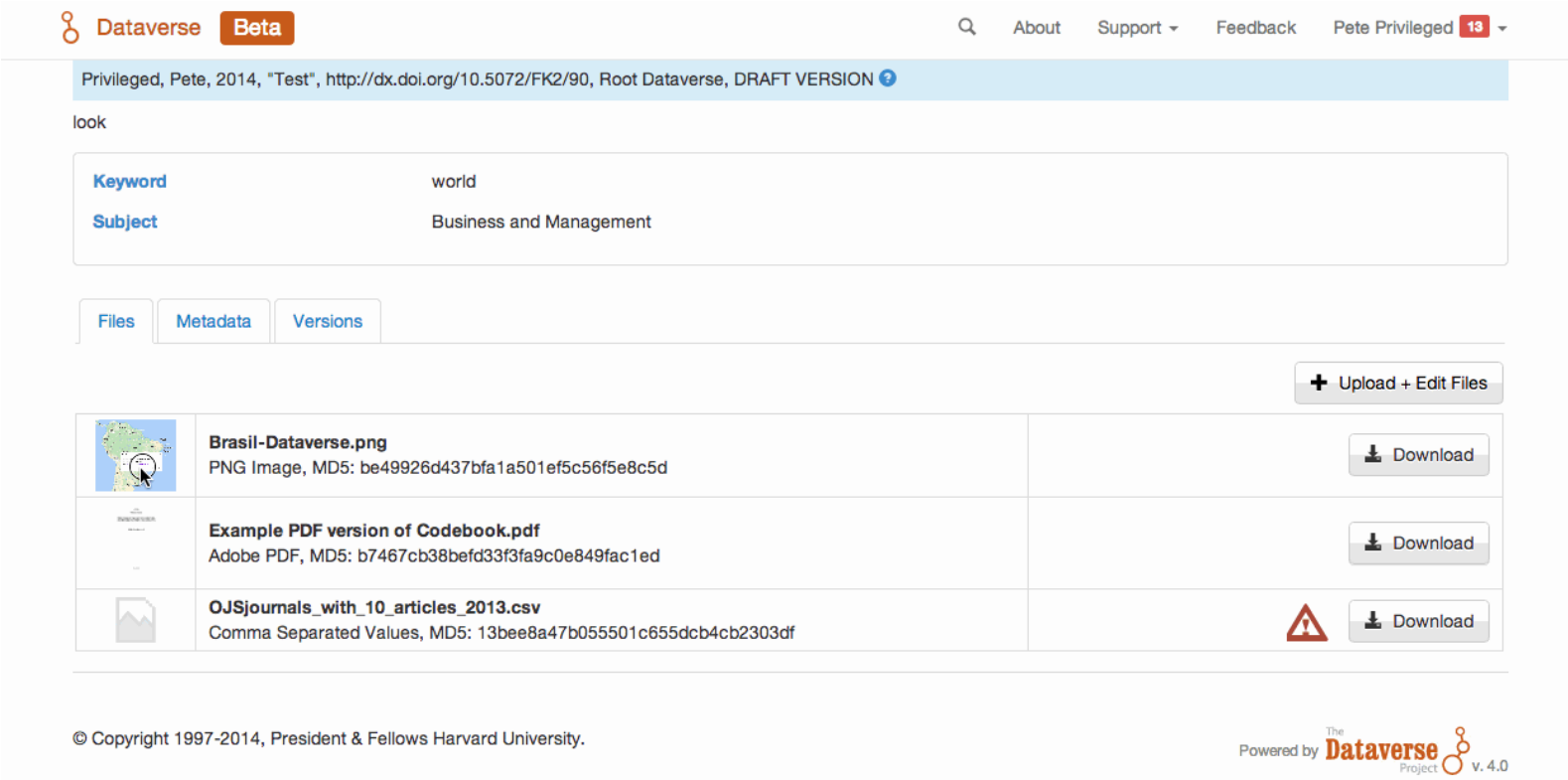


Featured Dataverses

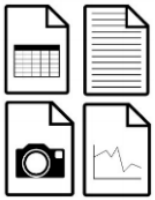


Improved File Upload & Handling

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



Rigorous Data Publishing Workflows



Upload

Draft Dataset

Note: A Published Dataset **cannot** be deleted (only deaccessioned, with reason included (i.e., legal)).

Publish Version 1

Authors, Title, Year, DOI, Repository, V1

Published Dataset v1

Publish Version 1.1: small metadata change (not citation); files not changed.

Published Dataset v1.1

Publish Version 2: File change (automatic); big metadata change (citation metadata).

Authors, Title, Year, DOI, Repository, UNF, V2

Published Dataset v2

See: Altman, M., & King, G. (2007) doi:10.1045/march2007-altman

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,…).

Astronomy and Astrophysics Metadata

Type

- Image
- Mosaic
- EventList
- Spectrum
- Cube

Facility

Instrument

Spatial Resolution

Spectral Resolution

Time Resolution

Bandpass

Central Wavelength (m)

Wavelength Range — **Minimum (m)** — **Maximum (m)**

Dataset Date Range — **Start** (YYYY-MM-DD) — **End** (YYYY-MM-DD)

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

CfA

SEAMLESS ASTRONOMY
Linking scientific data, publications, and communities

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

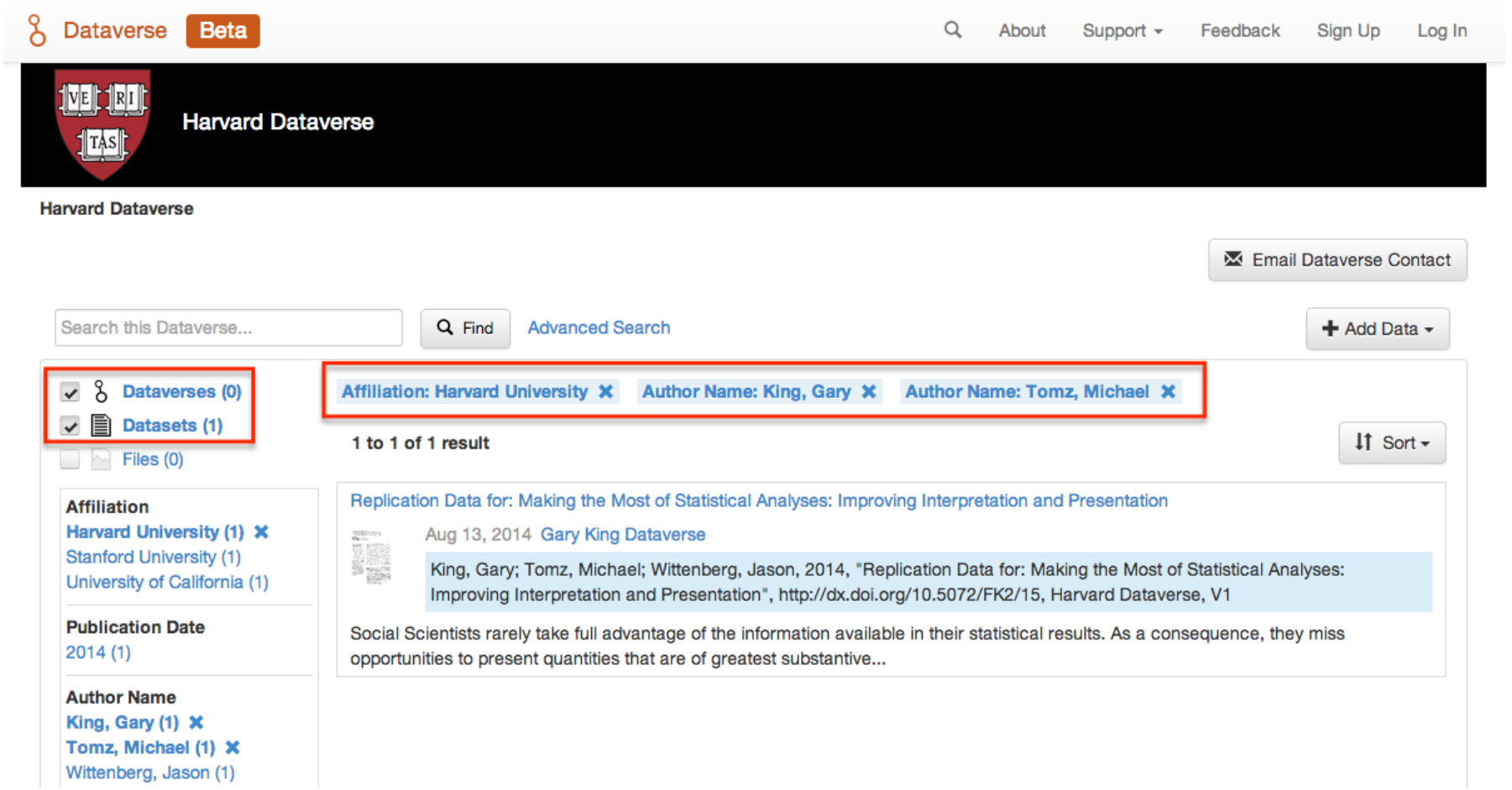
- Danio rerio
- Homo sapiens
- Mus musculus
- Rattus norvegicus

Cell Type

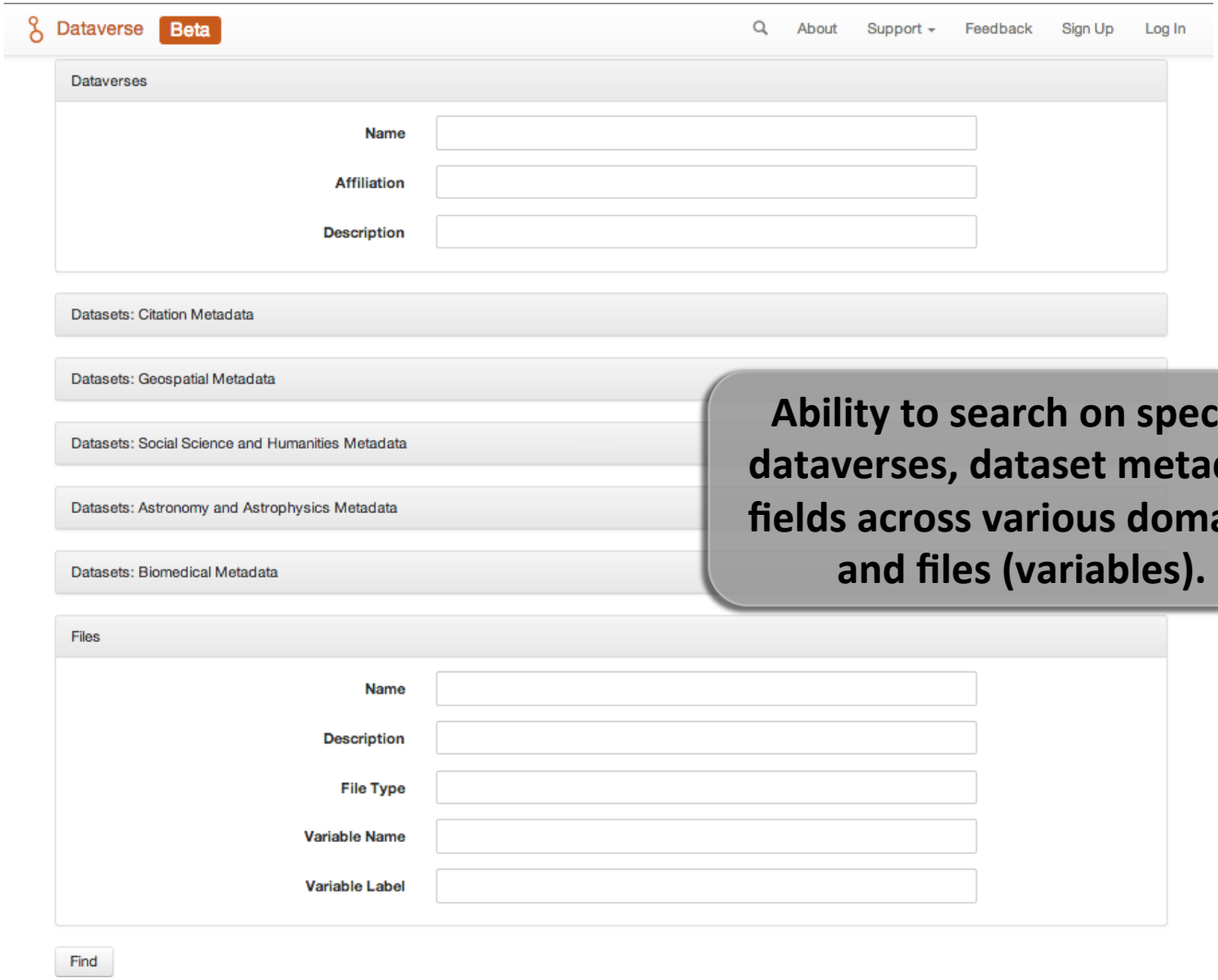
+

eagle-i network HSCI HARVARD STEM CELL INSTITUTE isatab

Enhanced Faceted Search



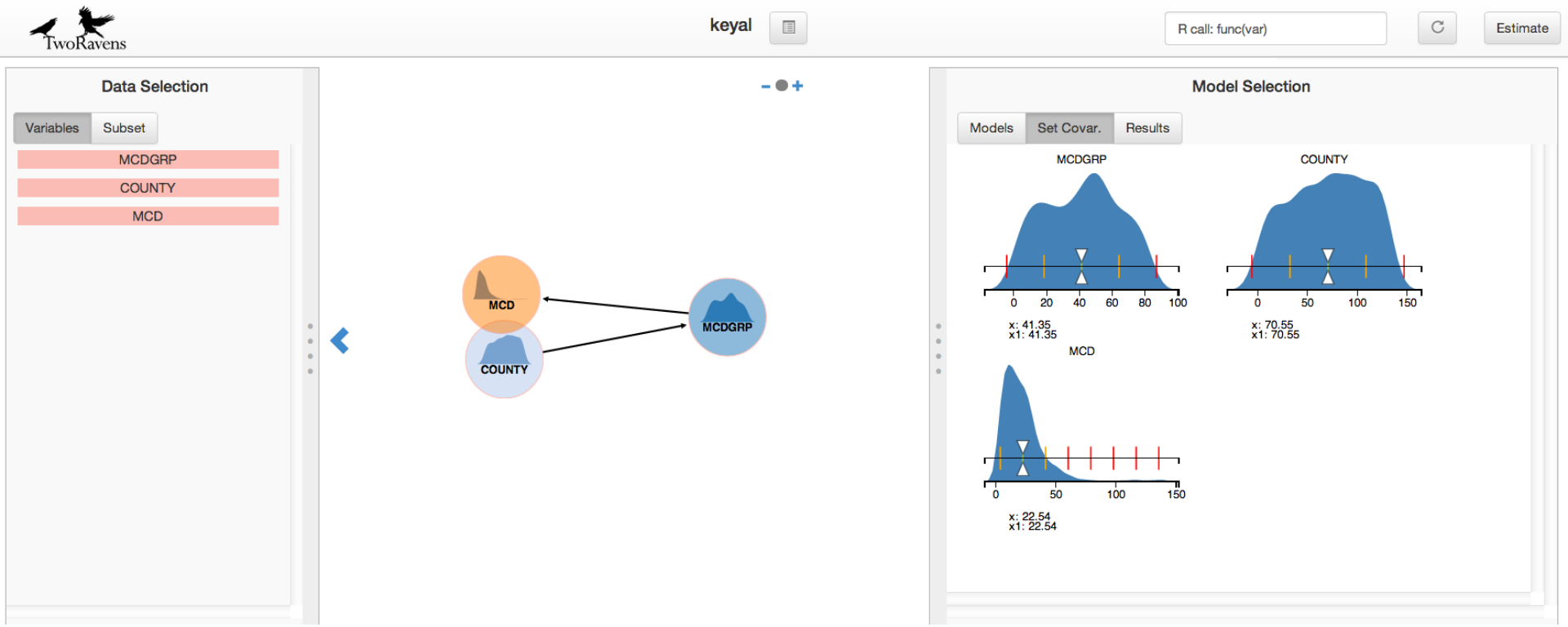
Expanded Advanced Search



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

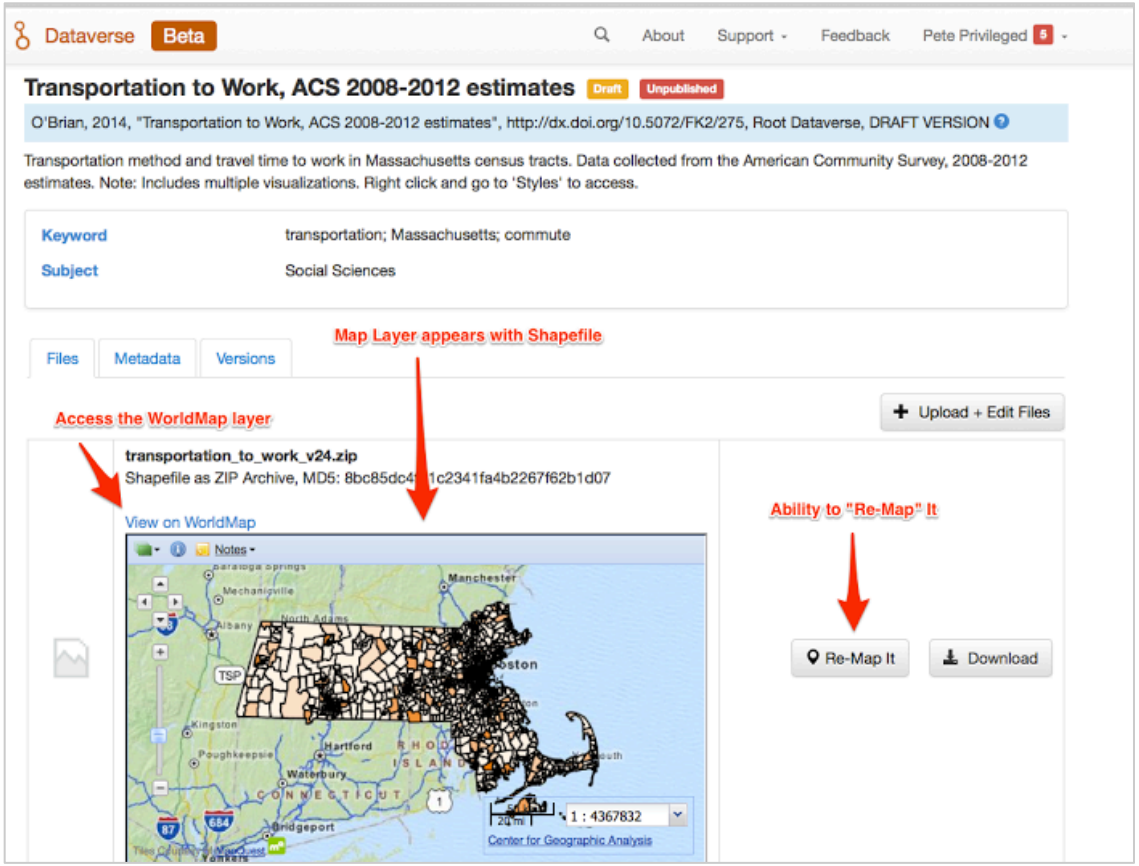
Visualize & Analyze Data: TwoRavens

- 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



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



Boston Area Research Initiative

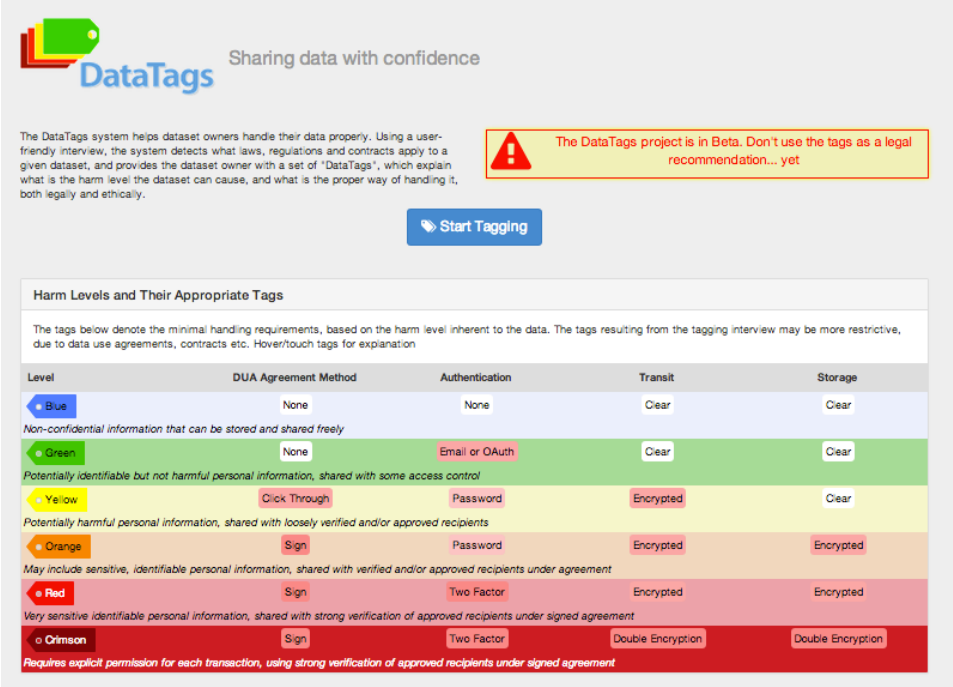
CENTER FOR GEOGRAPHIC ANALYSIS
HARVARD UNIVERSITY

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



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.

Credit: FlickrCommons

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