The Expanding Dataverse

Mercè Crosas, Director of Data Science, IQSS

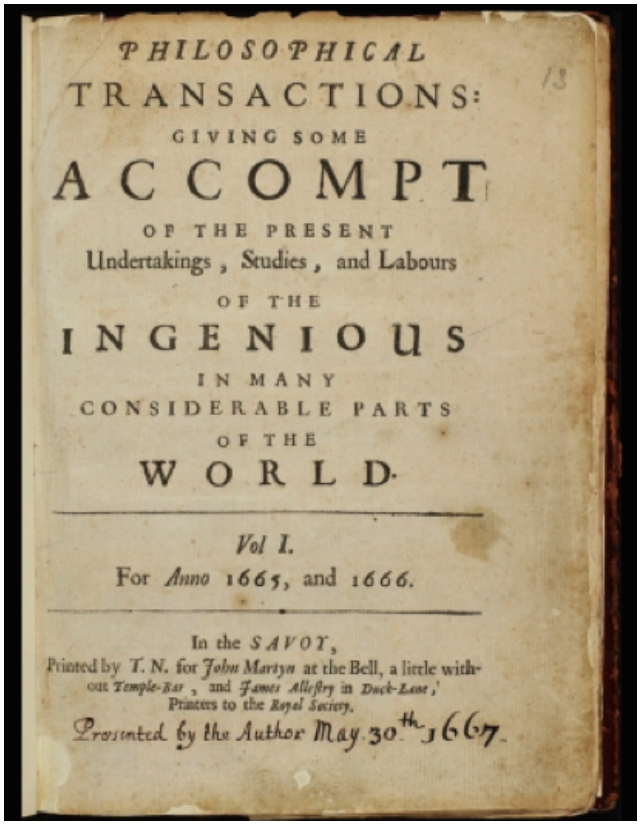
@mercecrosas

January 21, 2015, Lamont Library, Harvard University

Data Publishing: A form of Scholarly Communication

1665

Data, if any, were part of the printed publication



PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

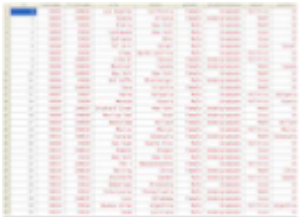
Vol I.
For Anno 1665, and 1666.

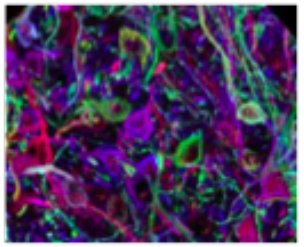
In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.
Presented by the Author May 30th 1667.

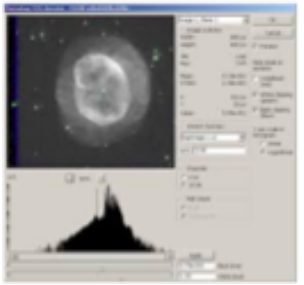
350 years of scientific publishing, with words and data

Now

Vast quantities of digital data (and code) cannot be part of the printed publication







Pillars of Data Publishing

To make data discoverable, accessible and reusable, we need:

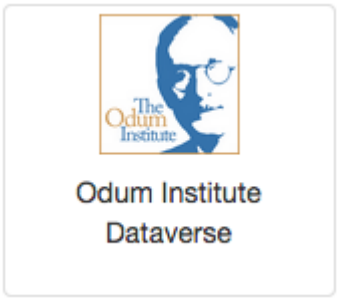
1. **Data Citation**, to reference and find data
2. **Data Repositories**, to host and access data
3. **Information about the data,** to understand and reuse them

Dataverse Software:
A Data Publishing framework

… for a wide range of repositories

Public, Generic Repositories

Institutional Repositories



Curated Data Archives Repositories

The Dataverse Project

Dedicated to sharing, archiving and citing research data.

Add Data

Find Data

Get Recognition

Dataverse Repositories

Be a part of the community by publishing your data to one of the Dataverse repositories or by setting up a new repository for your organization.

(Click cards for more information)

Harvard Dataverse

Odum Institute Dataverse

DANS - Dutch Dataverse

Fudan University Dataverse

University of Alberta Libraries Dataverse

Scholars Portal Dataverse

Abacus Dataverse

Heidelberg University HeiDATA Dataverse

UnBraL Fronteiras (Brazil) Dataverse

UiT Open Research Data

Dataverse 4.0: Enables and Enhances Data Publishing

- A **data citation** compliant with the Data Citation Principles
- Rich **metadata** to describe and find datasets from multiple domains
- Support for **public and restricted** data, open data license and terms of use
- Rigorous **workflows** to publish data, with support for new versions of the data

Data Citation

A Brief History of Citing Data

1906

Chicago Manual of Style:
author/creator, title, dates, publisher or distributor

1959

First scientific digital repositories
(e.g. World Data Center, ICPSR)

1979

ASBR ("Data File" type)
MARC (machine readable catalog)

Domain Repositories
(e.g., GenBank)

1999 - Now

Growth of Data Repositories
(e.g., NESSTAR, Dataverse, Dryad, Figshare, Zenodo)
DOI services for Data
(e.g., DataCite in 2009)

2014

Data Citation Principles
NISO-JATS revised to support data

Altman & Crosas, 2013, "The Evolution of Data Citation: From Principles to Implementation" IASSIST Quarterly

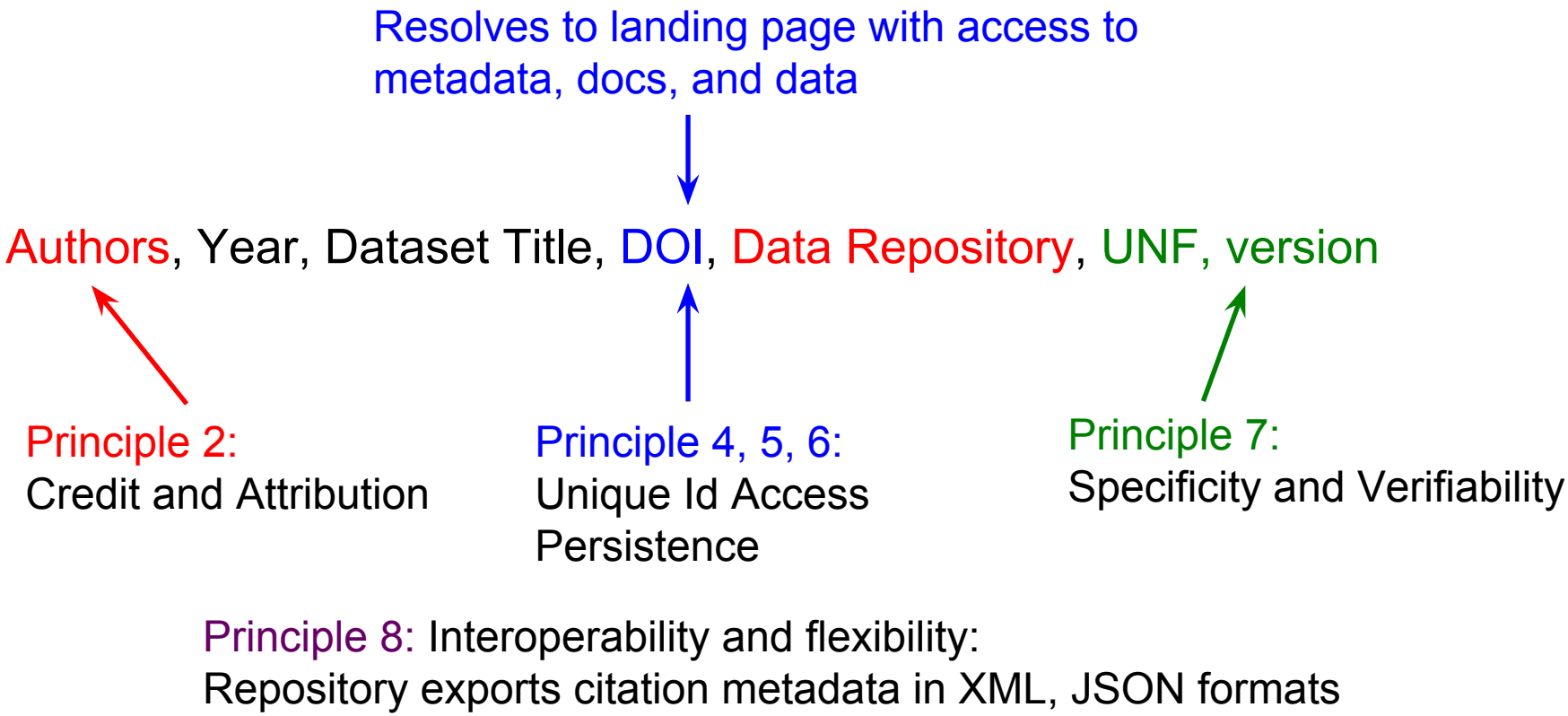
Joint Declaration of Data Citation Principles

1. **Importance**
2. **Credit and Attribution**
3. **Evidence**
4. **Unique Identification**
5. **Access**
6. **Persistence**
7. **Specificity and Verifiability**
8. **Interoperability and flexibility**

DC1
Data Citation Principles

https://www.force11.org/datacitation

Data Citation generated by Dataverse

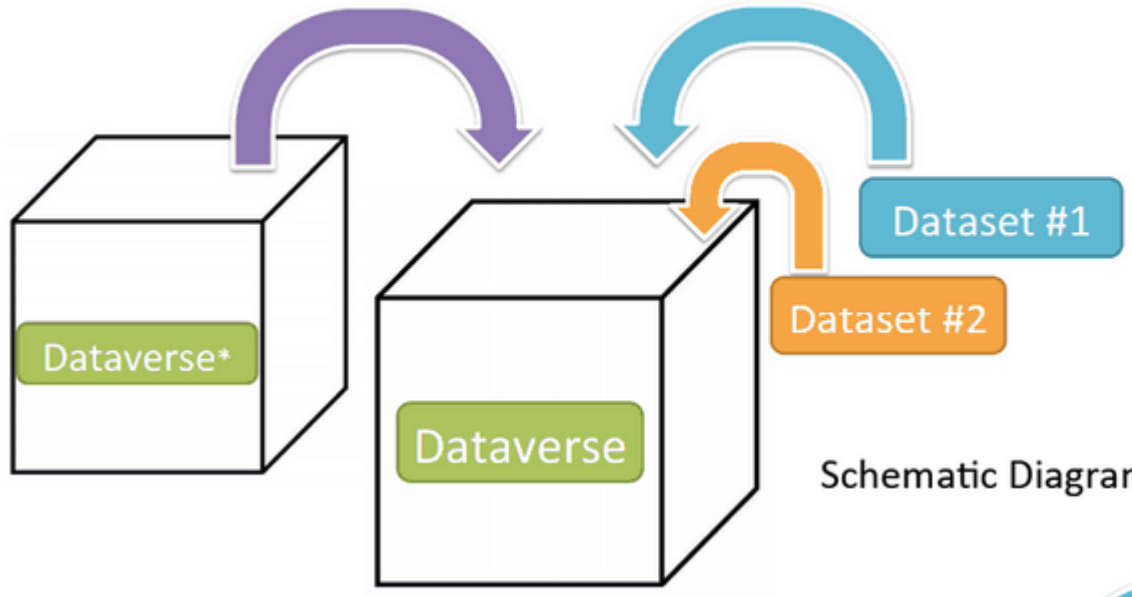


Principle 8: Interoperability and flexibility:
Repository exports citation metadata in XML, JSON formats

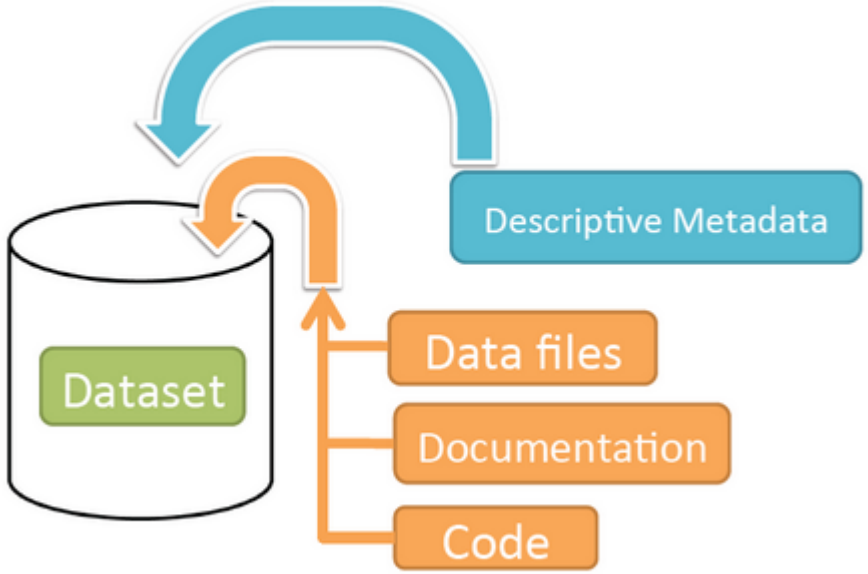
Altman & King, 2007. A Proposed Standard for the Scholarly Citation of Quantitative Data.

Metadata

Schematic Diagram of a **Dataverse** in Dataverse 4.0



Schematic Diagram of a **Dataset** in Dataverse 4.0



Container for your data, documentation, and code.

Three Metadata Levels

Generic Metadata

Includes data citation metadata fields (Examples: title, authors, persistent id, description)

Domain Specific Metadata

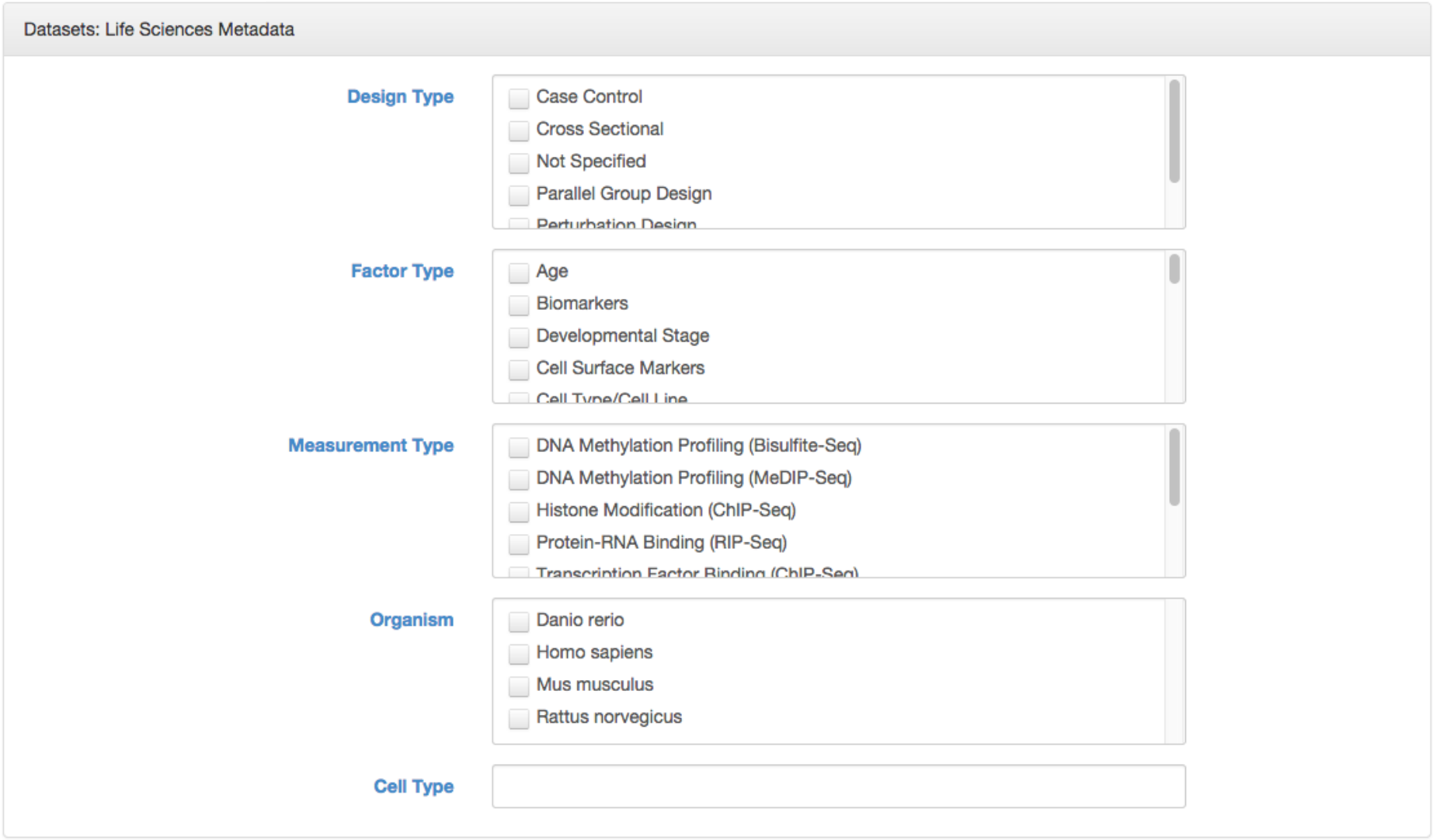
Examples:

- Social Science Metadata (DDI)
- Life Sciences (ISA-Tab)
- Astronomy (VO)

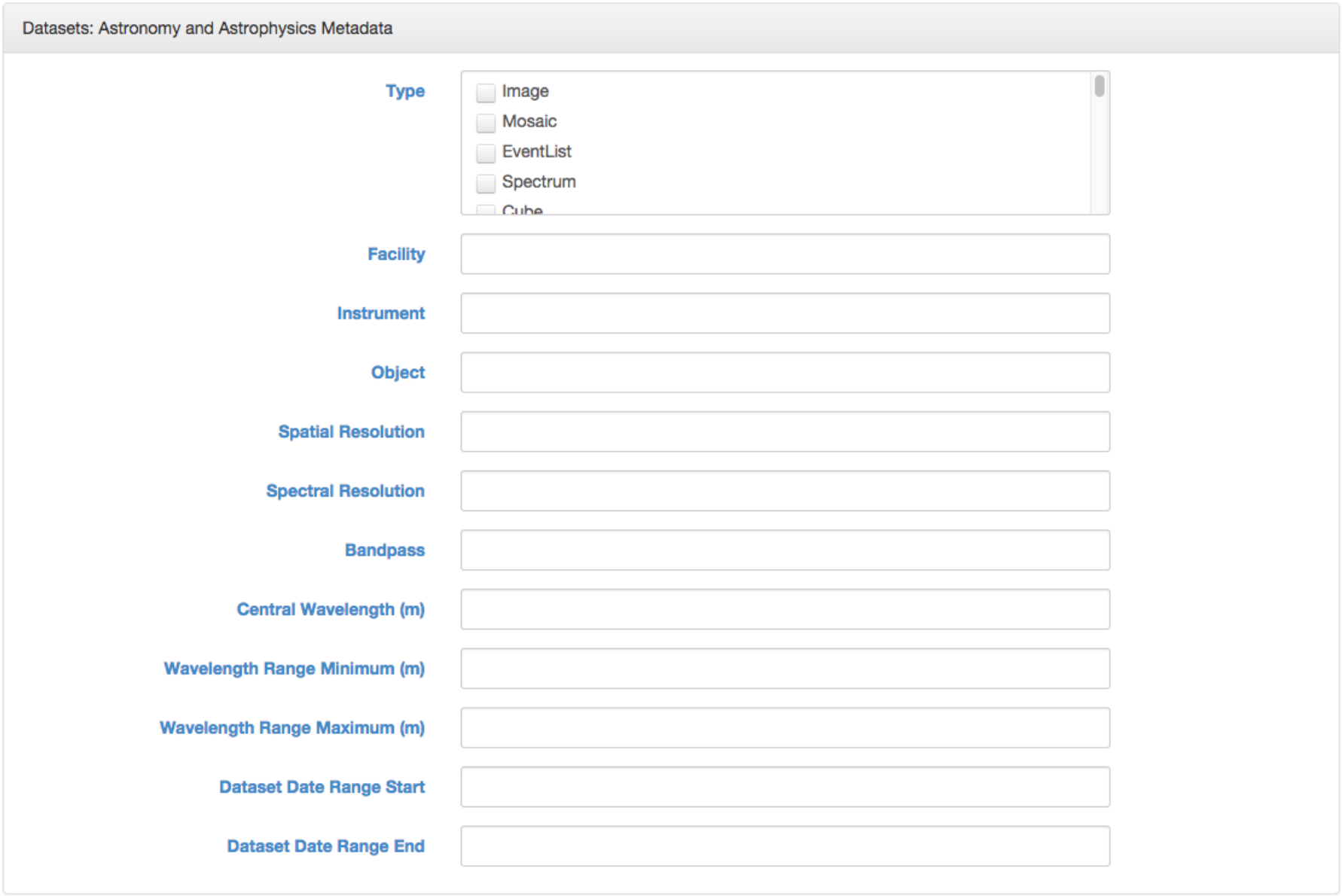
File Metadata

Examples (automatic):

- For Tabular Files: Column information
- For FITS Files: Header information



Example: Life Sciences Metadata



Example: Astronomy Metadata

Public vs Restricted

Terms, Licenses and Restrictions

Public Dataset

- CC0 License
- Metadata is public
- Files are public

Dataset with Restricted Files

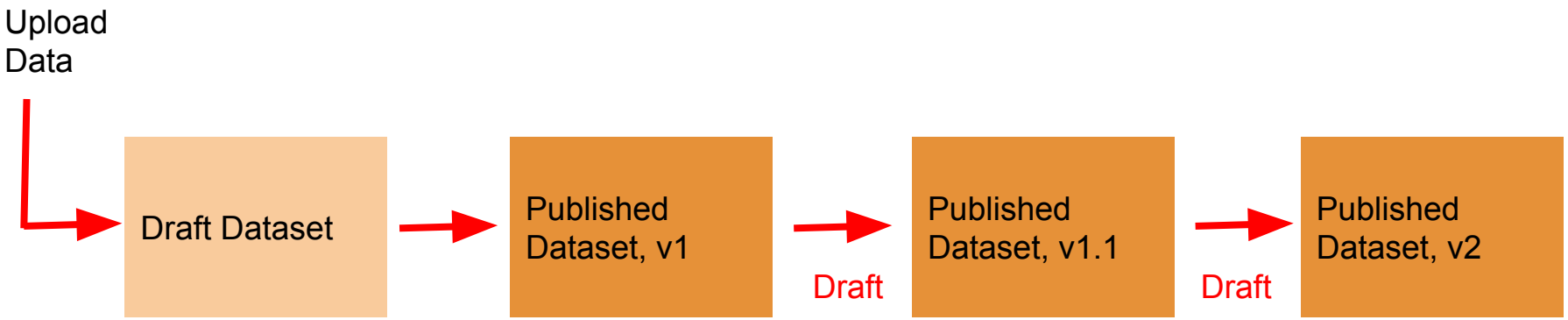
- CC0 License
- Metadata is public
- Files are restricted
- Access Terms are defined in dataset

Dataset with Terms of Use

- Metadata is public
- Terms of Use are defined in dataset (CC0 can’t apply)
- Files might be public or restricted

Workflows

Draft, Published and Versions



Dataset in review, can be shared with collaborators

Once published, dataset cannot be unpublished (only deaccessioned)

Data Citation becomes public

Minor version for small changes to dataset description

Data Citation doesn't change

Major version for new versions of data files

Data Citation changes

Multiple Roles for Multiple Workflows

Editor

Upload Data +
Edit Metadata

Manager

Upload Data +
Edit Metadata

+

Set File Restrictions +
License and Terms

Curator

Upload Data +
Edit Metadata

+

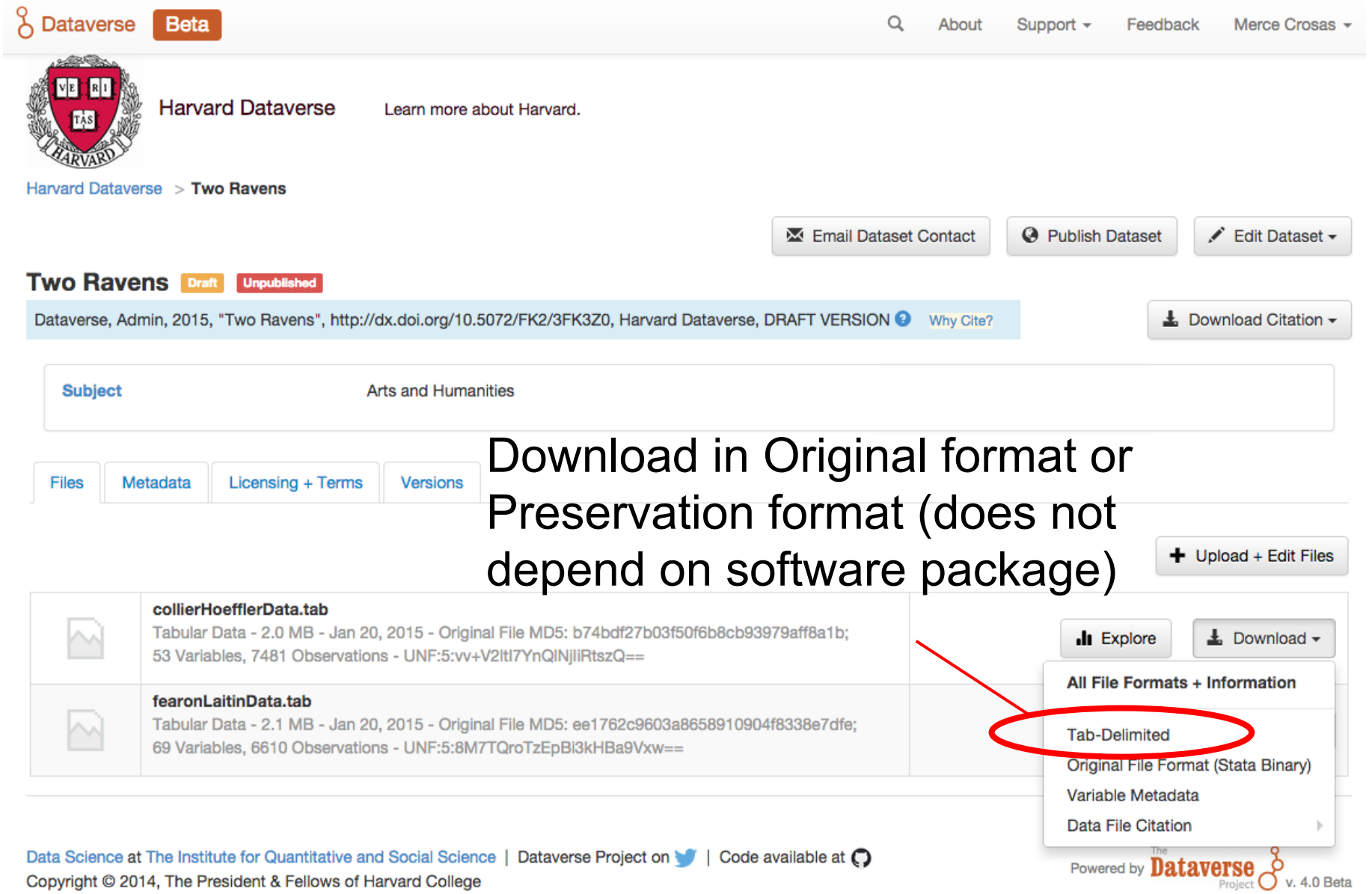
Set File Restrictions +
License and Terms

+

Grant Access +
Publish Dataset

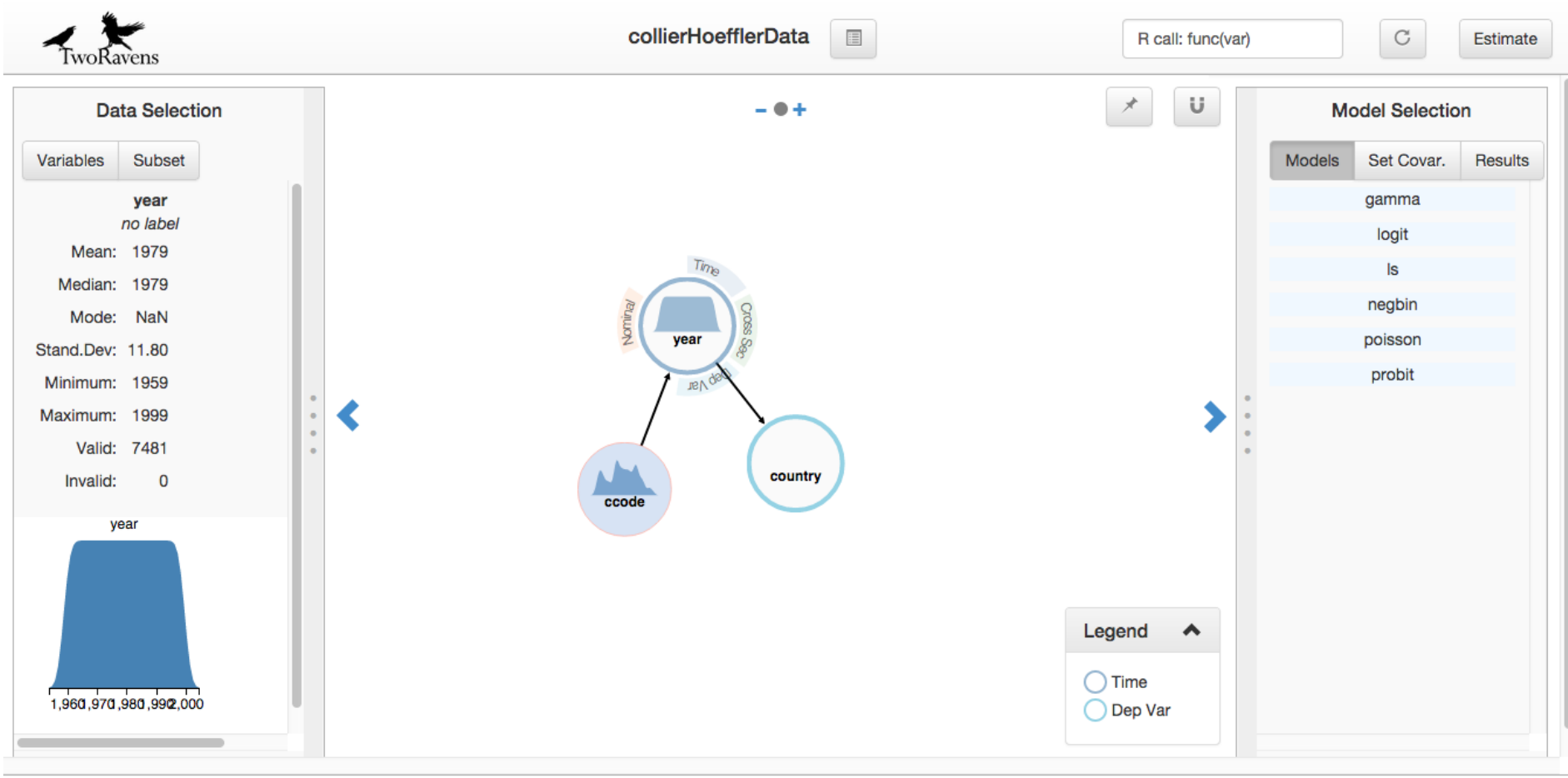
\+ Custom Roles

Data Processing, Analysis, and Visualizations

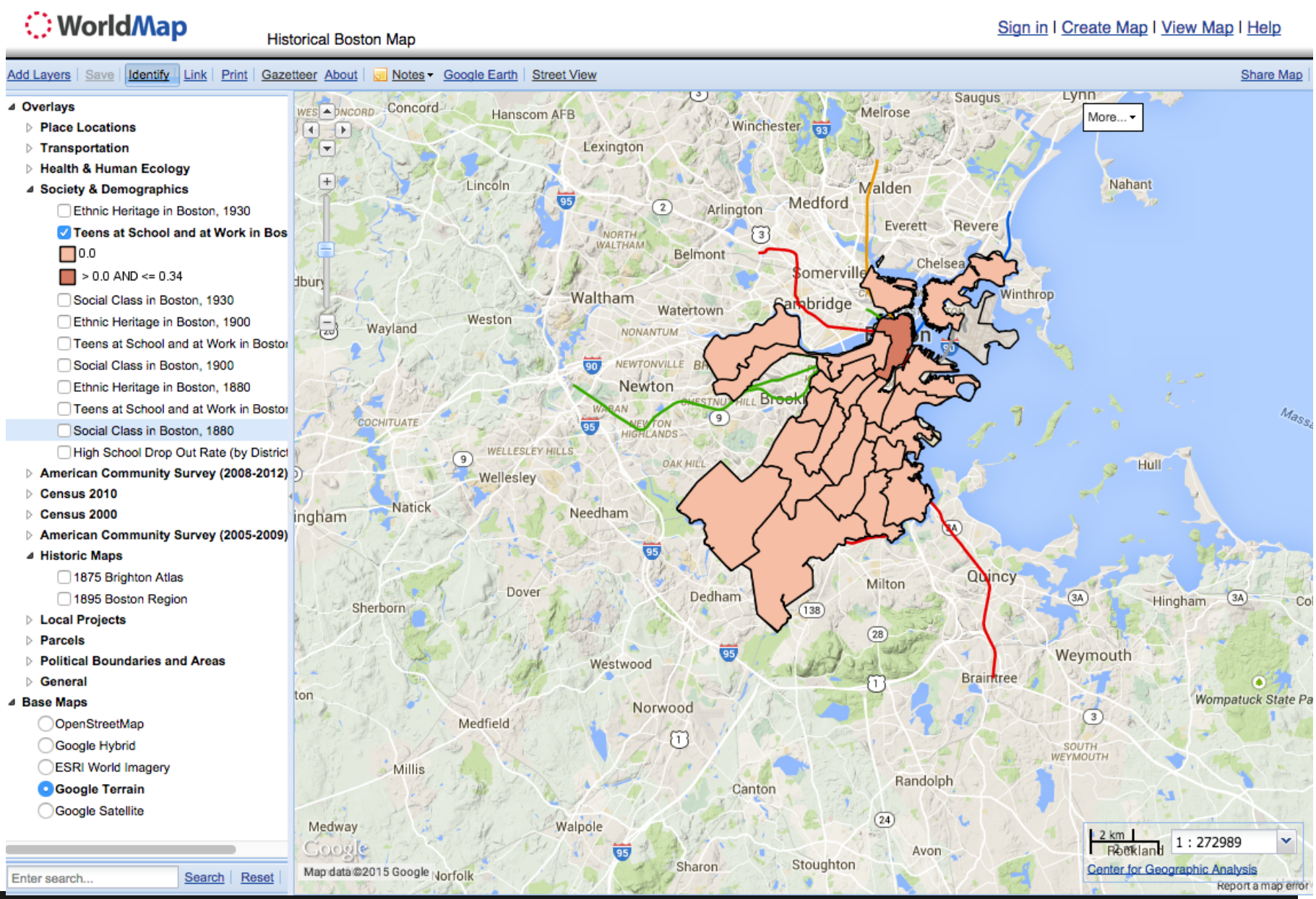


Download in Original format or Preservation format (does not depend on software package)

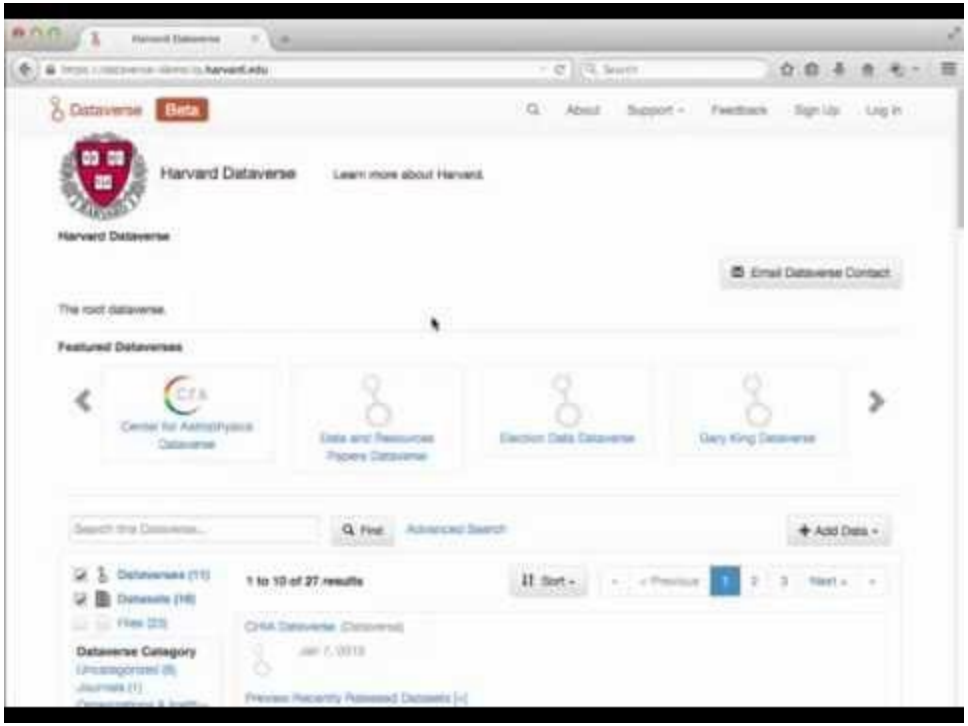
Tabular Data: Converted to Preservation format



Tabular Data: Explore and Analyze with TwoRavens



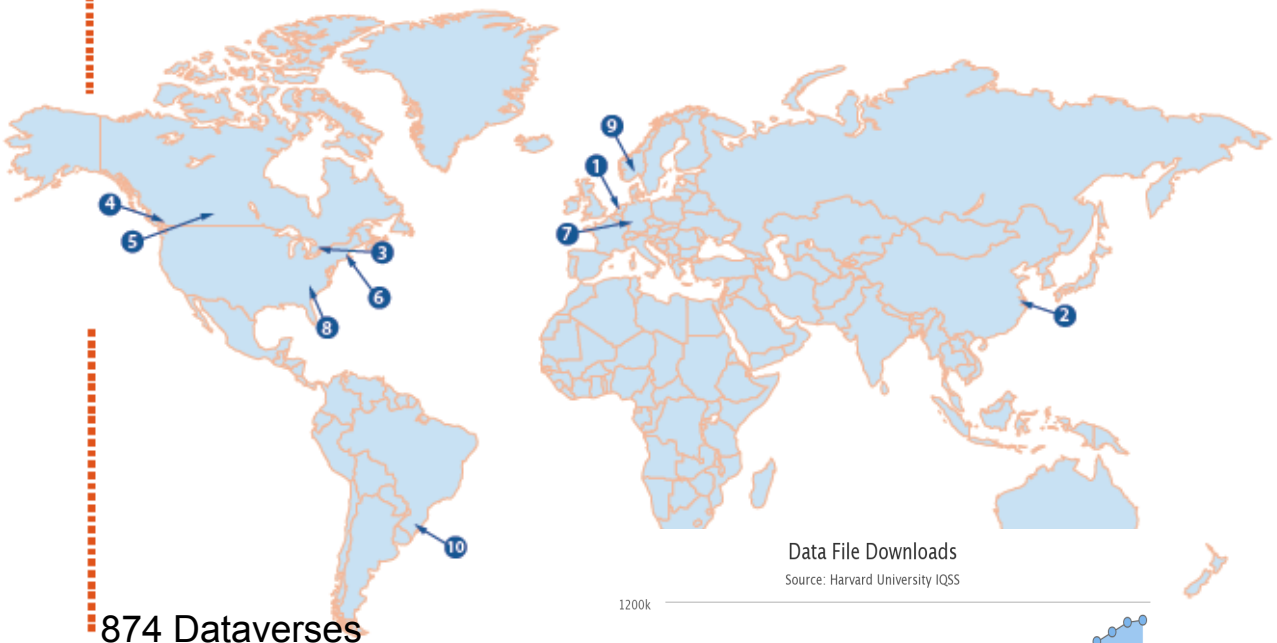
Geospatial Data: Visualize in WorldMap



Demo acknowledgement: Dwayne Liburd, Sonia Barbosa

Not only Expanding in Features, but also in Size

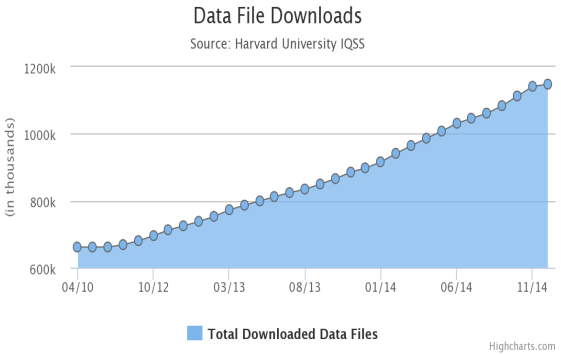
Federated Dataverse Installations:



1. DANS - Netherlands
2. Fudan University
3. OCUL Scholars Portal
4. Abacus
5. University of Alberta Libraries
6. Harvard Dataverse
7. Heidelberg University
8. ODUM Institute
9. UiT The Arctic University of Norway
10. UnBraL Fronteiras

@dataverseorg
1,480 Followers

874 Dataverses
55,539 Datasets
754,816 Files
1,173,733 Downloads



IQSS
The Institute for Quantitative Social Science
HARVARD UNIVERSITY

@IQSS
www.iq.harvard.edu
datascience.iq.harvard.edu

What’s coming

ODAP

The Open Data Assistance Program at Harvard

Many Harvard researchers are subject to open-data policies from the journals publishing their articles or the agencies funding their research. Many others simply want to open up their data to realize the benefits of transparency, collaboration, data citation, research acceleration, and reproducibility. ODAP is a program to help them.

ODAP will offer advice and instruction on how to deposit data files in the Harvard Dataverse. When privacy is an issue, ODAP will offer advice on how to make data files as open as privacy constraints will allow. Since anyone in the world may deposit in Dataverse, ODAP's online assistance should help researchers everywhere. However, when online assistance isn't enough, ODAP staffers and volunteers will offer personal assistance to Harvard faculty, students, fellows, and postdocs. We encourage other institutions to offer personal assistance to their own researchers as well, and can work with them on how to do that.

If you're interested in providing open access to your data, you should also be interested in providing open access to the research articles reporting your analysis and conclusions. If you're at Harvard, we welcome your research publications, especially your scholarly articles, in our open-access repository, Digital Access to Scholarship at Harvard (DASH). For more details, see the Office for Scholarly Communication. However, the present web site is about opening access to data files, not opening access to texts.

What the Harvard Community is Saying

"Data is one of the most vital resources of the 21st century. Fortunately, with Harvard's Dataverse, data associated with research can be stored and made accessible freely so others around the world can replicate studies and reuse it for new purposes. The Harvard Library will link publications in DASH, Harvard's institutional repository, to data in Dataverse, enriching the pool of open access information."

Sarah Thomas
Vice President for the Harvard Library
Roy E. Larsen Librarian for the Faculty of Arts and Sciences

Beyond 4.0

- Integration with other Systems:
  - DASH
  - ORCID
  - Journal Systems (in addition to OJS)
  - Archivematica
  - iRODS
- Support for Sensitive Data:
  - Secure Storage
  - DataTags
  - Analysis with Privacy Preserving Algorithms
- Data Citation with Dataset Provenance
- Expanding APIs!

1st Annual Dataverse Community Meeting

Institute for Quantitative Social Science (IQSS) at Harvard University
Tuesday, June 9, 2015 at 9:00 AM - Thursday, June 11, 2015 at 5:00 PM (EDT)
Cambridge, MA

The Dataverse Project

Registration Information

TYPE	END		QUANTITY
RSVP	Jun 9, 2015	Free	1

Register

Who's Going

Connect to see which of your Facebook friends are going to 1st Annual Dataverse Community Meeting.

Connect with Facebook

Share 1st Annual Dataverse Community Meeting

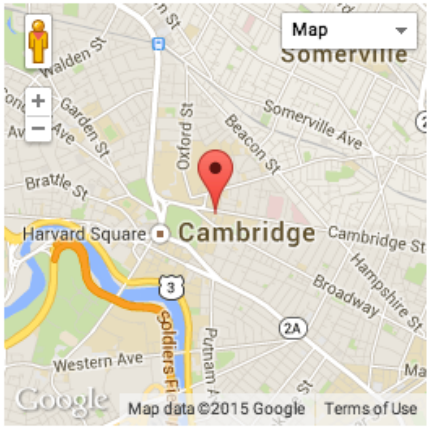
Email | Share | Tweet | Like Be the first of your friends to like this.

Event Details

You are cordially invited, as active contributors or interested members of the international Dataverse community, to take part in the 1st Annual Dataverse Community Meeting, which includes a Repository API workshop. The meeting will take place June 9-11, 2015 and will be hosted by the Institute for Quantitative Social Science (IQSS) at Harvard University.

Save This Event

When & Where



CGIS South, Tsai Auditorium
1730 Cambridge Street
Cambridge, MA 02138

Tuesday, June 9, 2015 at 9:00 AM -
Thursday, June 11, 2015 at 5:00 PM (EDT)

Add to my calendar

Thank You

mcrosas@iq.harvard.edu

@mercecrosas

http://datascience.iq.harvard.edu/team