

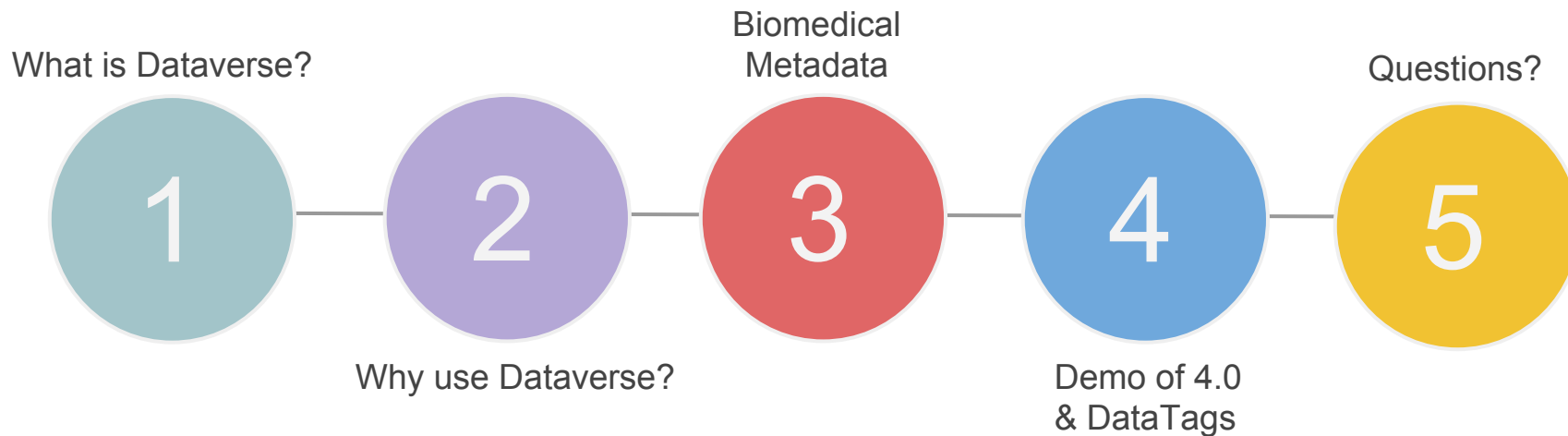
DATAVERSE

101: For Data Discovery Group

Elizabeth Quigley & Eleni Castro

Countway Library of Medicine - May 8, 2014

Agenda

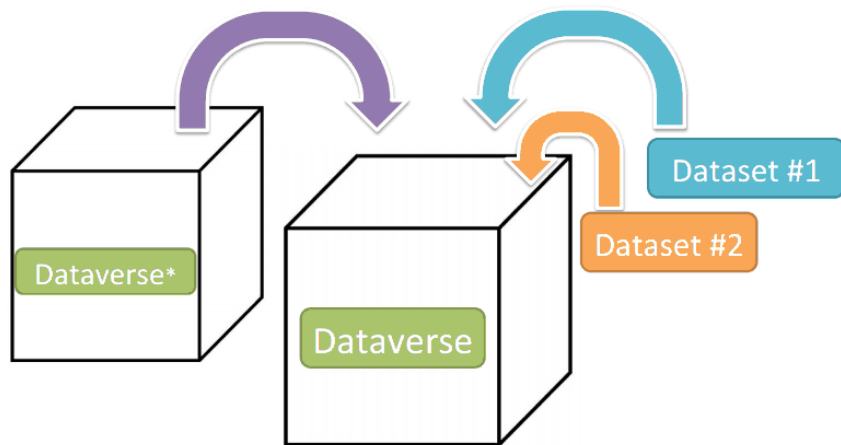


What is Dataverse?

A repository for sharing, citing, analyzing, and preserving research data.

What is a dataverse?

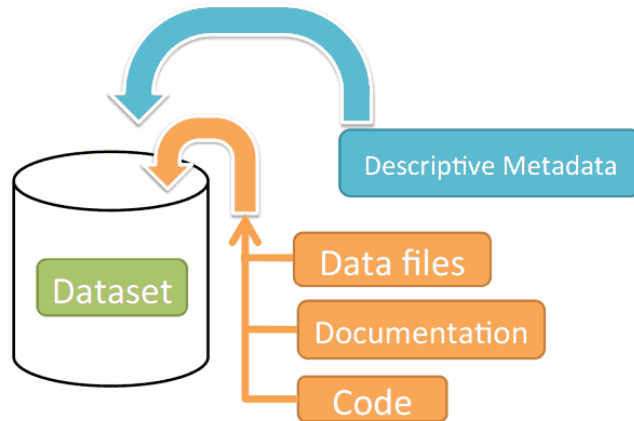
Schematic Diagram of a **Dataverse** in Dataverse 4.0



Container for your **Datasets** and/or **Dataverses***

* Dataverses can now contain other Dataverses (this replaces Collections & Subnetworks)

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



Container for your data, documentation, and code.

Why use Dataverse?

Upload datasets and files:

- Receive a formal data citation for a dataset and its files, with a persistent identifier (DOI).
- Provide as much descriptive metadata as possible:
 - for others to **discover** a dataset and its files
 - for others to **reuse** a dataset for their own research

Why use Dataverse?

Find and use data:

- Search for dataverses, datasets, and files (incl. variables)
- Browse dataverses, datasets, and files
- Download data
- Subset & Analysis within Dataverse using Zelig statistical analysis software (R)

Why use Dataverse?

Reasons to share data:

- Fulfill data management plan requirements (e.g., PLOS, NIH, NSF, etc).
- Get recognition and credit via data citations.
- Allow collaborators to contribute to your Dataverse (e.g, Journal, Project).
- Restrict data to your team until ready to publish
- Facilitate discovery and reuse of your data through extensive metadata.
- Enable reproducible research, which contributes to the validation and verification of science.
- ...



photo: [Flickr Commons](#)

Biomedical Metadata

Initially working with Stem Cell Commons research metadata/ontologies for Proof of Concept.

Design Type

- ☐ Case Control
- ☐ Cross Sectional
- ☐ Not Specified
- ☐ Parallel Group Design
- ☐ Perturbation Design

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

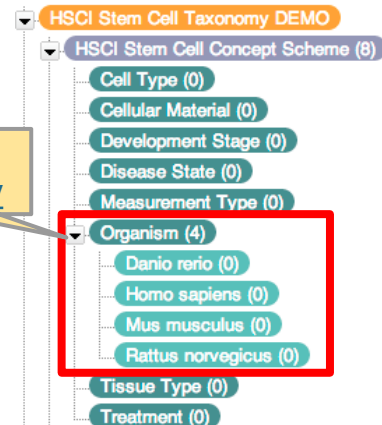
- ☐ Danio rerio
- ☐ Homo sapiens
- ☐ Mus musculus
- ☐ Rattus norvegicus

Cell Type



OBI
Ontology

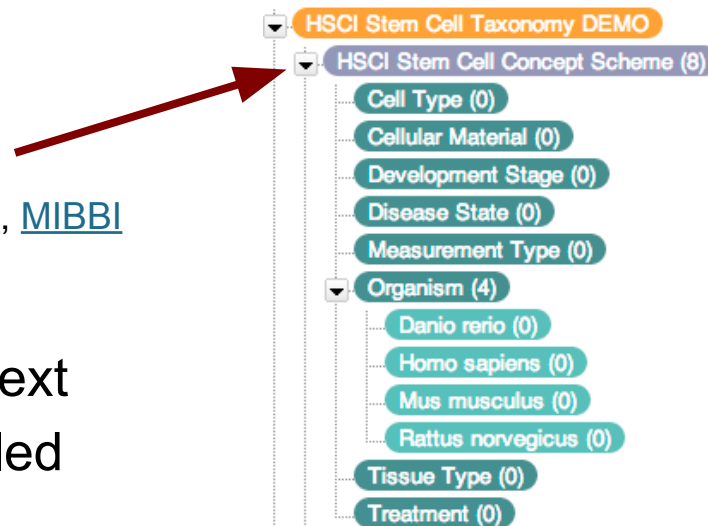
NCBI
Taxonomy



Bio Metadata Challenges

How feasible is it to support a one-size-fits-all solution for biomedical metadata?

1. Template Approach: work w/
different groups to support their
specific needs. (↑Time)
(e.g., HSCI [Stem Cell Commons](#), [FAIRPORT](#), [MIBBI](#)
(e.g., [fMRI](#))).
2. For broader support allow free text
(↓Quality) instead of just controlled
vocabularies (e.g., [OBO](#))



Transition from 3.6 to 4.0



Share, Cite, Reuse, Archive Research Data
Scientific data for reproducible research

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

Harvard Dataverse Network

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

The Harvard Dataverse Network is open to all scientific data from all disciplines worldwide. It includes the world's largest collection of social science research data. Learn more about the Dataverse Network.

Dataverses

124 Dataverses

Create Dataverse

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

RECENTLY RELEASED DATERVERSES

Elections Project	Feb 4, 2014
Blanca de Lizaur	Feb 3, 2014
Research of Archivasilp	Jan 31, 2014
Code Metrics	Jan 30, 2014
Timothy Ryan	Jan 30, 2014

View More >



Population Services International (PSI) Dataverses

Dataverses: 11

A global health organization that focuses on serious challenges such as a lack of family planning, HIV and AIDS, barriers to maternal health, and the greatest threats to children under five, including malaria, diarrhea, pneumonia and malnutrition.

• • • • •

Studies

52,970 Studies, 732,904 Files, 929,301 Downloads

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

RECENTLY RELEASED STUDIES

Block Group-Level Data, 1970: San Diego, California by University of California San Diego Library	Feb 12, 2014
Census of Population and Housing 1970, [San Diego, California]: Summary Statistic File 4A: Housing by NIA	Feb 12, 2014
HMS and MIS Data Summary by Kerdpairoj, Prad	Feb 12, 2014
Testing for near 1(2) trends when the signal to noise ratio is small [Dataset] by Jusellus, Katarina	Feb 12, 2014
Correlates of the Fear of Success in Black High School Women, 1974 by Jacqueline Fleming	Feb 11, 2014

View More >

MOST DOWNLOADED STUDIES

Replication data for: A Multivariate Model of Strategic Asset Allocation by John Y. Campbell, Yeung L. Chan, and Luis Viciara	18723
10 Million International Dyadic Events by Gary King, Will Lowe	17985
Replication data for: Asset Prices, Consumption, and the Business Cycle by John Y. Campbell	17661
Measuring the impact of microfinance in Hyderabad, India by Abhijit Banerjee, Esther Duflo, Rachel Glennerster, Cynthia Kinnan	15484
Textbooks and Test Scores by Paul Glewwe, Michael Kremer, Sylvie Moulin	15435

View More >

Transition from 3.6 to 4.0

What's changing?

- User interface and simpler workflows
- Support for more metadata
- Account page redeveloped with new functionality
- New interactive data exploration and analysis tool



Let's check out Dataverse 4.0 Beta.....

dataverse-demo.iq.harvard.edu

The logo for The Dataverse Network Project. It features the text "The Dataverse Network Project" in a serif font, with "Dataverse" in a larger, bold, orange font and "Network" in a smaller, bold, orange font. "The" and "Project" are in a smaller, grey font. To the right of the text is a stylized network diagram consisting of three grey circles connected by lines.

Tools for sharing privacy sensitive research data.

PRIVACY LAB

The Dataverse Network[®] Project



Questions?

To learn more about Dataverse and the Data Science Team at IQSS, visit datascience.iq.harvard.edu

Elizabeth Quigley: equigley@iq.harvard.edu & Eleni Castro: ecastro@fas.harvard.edu