Dataverse Software - New Features and Future Plans

Gustavo Durand

Technical Lead / Architect

IQSS, Harvard University

The Dataverse® Project

IQSS
The Institute for Quantitative Social Science

HARVARD LIBRARY

HARVARD UNIVERSITY
Information Technology

Introduction to Dataverse

Overview

- **An open-source platform to publish, cite, and archive research data**
- Built to support multiple types of data, users, and workflows
- Developed at Harvard's Institute for Quantitative Social Science (IQSS) since 2006
- Development funded by IQSS and with grants, in collaboration with institutions around the world
- Core team
  - @ IQSS - developers, designers, UX/UI, metadata specialists, curation team, leadership team
  - key contributors from the community

Dataverse Features

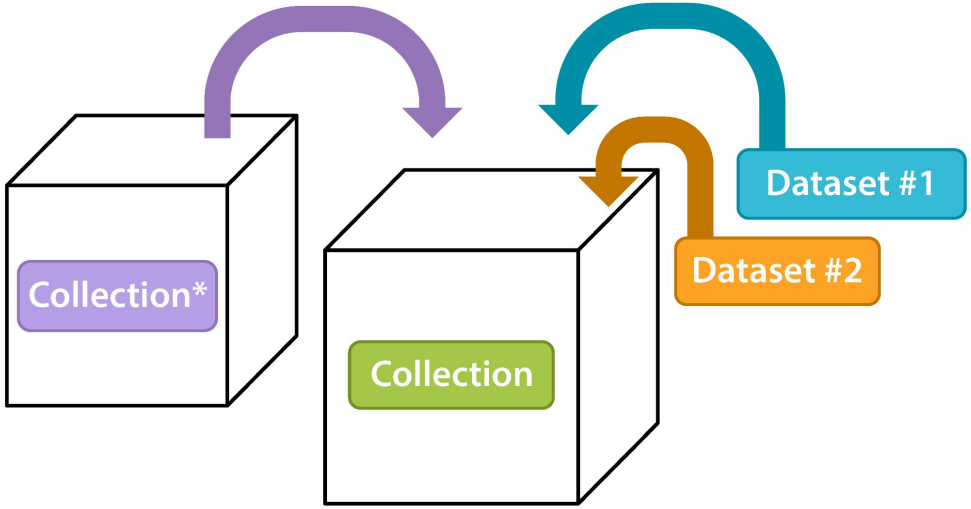
https://dataverse.org/software-features

- Main goal of core code is to focus on publishing (citing, sharing, versioning, etc.), FAIR Data principles
- Robust APIs to allow interoperability with “external tools” and other repositories / software

Dataverse Collections

- Ability to create Dataverse collections to organize datasets according to your needs
- Dataverses collections can also contain other collections, enabling any hierarchical structure
- Different rules can be applied for different Dataverse collections, e.g. for Metadata, Permissions, etc.

Schematic Diagram of a **Collection** in Dataverse Software 5.0

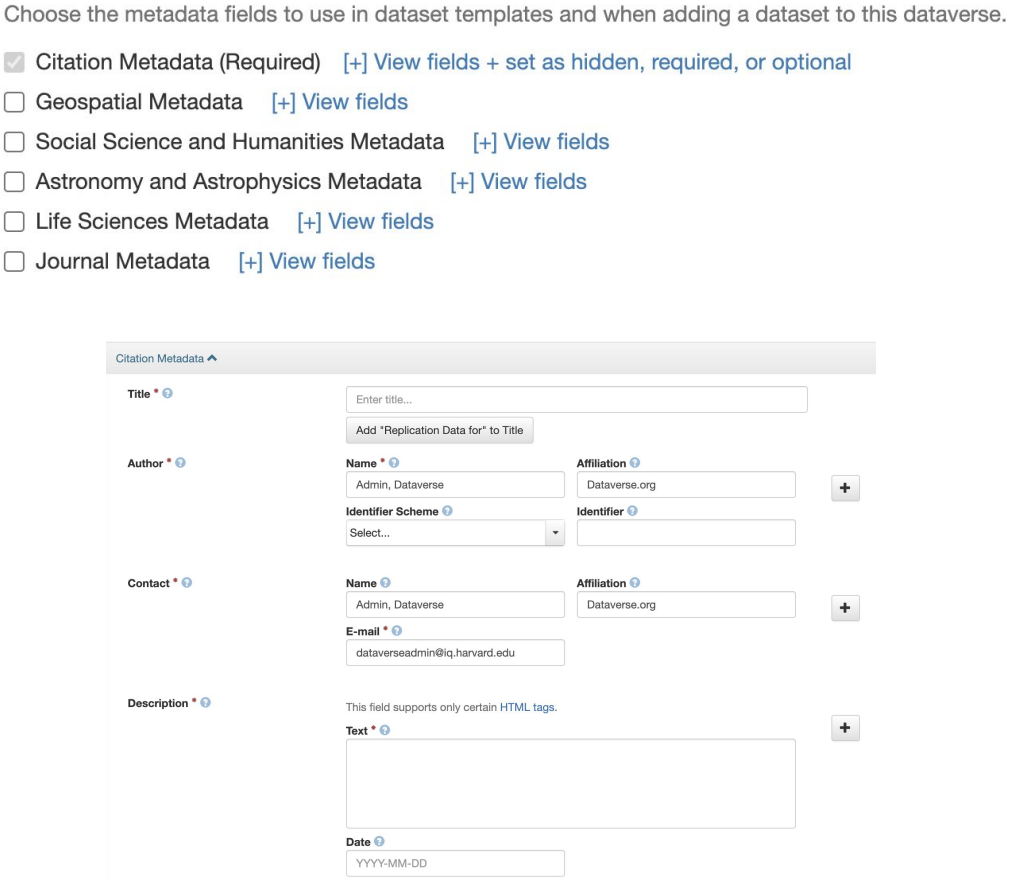


Container for your **Datasets** and/or **Collections***

* Collections can contain other Collections

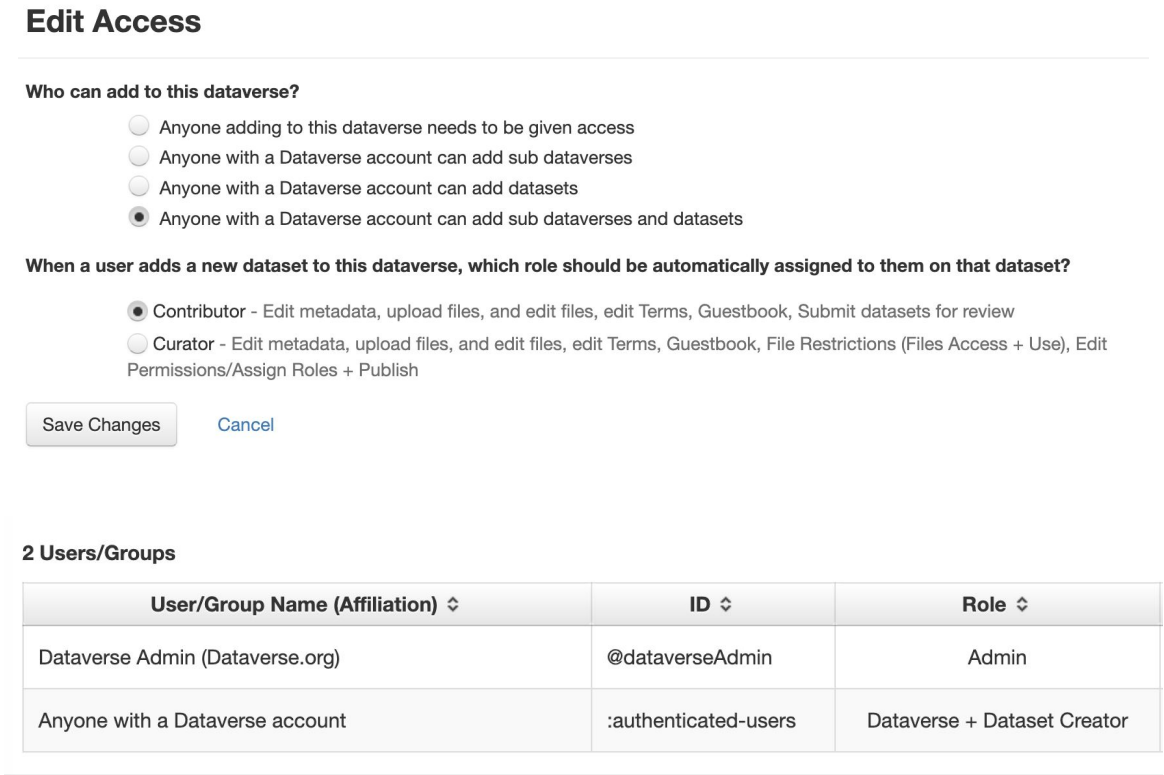
Dynamic Metadata

- Metadata is defined dynamically at the database level, allowing for modularly adding new Metadata blocks
- Supports:
  - single or multiple values
  - simple or compound values
  - controlled vocabularies



Flexible Permission System

- Supports multiple workflows by controlling who can add to your Dataverse collection, what they can, and what role they have on and created Datasets
- Roles are defined as a set of permissions to grant to users or to groups
- Groups can be defined statically or dynamically (e.g. users logging in from the same institution, via Shibboleth)



Robust APIs

- APIs for search, deposit, access, administration, metrics, etc.
- Additional APIs for harvesting (discovery) and interoperability with other systems
- External tools can be registered via APIs, so that Dataverse can provide links in the UI, then user is sent to tool to preview, explore, configure, and more

API Guide

Contents:

Dataverse Technology

Payara 5*

Java 11

Java EE8*

- Presentation: JSF (PrimeFaces), RESTful API
- Business: EJB, Transactions, Asynchronous, Timers
- Storage: JPA (Entities), Bean Validation

Storage: Postgres, Solr, File System / Swift / S3

Dataverse Community

Dataverse Community

- 155+ Github Contributors
- Hundreds of members of the Dataverse Community - developers, researchers, librarians, data scientists
  - Workshops & Trainings
  - UX/UI Testing & Interviews
  - Global Dataverse Community Consortium
  - Dataverse Google Group / Matrix / Community Slack
  - Dataverse Community Calls
  - Dataverse Community Meeting

Global Dataverse Community Consortium

- Supporting Dataverse repositories around the world

The Global Dataverse Community Consortium (GDCC) is dedicated to providing international organization to existing Dataverse community efforts, and will provide a collaborative venue for institutions to leverage economies of scale in support of Dataverse repositories around the world.

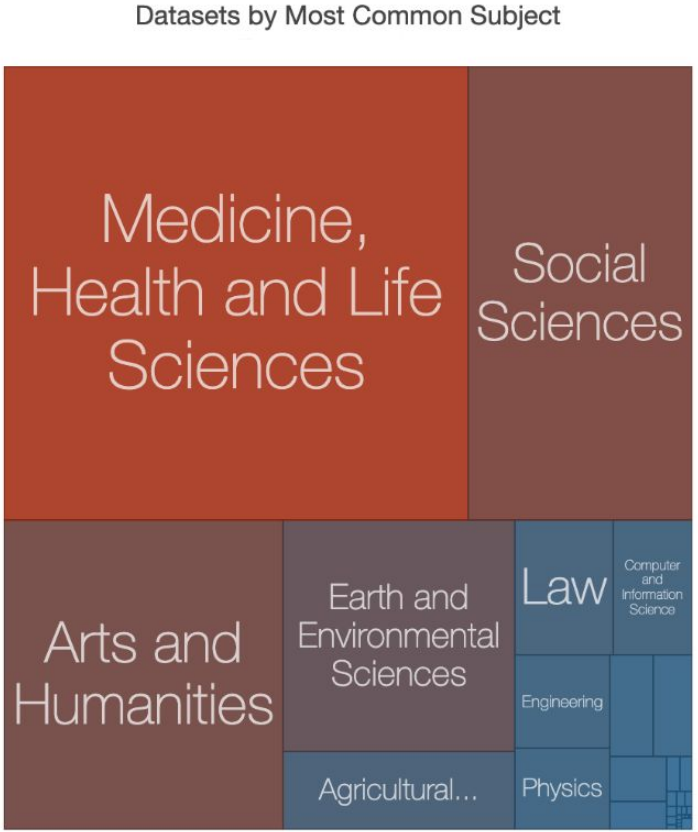
http://DataverseCommunity.Global

Dataverse Community

- **95** (self reporting) installations around the world, in **34** countries

The Data (dataverse.org/metrics)

- 95 installations
- 14,900 Dataverse Collections*
- 239,000 Datasets*
- 2,430,000 Files*
- 68,400,000 File Downloads*



* metrics collected from 60 installations
(running 4.9 and newer)

New Features

Harvard Data Commons

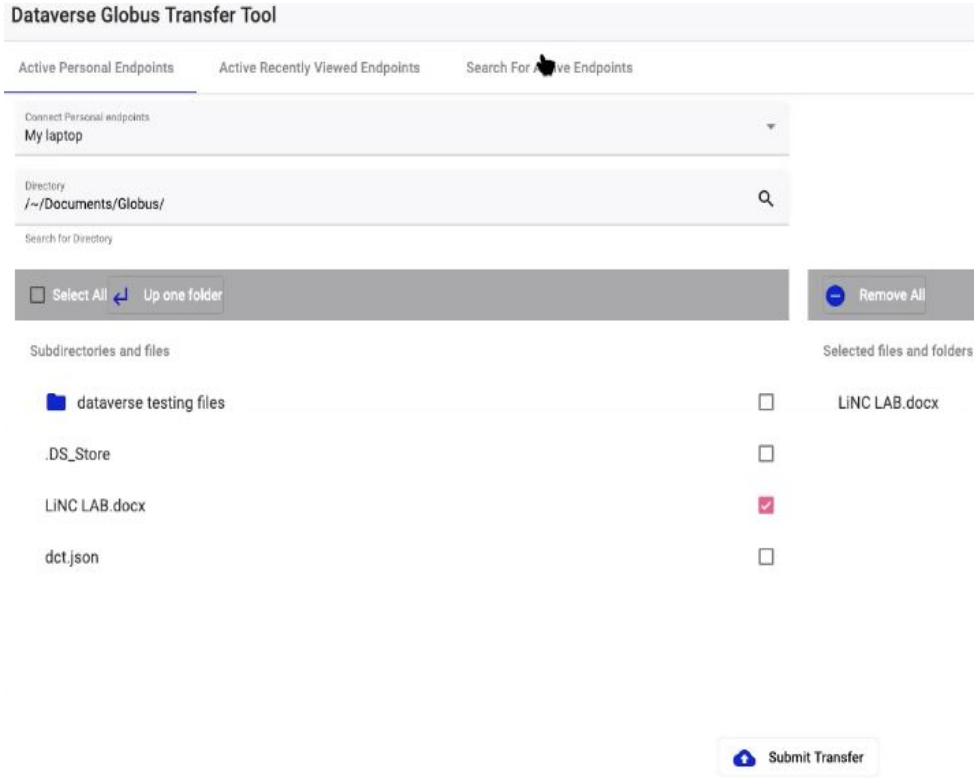
Automating the flow of research data from research computing environments to management, publication, discovery and preservation environments

Three Primary Objectives:

- Automating the technical pipeline between the research computing infrastructures and Dataverse
- Enhancing Dataverse to support machine-actionable workflows of various types
- Automating connections between research systems and key library systems used for archiving and publication

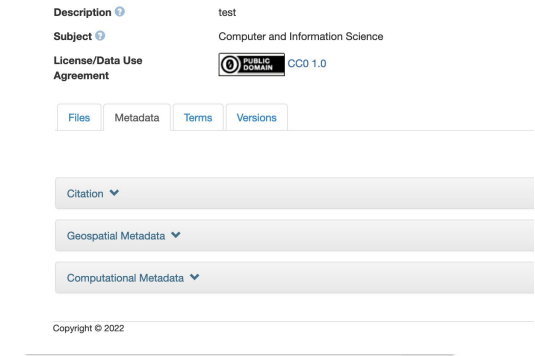
Globus Integration

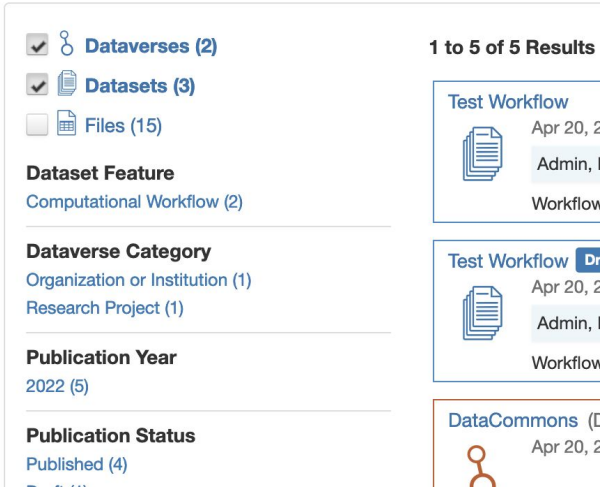
- Ability to add files (or just file metadata, leaving file at source) to Dataverse
- Developed as an external tool to be integrated into the Dataverse upload workflow



Computational Workflow Support

- External tools exist to support workflows and reproducibility
- Additional Metadata added to Dataverse to better support discoverability (e.g. the new “Dataset Feature” facet)
- Automatic checksum validation on BagIt file upload (based on the BagIt manifest)



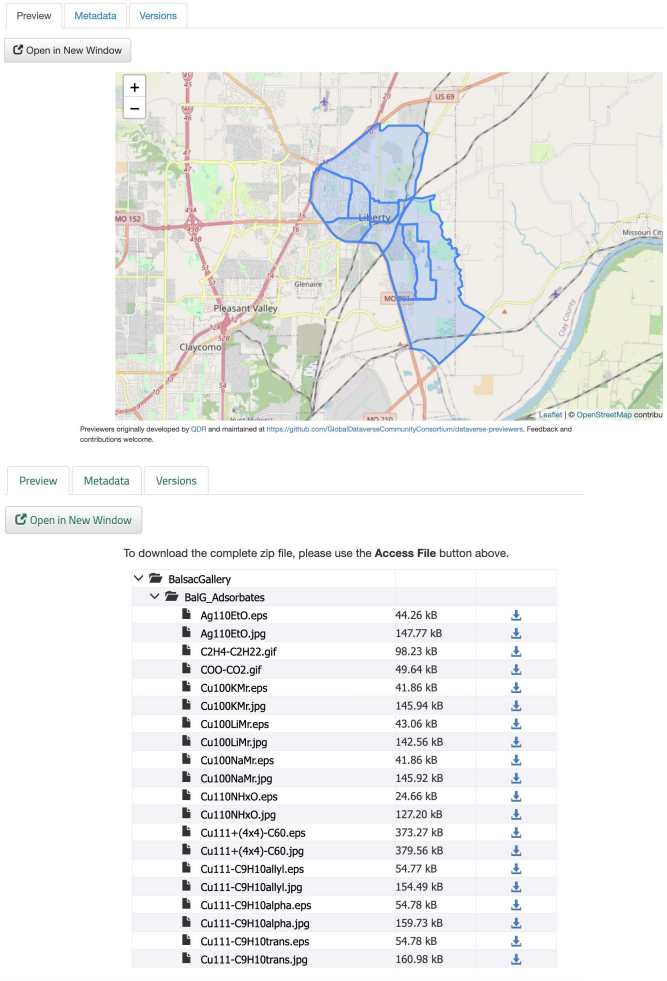


Improved Connections Between Systems

- New, more robust archiving features to S3
  - Includes new archive status API and admin display
- Bidirectional Notification of Related Resources
  - Ability to send and receive Linked Data Notification messages about relationships between datasets <--> papers / other resources
  - Will eventually be compliant with the COAR Notify protocol
  - Bonus feature: ability to add custom instructions to templates

Some Recent External Tools

- Map Previewer
  - Supports geoJson files
  - Available at the GDCC repo
- Zip File Previewer+
  - Uses the Range functionality in our Access api; so it's not just a viewer, it's an individual file unpacker and downloader too



And as of Last Week…

Dataverse 5.13!

5.13

- Schema.org Improvements (Some Backward Incompatibility)
- Folder Uploads via Web UI (dvwebloader, S3 only)
- Long Descriptions of Collections (Dataverses) are Now Truncated
- License Sorting
- Metadata Field Production Location Now Repeatable, Facetable, and Enabled for Advanced Search
- Support for NetCDF and HDF5 Files
- Support for .eln Files (Electronic Laboratory Notebooks)
- Improved Security for External Tools
- Geospatial Search (API Only)
- Reproducibility and Code Execution with Binder
- CodeMeta (Software) Metadata Support (Experimental)
- Mechanism Added for Stopping a Harvest in Progress
- API Endpoint Listing Metadata Block Details has been Extended
- Advanced Database Settings
- Support for Cleaning up Leftover Files in Dataset Storage
- OAI Server Bug Fixed
- . . .

Dataverse 5.13

https://github.com/IQSS/dataverse/releases/tag/v5.13

- Try it out at: https://demo.dataverse.org

Future Plans

Activities for Dataverse Team @ Harvard

- Harvard Dataverse Support
- Community Development Facilitation
- NIH GREI (Generalist Repository Ecosystem Initiative)
  - Individual Proposal
  - “Coopetition” Activities
    - Dataverse, Dryad, Figshare, Mendeley Data, Open Science Framework, Vivli
- Other Dataverse related projects and partnerships

NIH GREI Program Activities

- Remote Large Storage Support
- Controlled Vocabularies for Biomedical
- Discovery for DDI-CDI
- Software and Biomedical Workflows
- Harvesting and Sharing Metadata Across Repositories
- Usage Metrics - Make Data Count Support
- Revisiting of Sensitive Data Support
- Evaluation and Evolution of Architecture
- NIH Data Management Plans
- Training

Upcoming Release Plans

- In **May / June**, we are planning for **concurrent** releases of:
  - **5.14**
    - Consisting of usual new features, bug fixes, and community contributions
  - **6.0**
    - Functionally the same as 5.14, but running on latest Payara 6 server and other latest technologies
  - **Beyond 6.0**
    - Re-architecture project
      - Separate front end into separate SPA
      - Expand Dataverse back end APIs
      - Expand Modularity to further empower the community

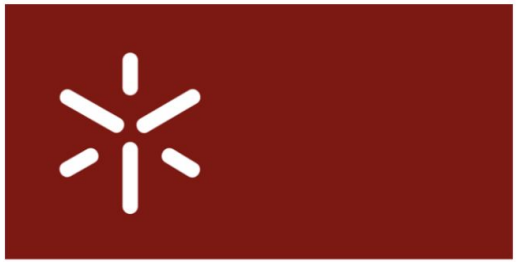
Dataverse Roadmap

https://www.iq.harvard.edu/roadmap-dataverse-project

- Strategic Goals
- Implementation, Planning, Future

Thank you

Dataverse Community Meeting 2023
June 5–7, 2023
University of Minho in Braga, Portugal



https://projects.iq.harvard.edu/dcm2023

The Dataverse® Project

Open source research data repository software

Researchers
Enjoy full control over your data. Receive *web visibility, academic credit,* and *increased citation counts.* A personal dataverse is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable to the research community, and satisfies data management plans. Want to set up your personal dataverse?

Journals
Seamlessly manage the submission, review, and publication of data associated with published articles. Establish an *unbreakable link* between *articles in your journal* and *associated data.* Participate in the open data movement by using Dataverse as part of your journal data policy or list of repository recommendations. Want to find out more about journal dataverses?

Institutions
Establish a research data management solution for your community. Federate with a growing list of Dataverse repositories worldwide for increased discoverability of your community's data. Participate in the drive to set norms for sharing, preserving, citing, exploring, and analyzing research data. Want to install a Dataverse repository?

Developers
Participate in a vibrant and growing community that is helping to drive the norms for sharing, preserving, citing, exploring, and analyzing research data. Contribute code extensions, documentation, testing, and/or standards. *Integrate research analysis, visualization* and *exploration tools,* or other research and data archival systems with Dataverse. Want to contribute?

https://dataverse.org
https://github.com/iqss/dataverse