Dataverse

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, UX/UI, metadata specialists, curation team, leadership team
  - key contributors from the community with full privileges as IQSS team

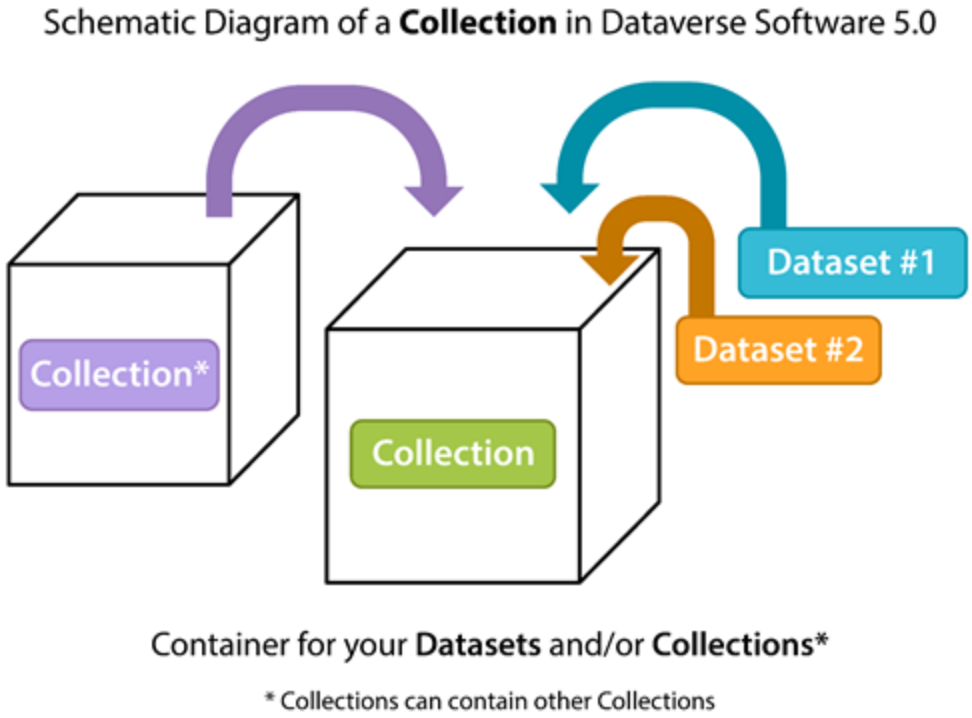
Dataverse Features

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

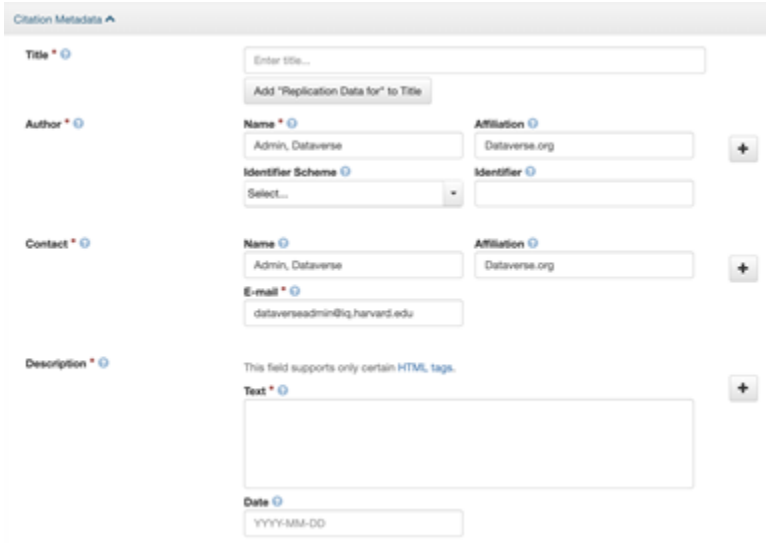
Dataverse Collections

- 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.



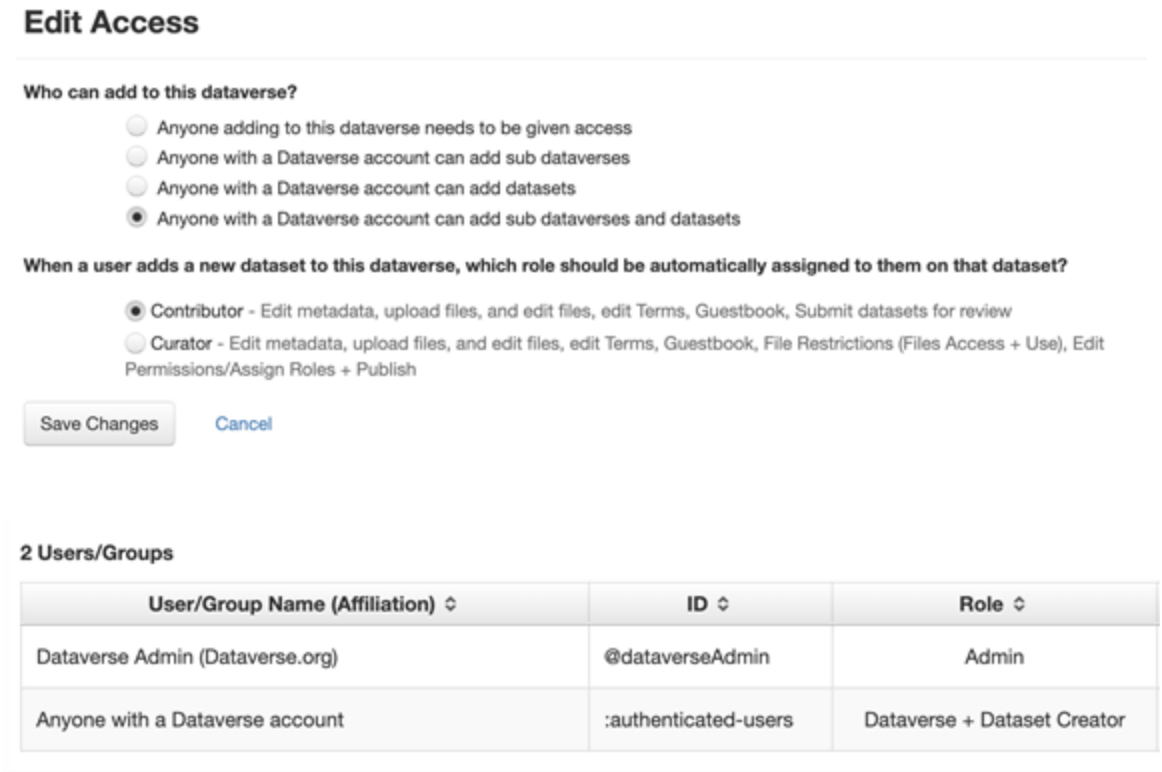
Dynamic Metadata

- 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 7

Java 21

Java EE10

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

Front end: Modern SPA (React) - **Coming Very Soon!**

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

Dataverse Community

Dataverse Community

- 220+ Github Contributors
- Thousands of members of the Dataverse Community - developers, researchers, librarians, data scientists
  - Workshops & Trainings
  - UX/UI Testing & Interviews
  - Global Dataverse Community Consortium
  - Dataverse Google Group / Zulip
  - Dataverse Community Calls
  - Dataverse Community Meeting (**next week!**)

The Dataverse Cup 🏆

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.

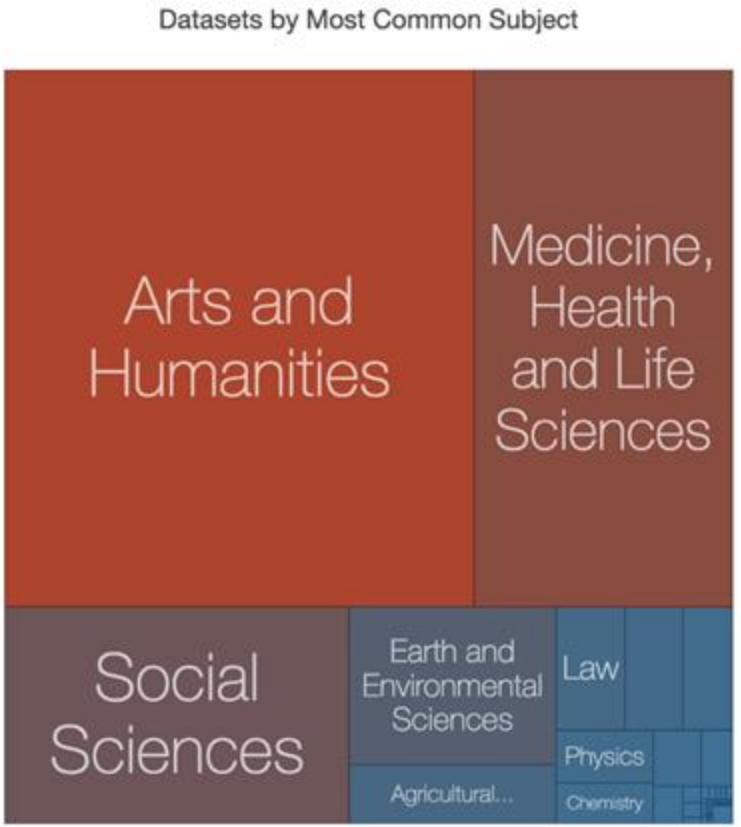
http://DataverseCommunity.Global

Dataverse Community

- 147 (self reporting) installations around the world

The Data (dataverse.org/metrics)

- 147 installations
- 22,900 Dataverse Collections*
- 497,000 Datasets*
- 12,200,000 Files*
- 230,000,000 File Downloads*



* metrics collected from 110 installations (running 4.9 and newer)

Recent Releases

Recent Releases

dataverse.lv currently running version **6.7.1**

Releases since then:

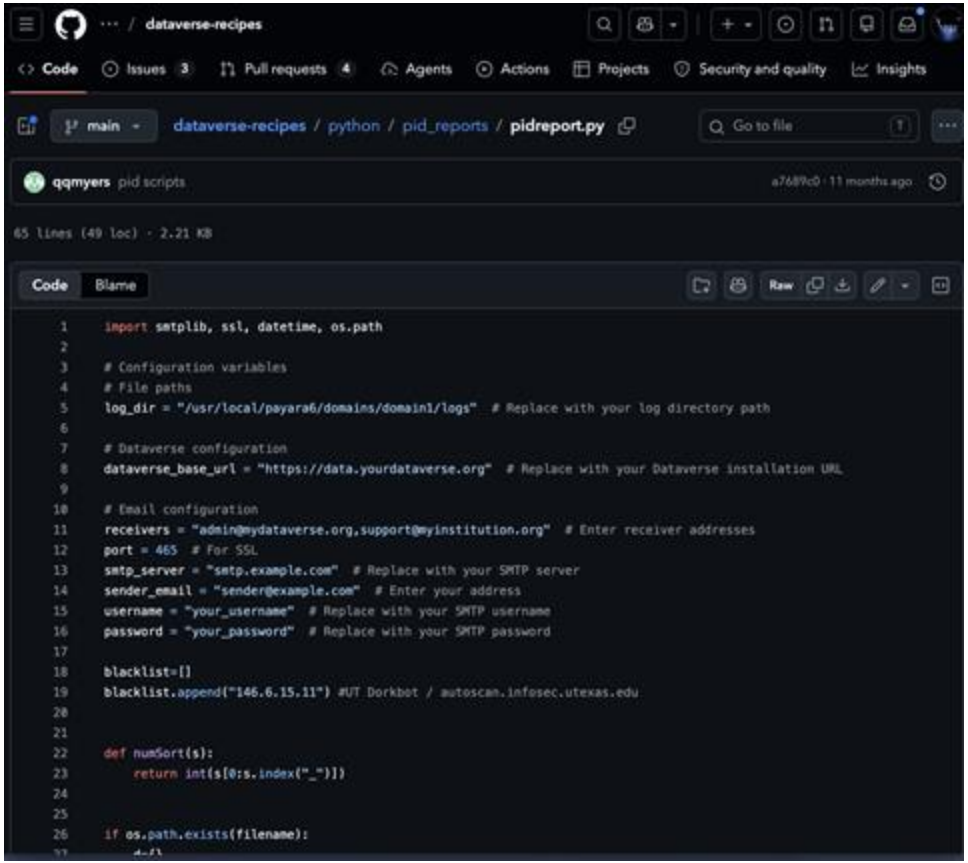
- **6.8** - September 2025
- **6.9** - December 2025
- **6.10 / 6.10.1** - March 2026

Dataverse 6.8

- Highlights - https://github.com/IQSS/dataverse/releases/tag/v6.8
  - Open OnDemand integration
  - Logs for diagnosing PID failures
  - Link permission split off from publish permission
  - New and improved APIs
  - Bug Fixes

Logs for diagnosing PID failures

- A new feature flag called **enable-pid-failure-log** can be enabled to help diagnose PID failures, e.g:
  - when users share PIDs from draft datasets/files
  - or share a URL to a dataset/file page with issues such as a trailing '.' character
- When set, Dataverse will log requests for dataset and file pages via persistentId that fail in monthly log files of the form PIDFailures_<yyyy-MM>.log
- The new log files can be used in concert with the **pidreport.py** script to generate and email monthly PID failure reports: https://github.com/gdcc/dataverse-recipes/blob/main/python/pid_reports/pidreport.py

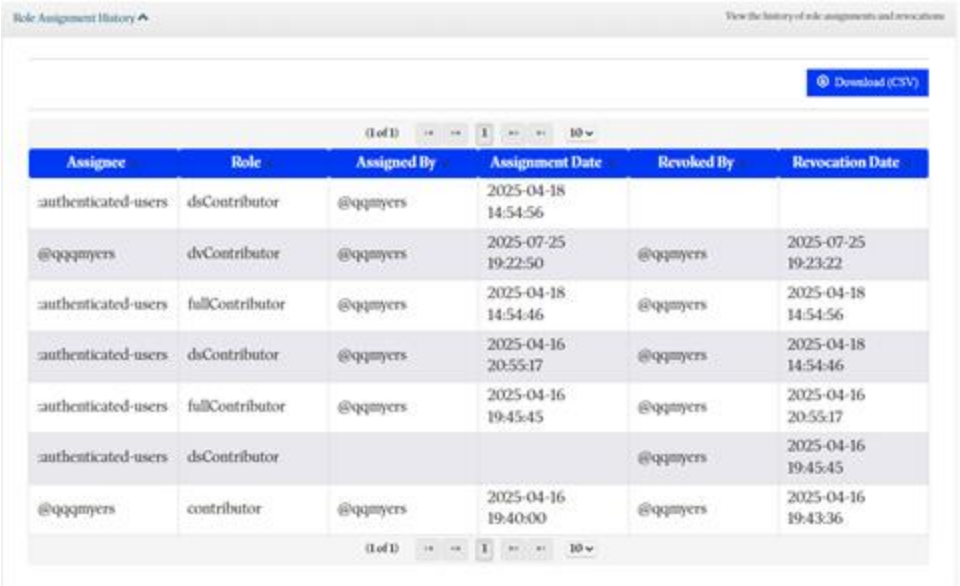


Dataverse 6.9

- Highlights - https://github.com/IQSS/dataverse/releases/tag/v6.9
  - Quickstart Guide
  - **Role Assignment History Tracking**
  - Scaling Dataverse with Data Size (Admin Guide)
  - Storage Quotas on Individual Datasets
  - Additional Licenses
  - DataCite Scaling
  - Support for COAR Notify Relationship Announcement
  - Infrastructure upgrade: Payara
  - New and improved APIs
  - Bug fixes

Role Assignment History Tracking

- Dataverse can now track the history of role assignments, allowing administrators to see:
  - who assigned or revoked roles,
  - when these actions occurred
  - which roles were involved
- This feature helps with auditing and understanding permission changes over time
- This feature is off by default but can be enabled with the "role-assignment-history" feature flag
- The information can also be downloaded via API in CSV and JSON formats: https://guides.dataverse.org/en/6.9/api/native-api.html#dataverse-role-assignment-history

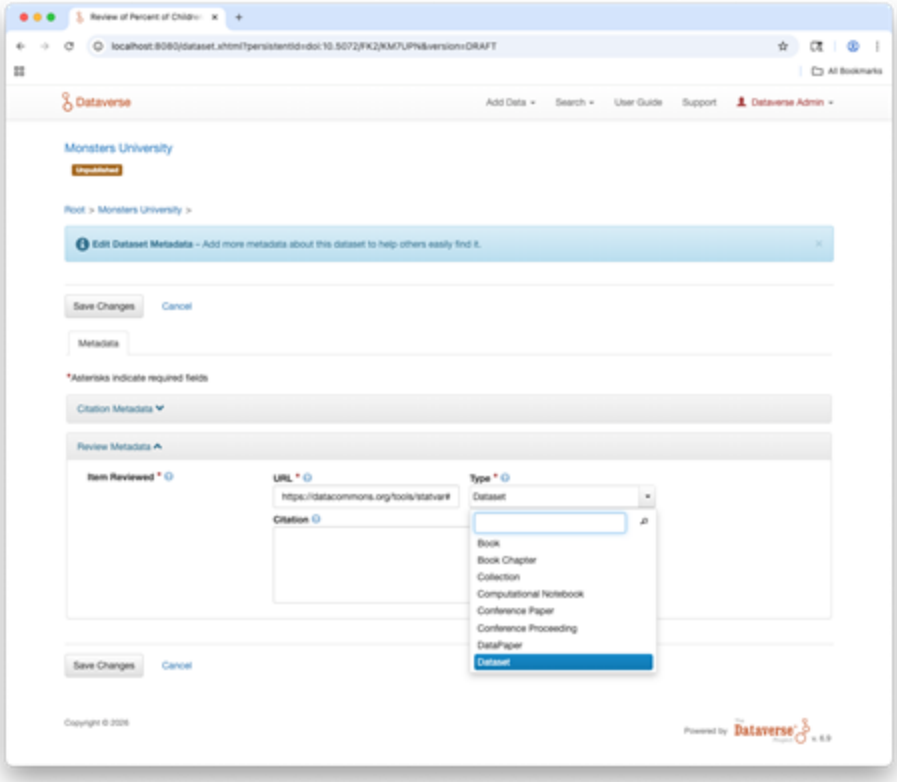


Dataverse 6.10 / 6.10.1

- Highlights - https://github.com/IQSS/dataverse/releases/tag/v6.10
  - Optionally require acknowledgment of a disclaimer when publishing
  - Optionally require embargo reason
  - Harvesting improvements
  - Croissant support now built in
  - Archiving, OAI-ORE, and BagIt export improvements
  - Support for REFI-QDA Codebook and Project files
  - **Review datasets**
  - Infrastructure: Payara has been upgraded from version 6 to 7
  - Infrastructure: Java has been upgraded from version 17 to 21
  - New and improved APIs, including filling in a guestbook when downloading files
  - Bug fixes

Review Datasets

- Review datasets are a specialized type of dataset that can be used to review resources (such as datasets) in the Dataverse installation itself or resources in external data repositories
- Review datasets build on the Dataset Types feature
- This feature is only available via API; the UI for this feature will only be available in a future version of the new React-based Dataverse Frontend
- Review Datasets are configured to use a new "review" metadata block
- When review datasets are published, different resourceType metadata is sent to DataCite
- **This feature is experimental**
- https://guides.dataverse.org/en/6.10/user/dataset-management.html#review-datasets



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, Open Science Framework, Mendeley Data, Vivli, Zenodo
- 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

- **6.11** planned for **June 2026**
- https://github.com/IQSS/dataverse/issues?q=milestone:6.11
- **Beyond**
  - Continue with timed releases every 3 months
  - Complete Modern Front end development / retirement of the JSF
  - Related projects: Dataverse Marketplace and Dataverse Hub
  - Continued work into more External Tools and AI integrations
  - External Search (Natural language)

Dataverse Roadmap

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

- Strategic Goals
- Roadmap
- Current and Past Projects

Thank you

Dataverse Community Meeting 2026
Barcelona Supercomputing Center

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