

Building Dataverse Communities that follow RDA Best Practices for Data Sharing and Management



Gustavo Durand, Jon Crabtree,
Philipp Conzett, Slava Tykhonov



<http://bit.ly/RDABOFDV>

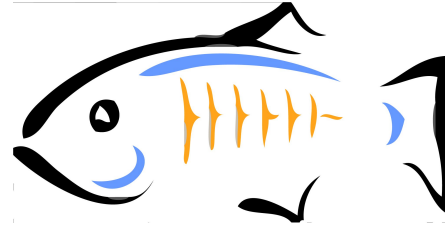
Introduction to Dataverse

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
- 12 on the core team - developers, designers, UI/UX, metadata specialists, curation team, leadership team

Dataverse Technology

Glassfish Server 4.1



Java SE8

Java EE7

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

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

Dataverse Features - Data

- Persistent IDs / URLs
 - DataCite
 - Handle
- Automatically Generated Citations with attribution
- Compliant with FAIR and data citation principles
- Domain-specific Metadata
- Versioning
- File Storage
 - Local
 - Swift (OpenStack)
 - S3 (Amazon)
- *DataTags for Sensitive Data*

Dataverse Features - Users

- Multiple Sign In options
 - Native
 - Shibboleth
 - OAuth (ORCID, Github, Google, *Microsoft*)
- Dataverses within Dataverses
- Branding
- Widgets

Dataverse Features - Workflows

- Permissions
- Access Controls and Terms of Use
- Publishing Workflows
- Private URLs
- Upload / Download Workflows
 - Browser
 - Dropbox
 - Rsync (for big data “packages”)
 - *Remote Storage (TRSAs)*

Dataverse Features - Interoperability

- APIs
 - SWORD
 - Native
 - Metrics
- Harvesting (OAI-PMH)
 - Client
 - Server
- Modular External Tools
 - Explore vs Configure
 - Scope: Dataset / Datafile

Dataverse Roadmap

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

- Strategic Goals
- Implementation, Planning, Future

Dataverse Community

Dataverse Community

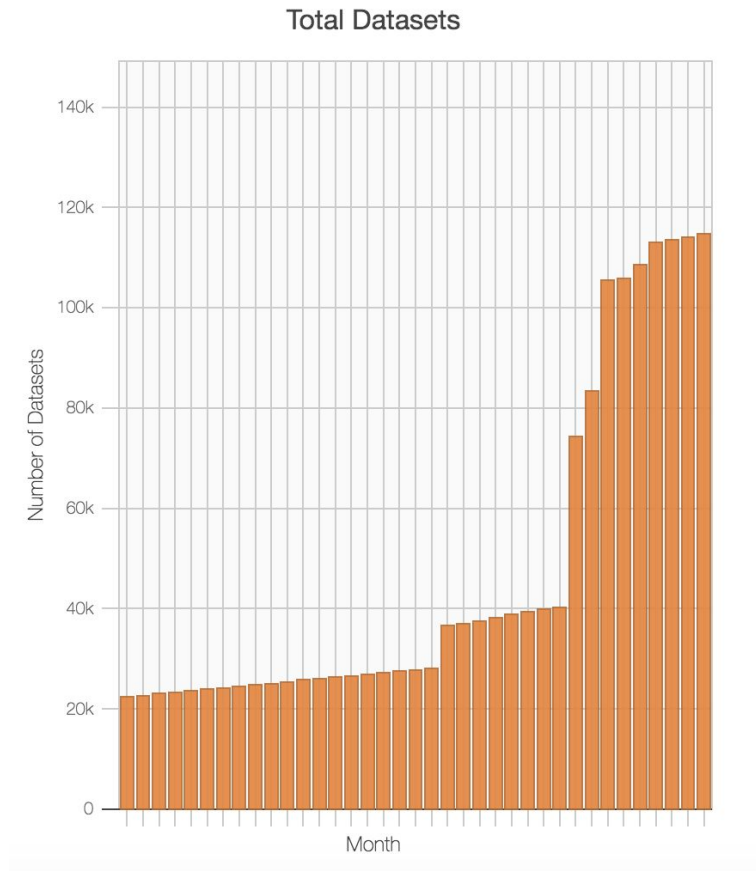
- **50** installations around the world



The Data (dataverse.org/metrics)

- 50 installations
- 5,500 Dataverses*
- 124,000 Datasets*
- 507,000 Files*
- 10,200,000 File Downloads*

* metrics collected from 26 installations



Dataverse Community

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

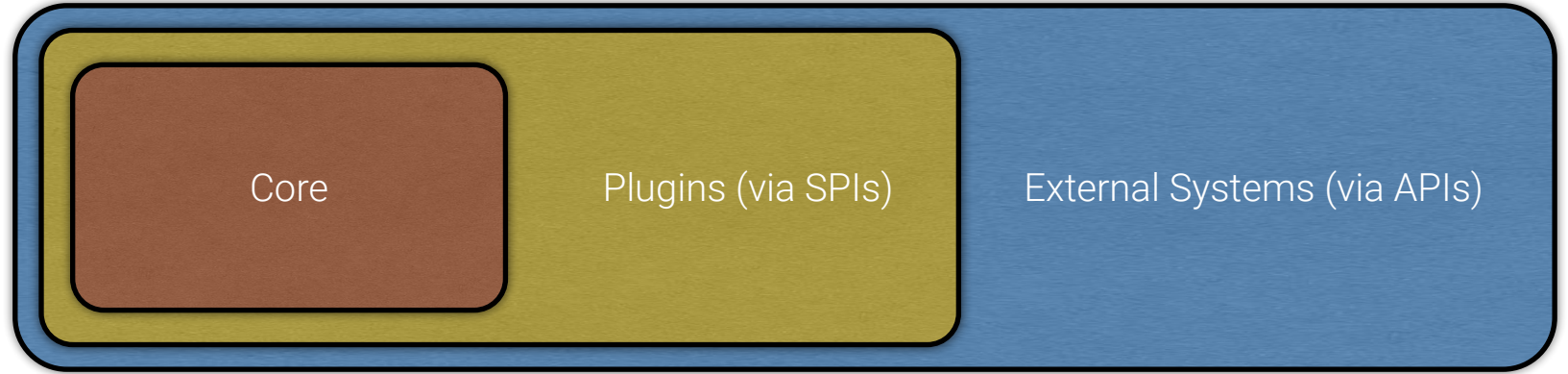
The Dataverse Cup 🏆



Community Development



Dataverse Ecosystem



Core - Contributing Code to the Dataverse Repo

- Let's talk early and often!
 - Preview vs Review
- We like small batches, but we'll follow your lead
- References
 - Developer's Guide
 - Style Guide
 - API Guide

SPIs / APIs - Why Modularity Matters

- Dataverse is a big application that serves many disciplines with various different needs
 - Almost no-one uses the full functionality
- Modular design allows:
 - Easier code contributions
 - Tailoring installations to institution needs
 - Smaller, more efficient, core
- SPIs - **Dataverse** calling **custom code**
- APIs - **custom code** calling **Dataverse**

Example Collaborations (Core)

- SBGrid Data
 - Large Data and Support
- Massachusetts Open Cloud
 - Big Data Storage and Compute Access (OpenStack)
- Provenance
 - W3C PROV
- Australian Data Archive (ADA)
 - Use Guestbook for Request Access

Example Collaborations (SPIs)

- SBGrid Data
 - Pre Publish Workflows
- DANS/CIMMYT/GESIS
 - Handles
 - da|ra

Example Collaborations (APIs)

- File Access APIs (External Tools)
 - Harvard SEAS - TwoRavens
 - Scholars Portal - Data Explorer, Data Curation Tool
 - QDR - File Previewers for pdfs, images, videos
- Deposit APIs
 - Open Journal Systems - OJS Plugin
- Client Libraries
 - ResearchSpace - Java
 - AUSSDA - python - pyDataverse

Odum Institute

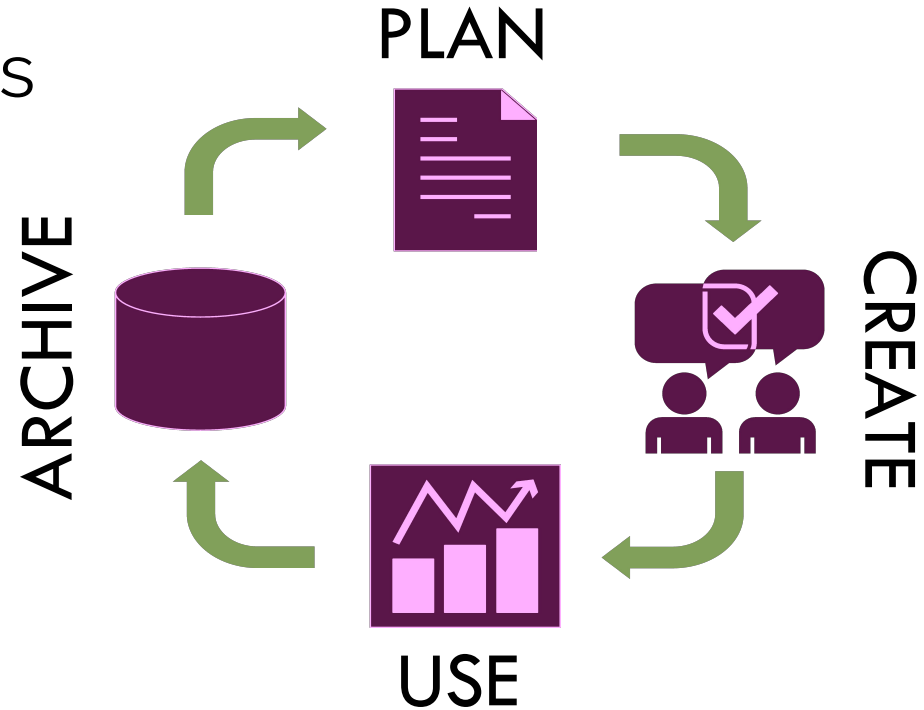


Who we are: The Odum Institute

- Founded in 1924 by Howard W. Odum
- Oldest university-based interdisciplinary social science research institute in the U.S.
- *Celebrating 95 years of service*
- *Our mission is to rigorously uphold the highest standards of scientific research across the University of North Carolina and the world, while simultaneously simplifying the research process for researchers we work with.*

Our Services: Across The Research Lifecycle

- Data Management Plans
- Data Collection
- Data Analysis
- Training & Education
- Consultations
- Facilities & Labs
- Cyberinfrastructure
- Data Archive





UNC Dataverse

UNC Dataverse

Hosted by the Odum Institute for Research in Social Science

Metrics

721,857 Downloads

Contact Share

Share, publish, and archive your data. Find and cite data across all research fields.



Search this dataverse...

Find Advanced Search

+ Add Data

Datasets (127)

Datasets (24,971)

Files (227,033)

Dataverse Category

Research Project (36)

Organization or Institution (16)

Researcher (11)

Journal (4)

Teaching Course (3)

More...

Metadata Source

Harvested (20,818)

UNC Dataverse (4,280)

1 to 10 of 25,098 Results

Sort

Monmouth University New Jersey Poll, Number 177

May 9, 2019 - Monmouth Polling Institute Dataverse



Monmouth University Polling Institute, 2019, "Monmouth University New Jersey Poll, Number 177", <https://doi.org/10.15139/S3/YD6SIU>, UNC Dataverse, V1, UNF:6XhgdzlQHmDD4nYS6mulfw== [fileUNF]

This survey was conducted among New Jersey residents likely to vote in the November election.

CMAQ Model Version 5.2 Output Data -- 2014 CONUS_12km

May 8, 2019 - CMAS Data Warehouse



US EPA, 2019, "CMAQ Model Version 5.2 Output Data -- 2014 CONUS_12km", <https://doi.org/10.15139/S3/XYW3HL>, UNC Dataverse, V1

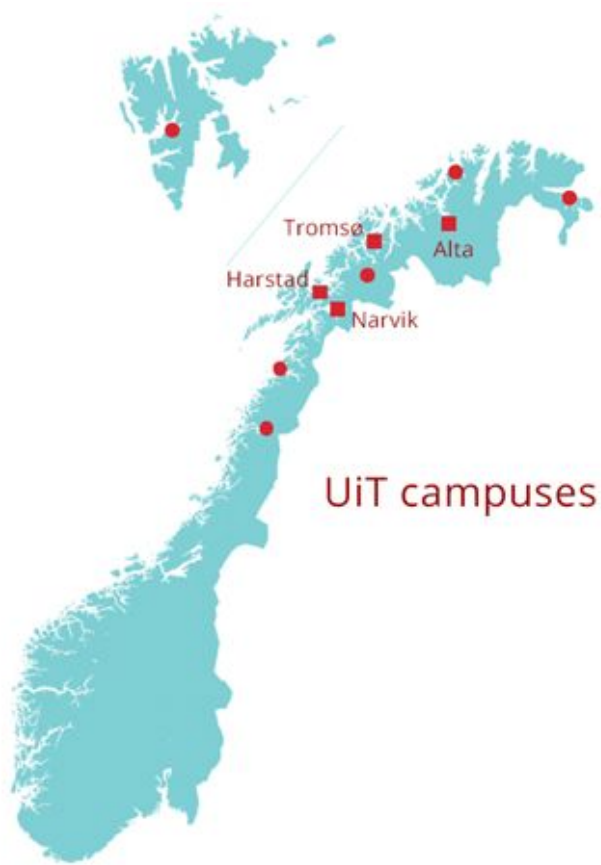
Data Summary: Community Multiscale Air Quality (CMAQ) Model Version 5.2 output data from a 2014 CONUS simulation. Note: The datasets are on a Google Drive. The metadata associated with this DOI contain the link to the Google Drive folder and instructions for downloading the data.

DataverseNO

What is DataverseNO?

- A **national, generic repository** for open research data
- For researchers from **Norwegian research institutions**
- Some collections within DataverseNO also accept data from researchers from other institutions.
- Aligned with the **FAIR** Guiding Principles for scientific data management and stewardship
- Owned and operated by UiT The Arctic University of Norway (<https://en.uit.no/>)
- **Repository:** dataverse.no | **Info:** info.dataverse.no

Who is UiT The Arctic University of Norway?



- Northernmost university in the world
- Established in 1968
- 6 campuses
- 8 faculties covering all major disciplines
- 3 511 employees (FTE)
 - 2 191 faculty
- 16 747 students
 - 118 PhD students per year
- 254 study programs
 - 90 Master's degree programs (30 international)
 - 7 PhD programs

How does DataverseNO work?

- **8 partner institutions**, all of them **Norwegian universities**
- Each partner institution has its own **institutional collection**
- **All data is curated** by research data support staff at partner institutions
- **Common policies and guidelines** apply for the entire repository (info.dataverse.no)
- Applied for **CoreTrustSeal** certification

DataverseNO-institusjoner
pr. mai-2019



Organization of DataverseNO

ORGANIZATIONAL DOCUMENTS

DataverseNO Policy Framework

DataverseNO Steering Documents

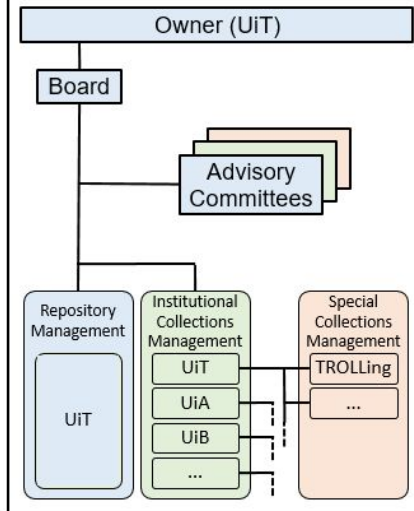
DataverseNO Administrator Guidelines

DataverseNO Curator Guidelines

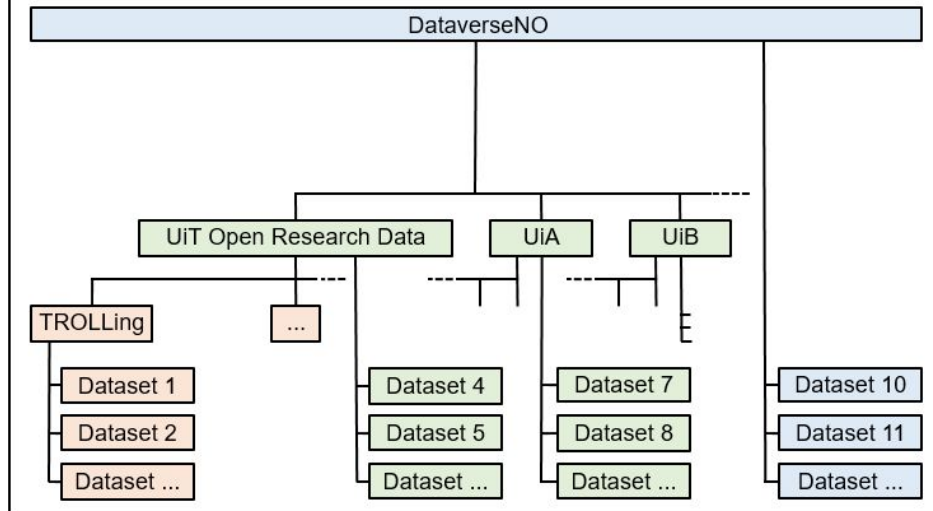
DataverseNO Deposit Guidelines

...

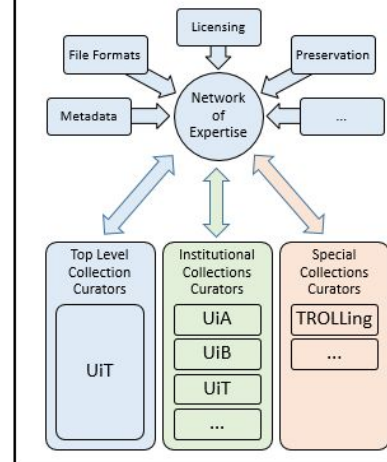
GOVERNANCE



REPOSITORY STRUCTURE



DATA CURATION



DESIGNATED COMMUNITY

Researchers within User Groups of Special Collections

Linguists

...

Researchers from Partner Institutions in Norway

HVL

INN

NMBU

NORD

NTNU

UiA

UiB

UiT

...

Researchers from Non-Partner Institutions in Norway

Join us at the European Dataverse Workshop 2020!



- **Venue:** UiT, Tromsø
- **Date:** January 23-24, 2020 --- during the Northern Lights season!
- More info on the workshop homepage:

<https://tinyurl.com/dataverse2020>

Dataverse in EU

SSHOC is EU Social Sciences and Humanities Open Cloud

- DANS-KNAW (Netherlands) established Dataverse as a service for Dutch Universities in May, 2014.
- The goal of SSHOC Dataverse project (CESSDA, DARIAH and CLARIN) is to create a reliable and production ready Open Source data infrastructure that everybody can install and reuse for their own needs and requirements.
- We're developing multilingual web interface and localizing metadata fields and developed data standardization technique based on APIs for CESSDA CVs, Topic Classification and CESSDA CV Manager services.

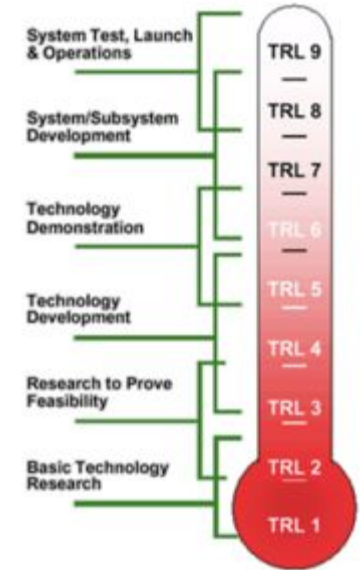


Production ready Dataverse infrastructure

- reliable and scalable Cloud service that can be deployed in Kubernetes
- out of the box installation on Google Cloud, Amazon AWS and Microsoft Azure
- can be connected to any research infrastructure by APIs
- distributed multilingual data infrastructure consisting of connected Dataverse nodes and forming a federated data portal
- repository already integrated with data previewers, external applications and VREs
- external controlled vocabulary support is the interoperability solution

Services in European Open Science Cloud (EOSC)

- EOSC requires the level 8 of maturity (at least)
- we need the highest quality of software to be accepted as a service
- clear and transparent evaluation of services is essential
- the evidence of technical maturity is the key to success
- the limited warranty will allow to stop out-of-warranty services



Dataverse App Store

Let's build different services out of tools!

Data preview: DDI Explorer, Spreadsheet/CSV, PDF, Text files, HTML, Images, video render, audio, JSON, GeoJSON/Shapefiles/Map, XML

Interoperability: external controlled vocabularies (CESSDA CV Manager)

Data processing: NESSTAR DDI migration tool

Linked Data: RDF compliance including SPARQL endpoint

Federated login as a service (OAuth/Shibboleth in the same installation)

Multilingual support

DataverseEU will run Weblate as a service for the user interface, metadata schema and SOLR translation.

We're developing an experimental but adjustable pipeline for multilingual support that allows to download and synchronize all translations available in Dataverse Consortium github and provides easy access for translators to keep all properties up-to-date.

Weblate as a Dataverse service

The screenshot shows the Weblate web interface. At the top, there are navigation links for 'Weblate', 'Dashboard', 'Projects', and 'Languages'. The breadcrumb path is 'DataverseEU / Bundle.properties ⚠ / bundle_DE (generated) / translate'. Below the breadcrumb is a navigation bar with 'All strings' (64 / 1568) and navigation arrows. A 'Zen' button is in the top right.

The main content area is a 'Translate' panel for the string 'htmlAllowedMsg'. It shows the source text 'htmlAllowedMsg' and the generated German translation 'Dieses Feld unterstützt nur bestimmte HTML-Tags.'. The context is 'htmlAllowedMsg'. There are buttons for 'Save', 'Suggest', and 'Skip'. Below the translation panel are tabs for 'Nearby strings' (11), 'Comments', 'Machine translation', 'Other languages', and 'History'.

Below the tabs is a table of nearby strings:

	Source	Translation	State
59	more	Mehr...	✓
60	less	Weniger...	✓
61	select	Auswählen...	⚠
62	selectedFiles	Ausgewählte Dateien	✓
63			⚠

On the right side, there are three panels: 'Things to check' (Trailing stop), 'Glossary' (no related strings), and 'Source information' (Screenshot context, Context, Flags, Source string age, Translation file).

Global Dataverse Community Consortium

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.



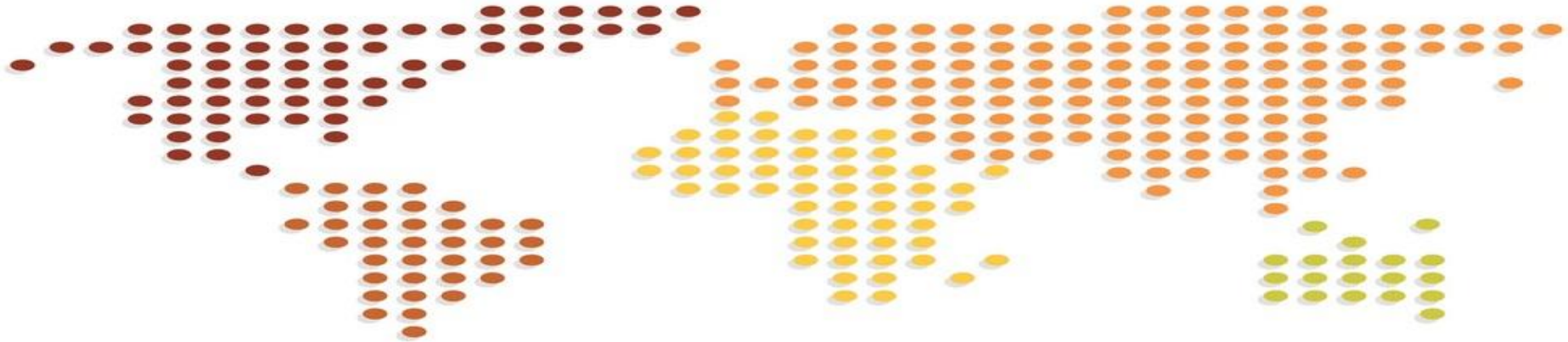
<http://DataverseCommunity.Global>



GDCC

The Global Dataverse Community Consortium
Supporting Dataverse repositories around the world.

- Home**
- About ▼
- Members
- Interest Groups
- Services
- Sign-Up Forms ▼
- News
- Events



Global Dataverse Community Consortium

Australian Data Archive

Consortio Madrono

DANS

DataverseNO

Fudan University

Gottingen eResearch Alliance

Harvard University

International Centre for Research in Agroforestry

Johns Hopkins University

Nanyang Technological University

Syracuse University

Texas Digital Library

University of California Los Angeles

University of Campinas

University of North Carolina Chapel Hill

University of Virginia

Australia

Spain

Netherlands

Norway

China

Germany

United States

Kenya

United States

Singapore

United States

United States

United States

Brazil

United States

United States

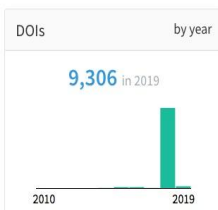
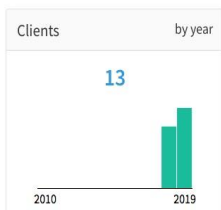
Membership Expanding

Initial Services

DataCite DOI Fabrica

The Global Dataverse Community Consortium (GDCC)

[Info](#) [Settings](#) [Clients](#) [Prefixes](#) [DOIs](#)



Welcome The Global Dataverse Community Consortium (GDCC) to the DOI Fabrica administration area.

The screenshot shows the GitHub profile page for the Global Dataverse Community Consortium. The page header includes a search bar and navigation links for Pull requests, Issues, Marketplace, and Explore. The profile name is "Global Dataverse Community Consortium" with a URL of http://DataverseCommunity.global and an email of Jonathan.Crabtree@unc.edu. Below the profile name, there are statistics for Repositories (1), People (6), Teams (0), Projects (0), and Settings. A search bar for repositories is present, along with a "Type: All" dropdown and a "New" button. The main content area displays a repository named "dataverse-language-packs" with a description "Repository for language files associated with Dataverse", 3 stars, 4 forks, and a note that it was updated 21 days ago. On the right side, there is a "People" section showing 6 members with their profile pictures and an "Invite someone" button. The footer contains copyright information for GitHub, Inc. and various links like Terms, Privacy, Security, Status, Help, Contact GitHub, Pricing, API, Training, Blog, and About.

New Potential Services?

Collaborative Code Development

Shared Programming Staff

Joint Documentation Initiative

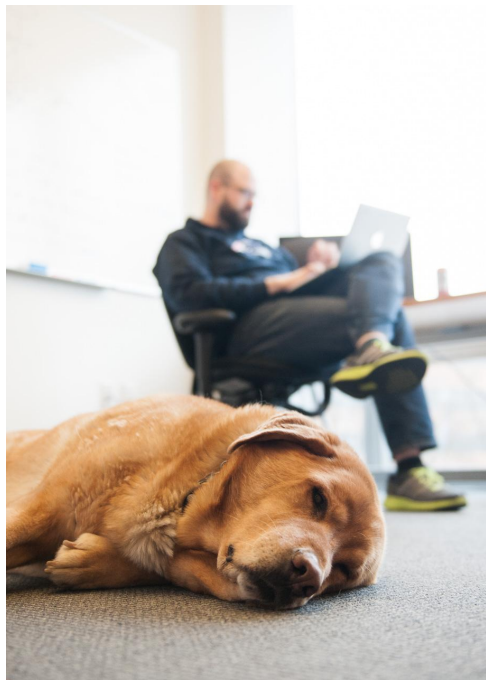
Collaborative Code Testing

Joint Funding Applications

Shared Community Policies

??????????

Thank you!



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>