



# The Evolution of Dataverse & Community Panel IASSIST & CARTO 2024

Thursday, May 30th, 2024, 11:30-12:30am

# Panelists



**Stefano Iacus**  
*Director of Data Science &  
Product Research, Managing  
Director Dataverse Project*



**Sonia Barbosa**  
*Associate Director, Dataverse, Data  
Curation, and The Murray Archive*



**Amber Leahey**  
*Service Director, Borealis (Canada)*



**Gustavo Durand**  
*Dataverse Technical Lead &  
Architect*



**Ceilyn Boyd**  
*Dataverse Development Project  
Manager*

# Agenda

1. Panelists Introduction
2. Dataverse Project History and Timeline(s)
3. Growth of the Global Dataverse Community
4. Dataverse Project Today & Tomorrow
5. Community Survey 2024: Results & Analysis
6. Panel Discussion & Audience Q&A

# Dataverse Project History

# Dataverse Project

## About the Dataverse Project

- An open-source repository 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)
- Established in 1997, Dataverse Network Launch in 2006
- Development funded by IQSS and with grants, in collaboration with institutions around the world
- 2022 - present: Re-Architecture
- Core team
  - @ IQSS - developers, UX/UI, metadata specialists, curation team, leadership team
  - Key contributors from the community with full privileges as IQSS team

## Dataverse Project Goals

- Grow the Dataverse community
- Empower the open source community to explore and implement new Dataverse applications, tools, and services
- Develop the capability to handle sensitive data and big data
- Expand data and metadata features for existing and new disciplines
- Expand archival and preservation features
- Increase interoperability through the implementation of standards
- Increase contributions from the open source development community
- Improve the Dataverse user experience

# Dataverse Project Timeline

**Founder & Principal Investigator:** Gary King, *Albert J. Weatherhead III University Professor* at Harvard University

## Milestones

- **1997: Project Begins; Focus on preserving and sharing Social Sciences data.** Harvard-MIT Virtual Data Center collaboration, Murray Archive data
- **2006: The Dataverse Network 1.0, Harvard Dataverse Repository was established.**
  - 2007: King, G. (2007). An Introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods & Research*, 36(2), 173–199.
  - 2007: Altman and King (2007). A Proposed Standard for the Scholarly Citation of Quantitative Data, *D-Lib Magazine*, 13, 3/4(March/April, 2007).
- 2008: ODUM Archive installation First installation of Datavers outside of Harvard
- **2015: First Dataverse Community Meeting held at IQSS Harvard**
- **2015: Data Citations.**
  - 2015: Altman, Borgman, Crosas & Matone, M. (2015). An introduction to the joint principles for data citation. *Bulletin of the Association for Information Science and Technology*, 41(3), 43–45.
- **2016: Dataverse 4.0, FAIR data.**
  - Dataverse 4.0 - [sample of features]
  - Wilkinson, et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3.
- **2023: 🎉 100th Dataverse installation, Dataverse 6.0, Dataverse Rearchitecture.**
- **2024: The Next Frontier: Large data sensitive data, and more.**

# Dataverse Project Timeline

**Dr. Gary King**

Founder & Principal Investigator

*Albert J. Weatherhead III University Professor*

Harvard University

**1997**

Project Begins

Focus on  
preserving &  
sharing social  
sciences data

Harvard-MIT  
Data Center  
collaboration

Murray Data  
Archive

**2006**

The Dataverse  
Network 1.0

Harvard  
Dataverse  
Repository  
established

**2007**

 Announcing  
the Dataverse  
Network

**2008**

Odum Archive  
becomes first  
Dataverse  
outside Harvard

**2015**

Data Citations

 First Open  
Dataverse  
Community  
Meeting  
Dataverse 4.0

**2016**

FAIR data

**2023**

 100th  
Dataverse  
installation

Dataverse 6.0  
Dataverse Re-  
architecture

 AI comes to  
Harvard  
Dataverse

**2024**

 The Next  
Frontier!  
Large data  
Computing on  
data

King, G. (2007). An Introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods & Research*, 36(2), 173–199.

Altman & King (2007). A Proposed Standard for the Scholarly Citation of Quantitative Data, *D-Lib Magazine*, 13, 3/4(March/April, 2007).

Wilkinson, et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3.

**1997**

Project Begins  
Focus on preserving & sharing social sciences data  
Harvard-MIT Data Center collaboration  
Murray Data Archive

**2006**

The Dataverse Network 1.0  
Harvard Dataverse Repository established


**2007**

 Announcing the Dataverse Network

**2008**

Odum Archive becomes first Dataverse outside Harvard

**2015**

Data Citations  
 First Open Dataverse Community Meeting  
Dataverse 4.0

**2016**

FAIR data

**2023**

 100th Dataverse installation  
Dataverse 6.0  
Dataverse Re-architecture  
 AI comes to Harvard Dataverse

**2024**

 The Next Frontier!  
Large data  
Computing on data

Altman, Borgman, Crosas & Matone. (2015). An introduction to the joint principles for data citation. *Bulletin of the Association for Information Science and Technology*, 41(3), 43–45.

# Technology Evolution

- **Prehistory: VDC**
  - written using XSLT (XSL Transformations)
- **History: Dataverse Network**
  - v1.0 - first published version using Java (Glassfish Web Server) for both backend and frontend
  - v3.0 - initial implementation of APIs for accessing DVN data
- **Today: Dataverse**
  - v4.0 - Complete rewrite of backend and frontend
    - Flexibility - Dataverse Collection Structure, Metadata, Permissions
    - APIs and External Tools (added in v4.8.5)
  - v5.0 - Infrastructure upgrade from Glassfish 4 to Payara 5
  - v6.0 - Latest infrastructure upgrade to Payara 6
- **Future**
  - SPA - Frontend written in React and communicating exclusively via APIs with backend
  - Marketplace - for easier configuration of modular components - exporters, metadata blocks, etc.
  - More Integrations and External tools!

V.3.0

Fudan University Datanverse Network  
复旦大学社会科学数据平台

POWERED BY THE **Datanverse Network** PROJECT v. 3.3

创建账户 登录

创建一个新的DATIVERSE  
创建一个Datanverse用于上传您的数据集和创建数据收藏集。

搜索课题 高级搜索 Tips

显示已发布Dataverses

资源类型

- 研究论文
- 学位论文
- 研究报告
- 政策法规
- 专项调查
- 统计年鉴

院系专业

- 社会发展与公共政策学院
- 经济学院
- 国际关系与公共事务学院
- 历史地理研究所
- 专家学者
- 专家

已发布Dataverses

Dataverses: 52 | 课题: 630 | 文件: 954

名称	机构	发布日期	活跃度
DVN课题组	复旦大学社会科学数据研究中心	2015-1-12	■■■■■
人口与经济	复旦大学	2014-12-16	■■■■■
公共管理	复旦大学	2014-12-16	■■■■■
区域人口研究论文	复旦大学	2014-12-16	■■■■■
人口迁移研究论文	复旦大学	2014-12-16	■■■■■
复旦大学能源流向与碳排放因子数据库	复旦大学能源研究中心	2014-12-16	■■■■■

Avahan Datanverse - Harvu x

thedata.harvard.edu/dvn/dv/Avahan/faces/StudyListingPage.xhtml?mode=1&collectionId=3866

BILL & MELINDA GATES foundation avahan INDIA AIDS INITIATIVE

DATASETS

Harvard Datanverse Network >

Avahan Datanverse

POWERED BY THE **Datanverse Network** PROJECT v. 3.3.2

创建账户 Harvard Affiliate Log In

Avahan provided funding and support to targeted HIV prevention programs in the six Indian states with the highest HIV prevalence, and along the nation's major trucking routes. Gathering and using data was critical for all of Avahan's goals - to continuously refine the program and its many moving parts, to inform other HIV prevention efforts including the national prevention program and its direction, to measure impact and to capture best practices. Data in this Datanverse represent the full range of data collected and used by the lead implementing partners, and some evaluation, knowledge building and capacity building partners in Avahan. The range of data encompasses routine program monitoring data, survey data used for monitoring and for evaluation, and special studies to better understand the HIV epidemic in the program areas.

Seeds of Discovery  
Maize · Wheat · DataVerse

MasAgro  
Modernización Sustentable de la Agricultura Nacional

CIMMYT Datanverse Network >

Seeds of Discovery Datanverse

POWERED BY THE **Datanverse Network** PROJECT v. 3.0

创建账户 Log In

UNIMPUTED GBS DERIVED SNPS FOR MAIZE LANDRACE ACCESSIONS REPRESENTED IN THE SEED-MAIZE GWAS PANEL

hdl:11529/10034

Version: 2 - Released: Tue Feb 10 10:30:33 CST 2015

CATALOGING INFORMATION

Data & Analysis Comments Versions

如果您使用这些数据, 请在您的学术参考文献中添加以下引用。为什么? cite?

Hearne, Sarah; Chen, Charles; Buckler, Ed; Mitchell, Sharon, 2014, "Unimputed GBS derived SNPs for maize landrace accessions represented in the Seed-maize GWAS panel", <http://hdl.handle.net/11529/10034> International Maize and Wheat Improvement Center [Dataverse] V2 [Version]

Data Citation

Citation Format Print

Study Global ID hdl:11529/10034

Authors Hearne, Sarah (CIMMYT); Chen, Charles (CIMMYT); Buckler, Ed (Cornell University); Mitchell, Sharon (Institute for Genomic Diversity (Genomic Diversity Facility))

Producer International Maize and Wheat Improvement Center (CIMMYT), CIMMYT

Production Date 2014

Production Place CIMMYT

Funding Agency La Secretaría de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación de México (SAGARPA Mexico)

Distributor International Maize and Wheat Improvement Center (CIMMYT), CIMMYT

Distributor Contact Seed Data Coordinator (CIMMYT), CIMMYT-Seed-data@cgiar.org

Production Date 2014

Deposit Date diciembre 12, 2014

Provenance Seeds of Discovery Datanverse

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

POWERED BY THE **Datanverse Network** PROJECT v. 3.0.2

创建账户 Log In

DVN Archive Datanverse Network

搜索 this Datanverse Network 搜索

Advanced Search Tips

Welcome! This dataverse network is for internal use and supports a specific group of datasets. For more general information, visit the project site or the Harvard Datanverse Network.

Dataverses

Dataverses is a container for research data studies, customized and managed by its owner.

RECENTLY RELEASED DATIVERSES

Dataset	Release Date
Poverty and Inequality	Apr 6, 2015
Hispanic Trends	Aug 1, 2013
California Welfare Laboratory	Nov 29, 2011
Recession Trends	Aug 9, 2011

View More >

Studies

Studies, Files, Downloads

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

RECENTLY RELEASED STUDIES

Dataset	Release Date
Dataset CWL-2016-07 by Stanford Center on Poverty and Inequality	Jul 16, 2016
Dataset HT-2016-06 v0 by Stanford University Center on Poverty and Inequality	Jun 30, 2016
Dataset POWN_2016_06_v0 by Stanford University Center on Poverty and Inequality	Jun 5, 2016
Dataset MGR-2014-04 v0 by Stanford University Center on Poverty and Inequality	May 2, 2014

## Authors can self-archive their studies with **extended cataloging options**

**Cataloging Information**

Show Required and Recommended Fields
  Show All Fields

Required Fields
  Recommended Fields
  A light blue background indicates that a Date format is required. Please
  Add a new row (e.g., when add multiple authors)
  Remove an entry

**Citation Information**

Title \*

Subtitle

Study ID \* hdl:1902.1/11839

Other ID

Other ID Agency

Author \*

Name \*

Affiliation

Producer \*

- Up to 100 metadata fields to describe each study, including:
  - geospatial information
  - data collection and methodology
  - terms of use and conditions

## A rich set of data analysis based on R statistical package

- Download a subset of variables
- Recode a variable
- Apply descriptive statistics or and advanced statistical models (from Zelig/R)

Download Subset | Recode and Case-Subsetting | Descriptive Statistics | **Advanced Statistical Analysis**

Selected Variables

ID  
012  
HH\_CELL

Select variables from table below (selected)

Variable Type  
 Discrete  
 Discrete  
 Discrete  
 Discrete  
 Discrete  
 Discrete

-- Choose a Statistical Model --  
 -- Choose a Statistical Model --  
 Categorical Data Analysis  
 Cross-Tabulation  
 Ecological Inference Models  
 Hierarchical Multinomial-Dirichlet Ecological Inference Model for R x C Tables  
 Event Count Models  
 Negative Binomial Reg for Event Count Dep Vars  
 Poisson Reg for Event Count Dep Vars  
 Models for Continuous Bounded Dependent Variables  
 Exponential Reg for Duration Dep Vars  
 Gamma Reg for Cont. Positive Dep Vars  
 Log-Normal Reg for Duration Dep Vars  
 Weibull Reg for Duration Dep Vars  
 Models for Continuous Dependent Variables  
 Fit an Analysis of Variance Model  
 ANOVA  
 GENDER probability or selection - gender  
 gender of R

## The Dataverse Network expansion outside Harvard/MIT is underway...



ODUM at University of North Carolina (In production)



Woods Hole Oceanographic Institution (In Testing)

ICPSR

Inter-University Consortium for Political and Social Research, University of Michigan (In testing)



Australian Social Science Data Archive (In evaluation)

South Africa National Archives and Records Service (NARS) (In evaluation)

Dataverse Network™

## A dataverse supports **ANY file type**, but offers additional services to quantitative data files

- Author can upload to the study any file type or format of audio, images, video, text, etc
- Additional features are available to subset and analyze quantitative data files (SPSS, STATA files)
- File metadata includes:
  - File name
  - description
  - category

**Cataloging Information** | **Documentation, Data and Analysis**

Access to some files is restricted

Download All Files Note: you will be prompted to save a single archive file.

1. Documentation

00022Colby-Responsibility-StudyDescription.pdf  
Adobe PDF - 10 KB - 0 downloads

2. Data

00022Colby-Responsibility-Data por  
SPSS Portable - 435 KB - 0 downloads

00022Colby-Responsibility-Data tab  
Tab Separated - 305 KB - 0 downloads + analyses

SUBSETTING 94 Cases 1445 Variables

3. Supplementary Documentation

00022Colby-Responsibility-CodebookAddendum.pdf  
Adobe PDF - 9 KB - 0 downloads

4. Detailed Usage Terms

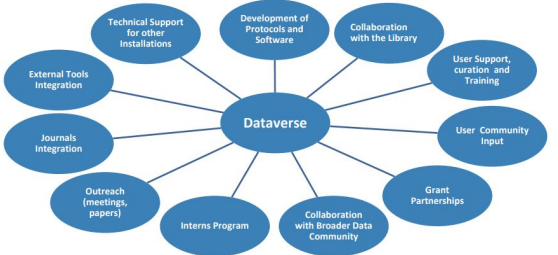
00022Colby-Responsibility-MemoOfAgreement.pdf  
Adobe PDF - 440 KB - 0 downloads

Dataverse Network™

# Dataverse 4.0

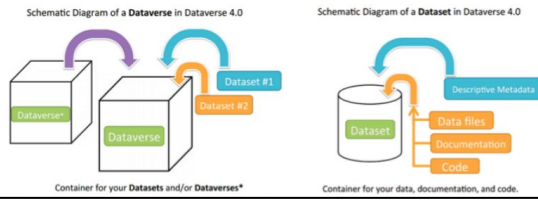
Dataverse 4.0 is completely rewritten and focused on improving usability, extending support to multiple disciplines, enhanced API support, and an improved permissions model. This list of new features and changes is not exhaustive, see the [user guides](#) and [GitHub project](#) for more information.

## Dataverse is now more than a software project and a data repository



## Dataverse 4.0

A full rewrite that improves usability defines a rigorous and standardized data publishing workflow, and leverages the latest technologies.

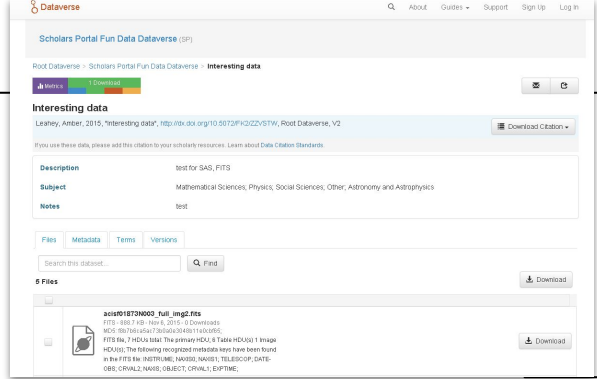


### Rich Set of Features

- Standard, persistent data citation
- Branding for each dataverse
- Standard, extensible metadata:
  - citation metadata
  - domain-specific metadata
  - file-level metadata
- Faceted search for all metadata
- Multiple levels of access control
  - CCO/ terms of use/ restricted
- **Multinla roles and permissions**

### Upgraded Technology

- UI improved by usability testing
- Built with open source solutions
- Enhanced UI framework
  - PrimeFaces and Bootstrap
- Widely used, community driven enterprise software platform
  - Java EE7 and Glassfish
- Reliable, scalable search platform
  - Solr
- **Web standard programmatic interfaces**



## 4.0

-Latest Release, 4.0.1, can be found here: <https://github.com/QSS/dataverse/releases/tag/v4.0.1>

### 4.0 Release Notes

Dataverse 4.0 is completely rewritten and focused on improving usability, extending support to multiple disciplines, enhanced API support, and an improved permissions model. This list of new features and changes is not exhaustive, see the [user guides](#) and [GitHub project](#) for more information.

### New Features:

- Improved add data and edit workflows.
- Improved restrict file and request access workflows.
- Both online and email-based workflow event notification.
- Improved messaging.
- Customizable search facets for each dataverse.
- Greater use of search facets for data discovery.
- Solr server for faster, more inclusive search results.
- Linked search allows adding search-based results to a dataverse.
- Simplified dataverse customization through themes and widgets.
- Featured dataverses allows highlighting selected dataverses.
- Default CCO waiver for datasets.
- Support for custom metadata fields
- New search and native APIs.
- Dataverse administrators can create user groups.
- Dataverse administrators can create nested dataverses.
- Support for Rdata ingest.
- New point and click online analysis tool, TwoRavens.
- Shibboleth single sign on support, currently experimental.

### Changes:

- Faceting, linked search, and sub dataverses replace collections.
- Extensive metadata domains replace controlled vocabulary.
- z39.50 support is ending.
- Widgets replace iframe customization.
- Dataverse widgets built in to OpenScholar
- Dataverse themes replace html header and footer customization.

### Technology Upgrade:

- **PrimeFaces, with Bootstrap**
- **Java 7**
- **Glassfish 4.1**

### Issues:

- Shibboleth support is [experimental](#).
- In some cases deleting a dataset fails.
- Linked datasets, dataverses, and searches are currently available to super users only.

### Installation:

- Please see our installation guide: <http://guides.dataverse.org/en/latest/installation/>

### Upgrading:

- Upgrading from a prior version requires considerable planning and testing. Please contact us for support at [support@dataverse.org](mailto:support@dataverse.org)

TwoRavens

cchs-82M0013-E-2012-Annual-component\_F1

Variable transformation ↺ ↻ Estimate

Data Selection

Variables Subset

- ADL\_01
- ACC\_12G
- ACC\_20
- SDCGCB12
- SDC\_5A\_1
- DHH\_DOWN
- SDCGLHM
- WTS\_M

Model Selection

Models Set Covar. Results

- ls
- logit
- probit
- poisson
- normal
- gamma
- negbinom
- exp
- lognorm
- tobit
- quantile
- logtgee
- probitgee
- zgammagee
- znormalgee
- poissongee

# Integrations over the years...

## Two Ravens

Dataverse

À propos Guides\_fr Soutien

English French Spanish

Dataverse Admin

Root Dataverse > Nouveau dataverse

**Dataverse \***  Dataverse

**Identifiant \***

**Catégorie \*** Sélectionner un/une...

**Courriel \***  +

**Affiliation**

**Dataverse hôte / Héberger Dataverse?** Root

**Description**  Ce champ prend seulement en charge certains/certaines ??

**Champs de métadonnées**

- Sélectionner les champs de métadonnées à utiliser dans les modèles d'ensemble de données et au moment d'ajouter un ensemble de données à ce dataverse.
- Utiliser les champs de métadonnées de Root Dataverse
  - Métadonnées de la référence bibliographique (Obligatoire) [\[+\] Voir les champs](#)
  - Métadonnées géospatiales [\[+\] Voir les champs](#)
  - Métadonnées sur les sciences sociales et les sciences humaines [\[+\] Voir les champs](#)
  - Métadonnées liées aux sciences de la vie [\[+\] Voir les champs](#)
  - Métadonnées liées à la revue [\[+\] Voir les champs](#)

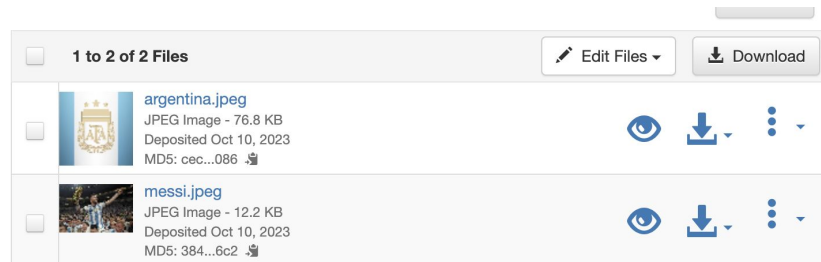
## Dataverse Internationalization

# Dataverse Community Integrations & Tools

- Tools that talk to Dataverse
  - generally used to deposit data into Dataverse (via Deposit API)
  - usually don't require anything special to be set up in the Dataverse repository
- Tools that Dataverse talks to
  - user starts on Dataverse and is directed to the external tool
    - require manifest files
  - have predefined areas in the UI where these would plug into (**Explore** tools)
  - **OR**, are embedded into the Dataverse UI directly (**Preview** tools and **Query** Tools)
- Tools that do both
  - user starts on Dataverse and is directed to the external tool
    - require manifest files
  - also have predefined areas in the UI where these would plug into (**Configure** tools)
  - will also send something back to Dataverse, so need an API token that has "write" privileges

# File Previewers









- A set of tools that display the content of files, allowing them to be viewed without downloading the file, including
- audio
- html
- Hypothes.is annotations
- images
- PDF
- text
- video
- tabular data
- spreadsheets
- GeoJSON
- Zip files
- NcML files
- Previewers are available through the preview (eye) icon on Dataset pages
- And also embedded as a tab on Datafile pages



1 to 2 of 2 Files

Edit Files

Download

<input type="checkbox"/>		<b>argentina.jpeg</b> JPEG Image - 76.8 KB Deposited Oct 10, 2023 MD5: cec...086	  
<input type="checkbox"/>		<b>messi.jpeg</b> JPEG Image - 12.2 KB Deposited Oct 10, 2023 MD5: 384...6c2	  

Preview

Metadata

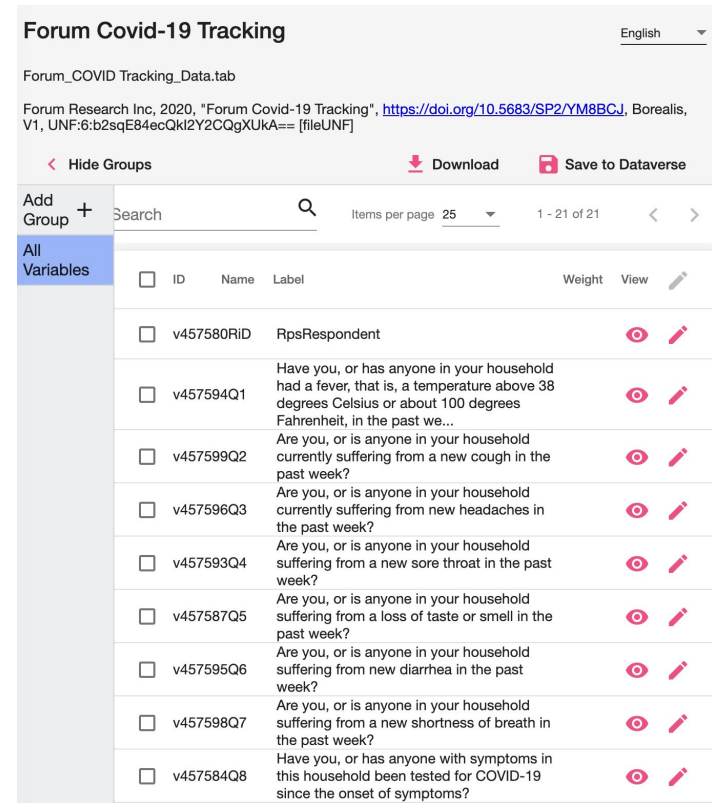
Versions

Explore on View Image



# File Exploration, Configuration, and Query Tools

- File level **explore** tools provide a variety of features from data visualization to statistical analysis
- File level **query** tools allow the user to ask questions (e.g. natural language queries) of a data table's contents without having to download the file
- File level **configure** tools allow (authorized) users to send metadata about the file back to Dataverse

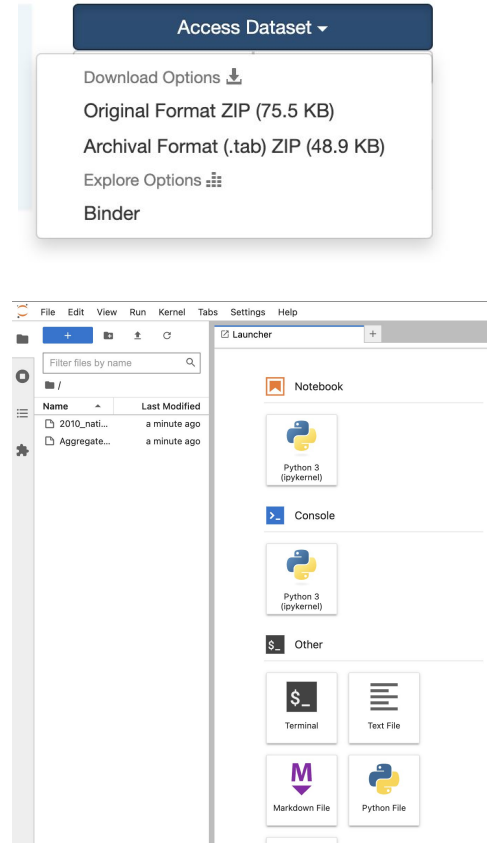


The screenshot displays the 'Forum Covid-19 Tracking' interface. At the top, it shows the title 'Forum Covid-19 Tracking' and the language 'English'. Below this, the file name 'Forum\_COVID Tracking\_Data.tab' is listed, along with a copyright notice for Forum Research Inc. (2020) and a DOI link. Action buttons for 'Hide Groups', 'Download', and 'Save to Dataverse' are visible. The main area shows a table of variables under the 'All Variables' tab. The table has columns for 'ID', 'Name', 'Label', 'Weight', and 'View'. Each row includes a checkbox, an ID, a name, a detailed label, and icons for viewing and editing.

<input type="checkbox"/>	ID	Name	Label	Weight	View	
<input type="checkbox"/>	v457580RID	RpsRespondent				
<input type="checkbox"/>	v457594Q1		Have you, or has anyone in your household had a fever, that is, a temperature above 38 degrees Celsius or about 100 degrees Fahrenheit, in the past we...			
<input type="checkbox"/>	v457599Q2		Are you, or is anyone in your household currently suffering from a new cough in the past week?			
<input type="checkbox"/>	v457596Q3		Are you, or is anyone in your household currently suffering from new headaches in the past week?			
<input type="checkbox"/>	v457593Q4		Are you, or is anyone in your household suffering from a new sore throat in the past week?			
<input type="checkbox"/>	v457587Q5		Are you, or is anyone in your household suffering from a loss of taste or smell in the past week?			
<input type="checkbox"/>	v457595Q6		Are you, or is anyone in your household suffering from new diarrhea in the past week?			
<input type="checkbox"/>	v457598Q7		Are you, or is anyone in your household suffering from a new shortness of breath in the past week?			
<input type="checkbox"/>	v457584Q8		Have you, or has anyone with symptoms in this household been tested for COVID-19 since the onset of symptoms?			

# Dataset External Tools

- Dataset level **explore** tools allow the user to explore all the files in a dataset - common use case is reproducibility
  - **WholeTale** - creates reproducible research packages based on popular tools such as Jupyter and RStudio
  - **Binder** - spins up custom computing environments in the cloud (including Jupyter notebooks)
- Dataset level **configure** tools allow (authorized) users to send metadata about the dataset back to Dataverse
  - **Turbo Curator** - uses Open AI's ChatGPT & ICPSR best practices to provides recommendation to enhance metadata & generate meaningful titles, descriptions, and keywords



# The Global Dataverse Community

2015-Today

# DATAVERSE REPOSITORIES - A WORLD VIEW

118 Installations



Powered by the  
DataVerse Project

**UNR**  
Universidad  
Nacional  
de Rosario

**ARCA** dados  
FIOCRUZ

**World Agroforestry (ICRAF)**

**ICRISAT**  
International Crops  
Research Institute  
for the Semi-Arid  
Tropics

陽明交大  
**NYCU** DataVerse



# Repository of open research data of the National Academy of Sciences of Ukraine

Search datasets in DataverseUA

Search

Go to the repository

Data placement request

Recently published datasets

## Advantages

DataverseUA is a public platform of the Open Data Repository, available to researchers of the institutes of the National Academy of Sciences of Ukraine for the purpose of hosting, sharing and reusing research data. This helps to ensure the multiple use of research data, its availability, the ability to interact with different types of data and the implementation of operational search (FAIR principles).

DataverseUA supports creating custom terms of use and restrictions to control access to your research data. DataverseUA provides guaranteed access to data sets, permanent identifiers and special conditions for saving research data, creating backup copies for long-term storage.

Will post research data on DataverseUA, making it more accessible to the research community on the Internet. You can count on the Dataverse platform to work well with all types of data and provide access to advanced search capabilities, allowing your researchers to find the data they need and extract valuable insights from the data. The Dataverse platform is secure, interoperable, scalable and

<https://opendata.nas.gov.ua/>

DataverseUA

Metrics 2 Downloads

Contact Share

Data repository of the institutes of the National Academy of Sciences of Ukraine.



Search this dataverse...

Advanced Search

Datasets (4)

Files (0)

**Dataverse Category**

Organization or Institution (2)

Department (1)

Research Group (1)

**Publication Year**

2023 (2)

2022 (2)

**Subject**

Chemistry (2)

Physics (2)

1 to 4 of 4 Results

Sort

**UHV ANALYSIS SYSTEM Center** (Frantsevych Institute for Problems of Materials Science of NAS of Ukraine)

Sep 19, 2023 Frantsevych Institute for Problems of Materials Science of NAS of Ukraine

Center for collective use of appliances "High-vacuum analytical system UHV-ANALYSIS-SYSTEM" is designed to conduct research in the field of electronic structure, elemental and phase composition of the surface of solids without their destruction. This center was created on the basis...

**SPM&RS Center** (Kurdyumov Institute for Metal Physics of the NAS of Ukraine)

Sep 19, 2023 G.V. Kurdyumov Institute for Metal Physics of the NAS of Ukraine

The collective use center "Centre of scanning probe microscopy and resonance spectroscopy" (SPM&RS-Centre) was created in 2012 in the structure of the GV Kurdyumov Institute for Metal Physics National Academy of Sciences of Ukraine for the most rational use of unique equipment...

**Frantsevych Institute for Problems of Materials Science of NAS of Ukraine** (National Academy of Sciences of Ukraine)

February 9, 2022

Frantsevych Institute for Problems of Materials Science of NAS of Ukraine is a leading center of scientific and technical developments in the field of theoretical foundations of new materials formation, technology for production and manufacture of products from these materials w...

**GV Kurdyumov Institute for Metal Physics of the NAS of Ukraine** (National Academy of Sciences of Ukraine)

January 7, 2022

GV Kurdyumov Institute for Metal Physics of the NAS of Ukraine (hereinafter, Institute) was founded in 1945. It is one of the largest scientific centers of fundamental research in the field of metal physics in Ukraine and Europe. The scientists of the Institute carry out re...

# Global Research Community...

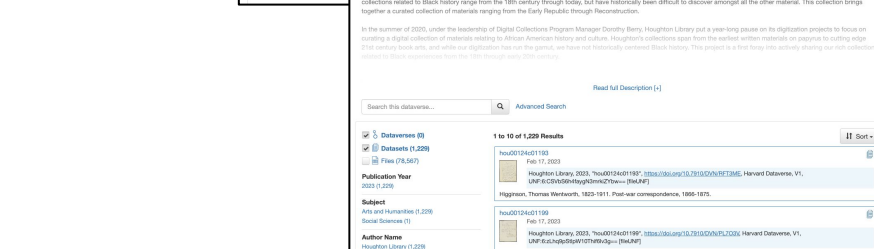
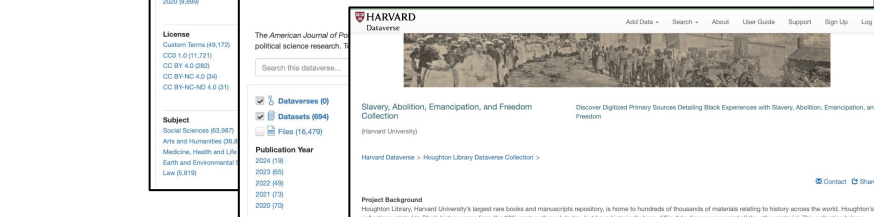
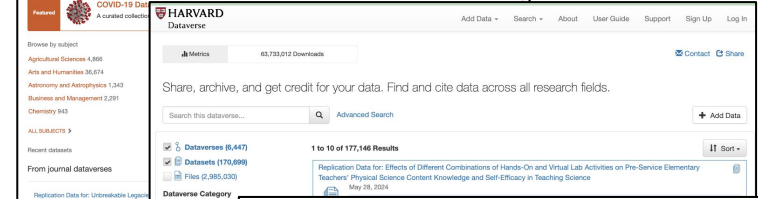
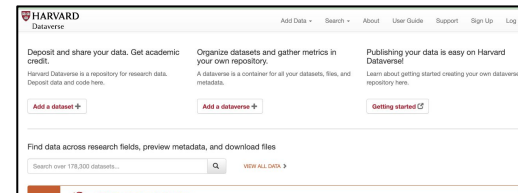
## ANU Poll 47 (August 2021): Mental health and schooling of children during COVID 19

## Impacts of COVID-19 on Canadians, 2020: Perceptions of Safety

## Dental emergency care in Spain during the state of alarm due to COVID-19 pandemic

# Harvard Dataverse (USA)

- Launched in 2006, Open to the world-wide research data sharing community
- Provided by IQSS and supported by Harvard Library, HUIT, and Research Data Computing at Harvard University
- Self-deposit and publishing workflow
- Member of [NIH-GREI](#)
- Member of [GDCC](#)
- Paid curation services
- Large data support
- Collaborations with [Harvard Data Commons](#)





# Borealis (Canada)

- Launched in 2011, Borealis is a national research data repository
- Provided by academic libraries and institutions, Canadian regional library consortia, and the Digital Research Alliance of Canada
- Growth of the data repository and integrations to support wide range of users at many institutions
  - Bilingual installation (English & French)
  - Customized institutional collections & deposit authentication workflows
  - Discovery, data exploration, data curation tools (Odesi.ca, Dataverse Data Explorer & Curation Tools, collections metrics, and more)
  - GDCC member and contributing developer

**POLYTECHNIQUE MONTREAL**  
UNIVERSITÉ D'INGÉNIERIE

Dataverse de Polytechnique Montréal  
(Polytechnique Montréal)

Borealis > Dataverse de Polytechnique Montréal >

**A Worldwide Historical Dam Failure's Database**  
Version 1.1

**ACADIA UNIVERSITY**

Acadia University  
(Acadia University)

Borealis > Acadia University >

**Foliar Endophytic Fungi from Geum peckii in Canada**  
Version 1.0

Sarah Geum  
UNF:6:aq4Tpnq+3jPILZMh22A== [fileUNF]

Cite Dataset

Borealis > Dalhousie University Dataverse > Coastal Hydrology Lab > BantingFellowship >

**Global groundwater warming due to climate change**  
Version 1.0

Benzi, Susanne, 2024, "Global groundwater warming due to climate change", <https://doi.org/10.5683/SP3/G4VEEQ>, Borealis, V1, UNF:6:16aF1rXOVL3kMOD26gbsA== [fileUNF]

Cite Dataset • Learn about [Data Citation Standards](#).

Access Dataset •  
Contact Owner • Share

Dataset Metrics •  
150 Downloads •

Borealis > Cape Breton University Dataverse > Boat Harbour Project >

**Fish Sampling Results: species, location and catch per unit effort**  
Version 6.0

Oakes, Ken, 2020, "Fish Sampling Results: species, location and catch per unit effort", <https://doi.org/10.5683/SP2/PEHA00>, Borealis, V5, UNF:6:aq4Tpnq+3jPILZMh22A== [fileUNF]

Cite Dataset • Learn about [Data Citation Standards](#).

Access Dataset •  
Contact Owner • Share

Dataset Metrics •  
134 Downloads •

**Description** • Final results of fish survey done in Boat Harbour, Sept/ Oct 2019.

**Subject** • Earth and Environmental Sciences

**Keyword** • Fish, Boat Harbour, Nova Scotia

**Related Publication** • Hoover, Z., Adesida, A., Carrier, A.J., Francis, L., Hoover, J., Pham, M.N., Nicholson, A., Panneerselvam, E., Zhang, X., Oakes, K (2020), Boat Harbour Fish Population Assessment, Cape Breton University.

**License/Data Use Agreement** • **CC BY 4.0**

# The State of the Dataverse Community is strong!

Dataverse Community (users, developers, researchers, librarians, data scientists and more)

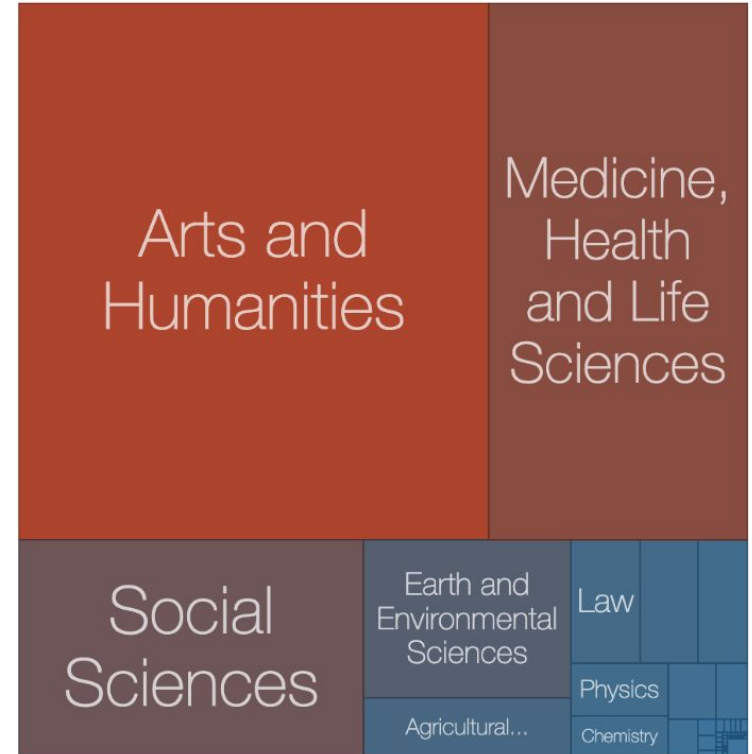
- 118+ installations across the world
- 35 countries
- ~1000 subscribers to the Google Group
- ~100 people in chat.dataverse.org (Zulip)
- 253 “DV Community” Slack user
- 1600+ GitHub issues opened by the community
- 1300+ pull requests made by the community
- 185+ Github contributors to the main code base
- 186 videos on DataverseTV
- 189 community calls since 2016
- Many working groups, interest groups, community groups, meetings, etc.

# The Data ([dataverse.org/metrics](https://dataverse.org/metrics))

- 118 installations
- 16,600 Dataverse Collections\*
- 426,000 Datasets\*
- 6,120,000 Files\*
- 86,500,000 File Downloads\*

\* metrics collected from 96 installations  
(running 4.9 and newer)

Datasets by Most Common Subject

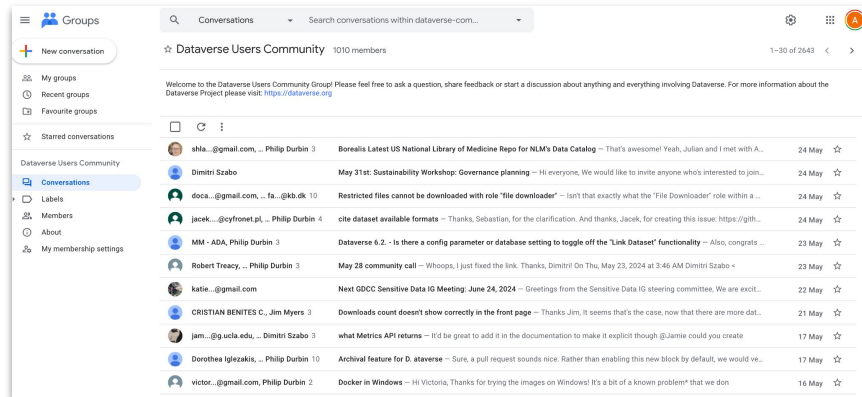


# Dataverse Google Group

## Recent Discussions:

- Technical setup & development questions
- Community calls & meetings
- Integrations & externals tools
- Preservation & FAIR Principles
- Community best practices
- Geospatial data integration
- User & Developer Guides
- Global Dataverse Community Consortium (GDCC)
- Surveys & community feedback

## Dataverse Google Group



<https://groups.google.com/g/dataverse-community>

# Dataverse Community Meetings (2015-2025)

## Annual [Dataverse Community Meeting](#)

- 2015 - IQSS, Harvard: [Common Models and APIs for Data Publishing and Citation Workshop](#)
- 2016 - Harvard Medical School, Harvard: [Fostering the Dataverse Community](#)
- 2017 - IQSS, Harvard: [Dataverse and the Data Lifecycle](#)
- 2018 - IQSS, Harvard: [Dataverse in the Age of Data Science](#)
- 2019 - IQSS, Harvard: Data Sharing: [Supporting Sustainability and Growth](#)

## 1st virtual meetings:

- 2020 - Global Virtual Conference: [Global COVID-19 Data Sharing](#)
- 2021 - Global Virtual Conference: [Automated CI/CD Testing, Installation and Deployment of a Dataverse Installation on a Cloud](#)
- 2022 - Global Virtual Conference: [Indigenous Data Sovereignty](#)

## 1st Regional Community Meetings:

- 2023 - University of Minho, Portugal: [Sharing data for future generations - sustainability, trust, and community building](#)
- 2024 - Texcoco, Mexico (CIMMYT): [Harvesting data from the field to the cloud](#)

# 10 years of community meetings!



[dataverse.org/events](https://dataverse.org/events)

# DCM @CIMMYT in March 2024

- Presentations from IQSS, GDCC, DANS, Borealis, DataverseNO, FIOCrúz, and many Central and South America installations and communities
- CIMMYT Agricultural science community
- Developer meet-up
- PID Workshop



# **Dataverse Community Meeting 2025**

**University of North Carolina at  
Chapel Hill**





The Global Dataverse Community Consortium  
*Supporting Dataverse repositories Around the World*

# Global Dataverse Community Consortium (GDCC) 2018-today

- What is the GDCC?
  - A Community Steering Committee committed to sustainable community development and technical Dataverse support with currently 40 members from 21 countries
- GDCC hosts open Dataverse tools and integration projects in Github
  - File Previewers (e.g. TXT, Markdown, HTML, Images, Videos, Documents, Excel, SPSS, CSV, GeoJSON, Shapefile, GeoTIFF, NetCDF and HDF5, etc.)
  - PyDataverse (e.g. migration scripts, code workflows, etc.)
  - Globus Dataverse tools and more
- In the many WGs and operating as a small organization with ties to the community to support project staff and resources

# GDCC & Community Spaces

## Chairs/Board Members

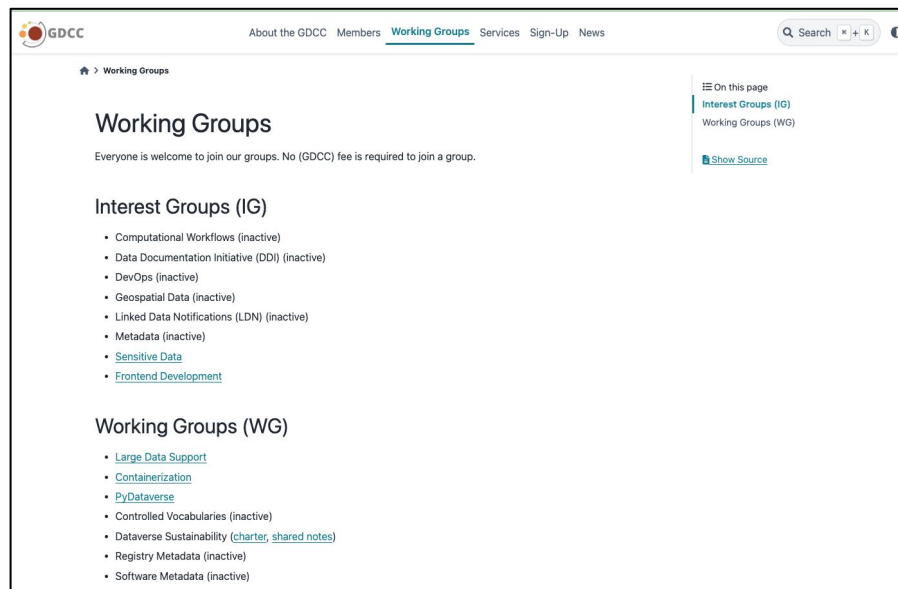
### Steering Committee:

- Philipp Conzett from UiT The Arctic University of Norway (chair)
- Jonathan Crabtree from the University of North Carolina (treasurer)
- Stefano Iacus from the Institute for Quantitative Social Science (IQSS) at Harvard University
- Steven McEachern from the Australian Data Archive

### In addition attending committee meetings:

- Jim Myers, Senior Developer and Architect for GDCC and community contributions
- Ceilyn Boyd, Dataverse Development Project Manager, IQSS, Harvard

## Working Groups

A screenshot of the GDCC website's 'Working Groups' page. The page has a white background with a blue header. The header includes the GDCC logo, navigation links for 'About the GDCC', 'Members', 'Working Groups' (which is highlighted), 'Services', 'Sign-Up', and 'News', and a search bar. Below the header, there is a breadcrumb trail 'Home > Working Groups'. The main heading is 'Working Groups'. A welcome message states: 'Everyone is welcome to join our groups. No (GDCC) fee is required to join a group.' There are three sections: 'Interest Groups (IG)' with a list of groups including 'Computational Workflows (inactive)', 'Data Documentation Initiative (DDI) (inactive)', 'DevOps (inactive)', 'Geospatial Data (inactive)', 'Linked Data Notifications (LDN) (inactive)', 'Metadata (inactive)', 'Sensitive Data', and 'Frontend Development'; 'Working Groups (WG)' with a list including 'Large Data Support', 'Containerization', 'PyDataverse', 'Controlled Vocabularies (inactive)', 'Dataverse Sustainability (charter, shared notes)', 'Registry Metadata (inactive)', and 'Software Metadata (inactive)'. On the right side, there is a sidebar with a search icon and the text 'On this page' followed by 'Interest Groups (IG)' and 'Working Groups (WG)', with a 'Show Source' link below.

<https://www.gdcc.io/working-groups.html>

# Recent New GDCC Members

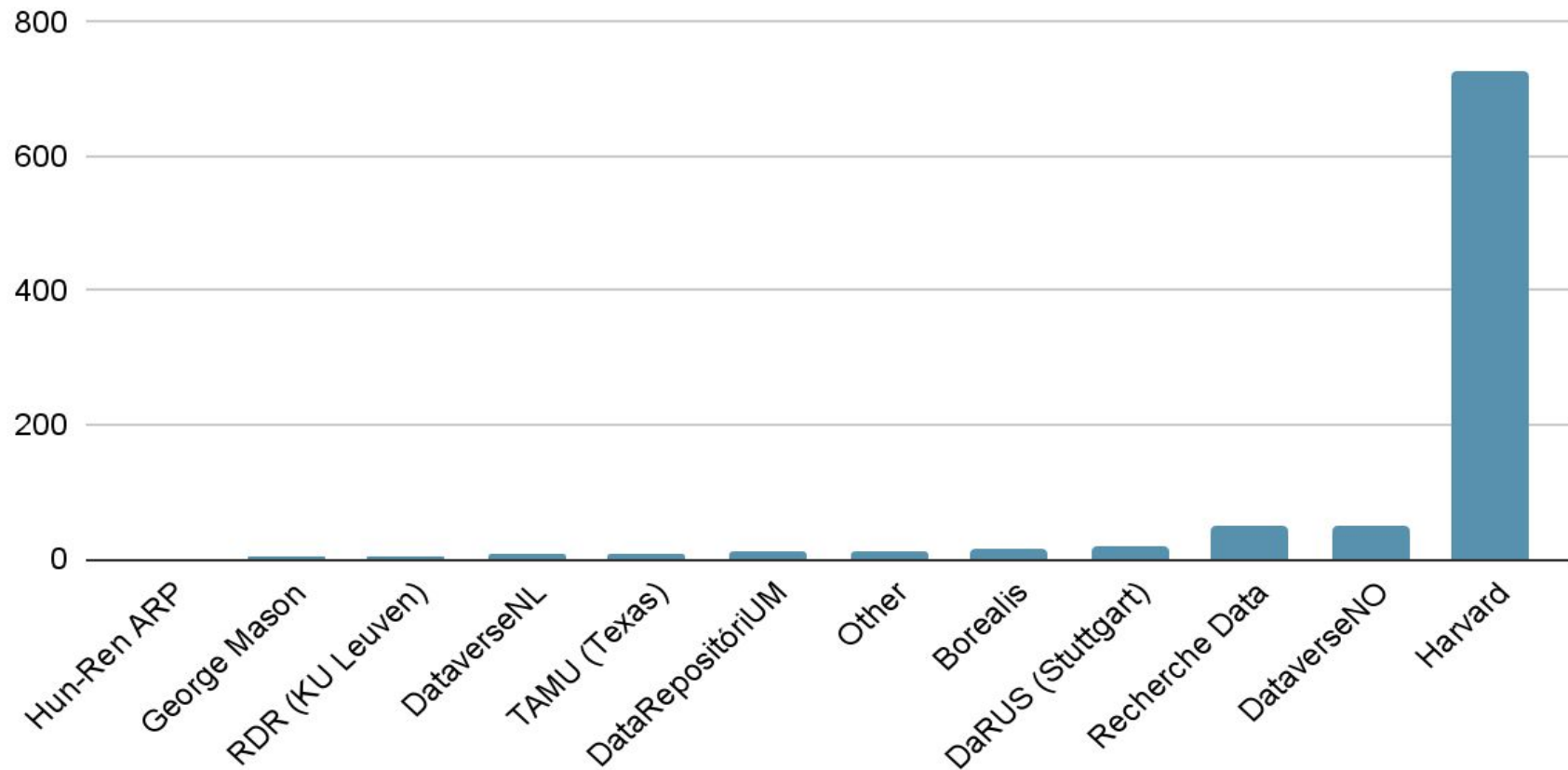
- NIDDK IBD Genetics Consortium. USA
- Agencia Nacional de Investigacion e Innovacion (ANII) Uruguay
- State Research Institution "Kyiv Academic University" Ukraine
- Universidad del Valle de Guatemala

# Community User Survey, 2024 Preliminary Results & Analysis

# Community User Survey 2024

- First sample from Harvard Repository only, early 2024
- Second sample sent across all Dataverse installations, May 2024

# Number of responses



# How likely will you recommend the use of this repository?

0 = Not at all likely ← → 10 = Extremely likely



### Detractors

(score 0-6) are unhappy customers who can damage your brand and impede growth through negative word-of-mouth.



### Passives

(score 7-8) are satisfied but unenthusiastic customers who are vulnerable to competitive offerings.



### Promoters

(score 9-10) are loyal enthusiasts who will keep buying and fuel growth by referring others.



$$\text{Green smiley} \% - \text{Red angry face} \% = \text{NPS}$$

Harvard Dataverse

# How likely will you recommend the use of this repository?

0 = Not at all likely ← → 10 = Extremely likely



### Detractors

(score 0-6) are unhappy customers who can damage your brand and impede growth through negative word-of-mouth.



### Passives

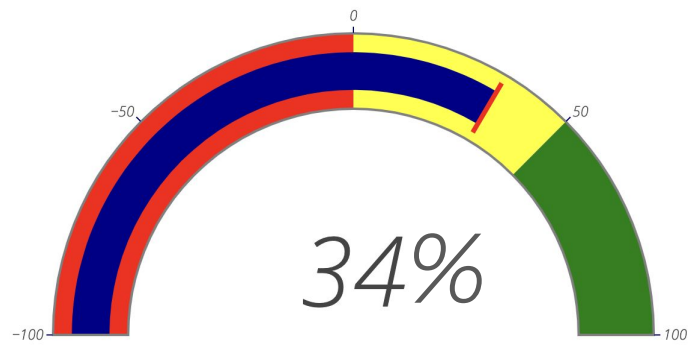
(score 7-8) are satisfied but unenthusiastic customers who are vulnerable to competitive offerings.



### Promoters

(score 9-10) are loyal enthusiasts who will keep buying and fuel growth by referring others.

$$\text{Green smiley} \% - \text{Red angry face} \% = \text{NPS}$$



Harvard Dataverse

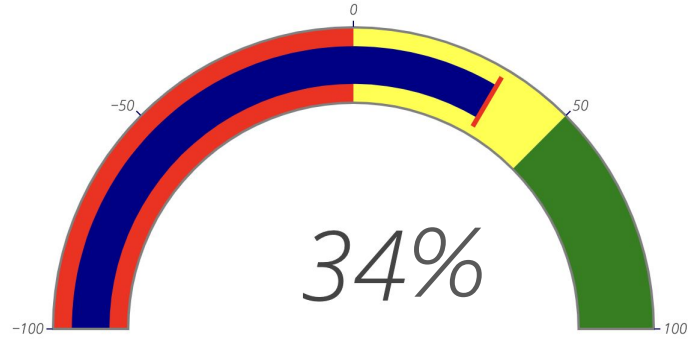
Good or Bad ?

# How likely will you recommend the use of this repository?

0 = Not at all likely    ← →    10 = Extremely likely

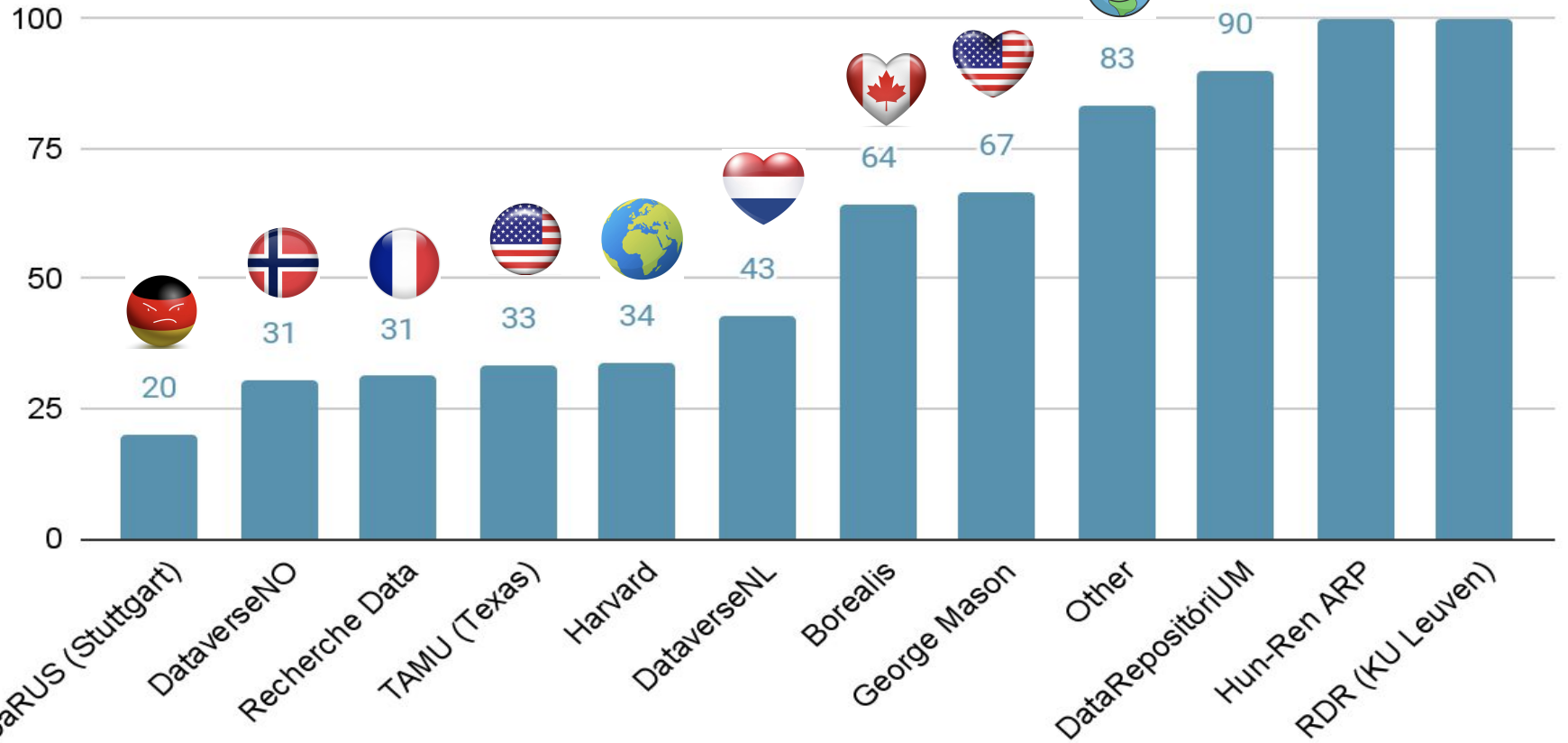
Industry	Professional services	Technology companies	Consumer goods and services
Average NPS	43	35	43
Median NPS	50	40	50

Source: © survey.com

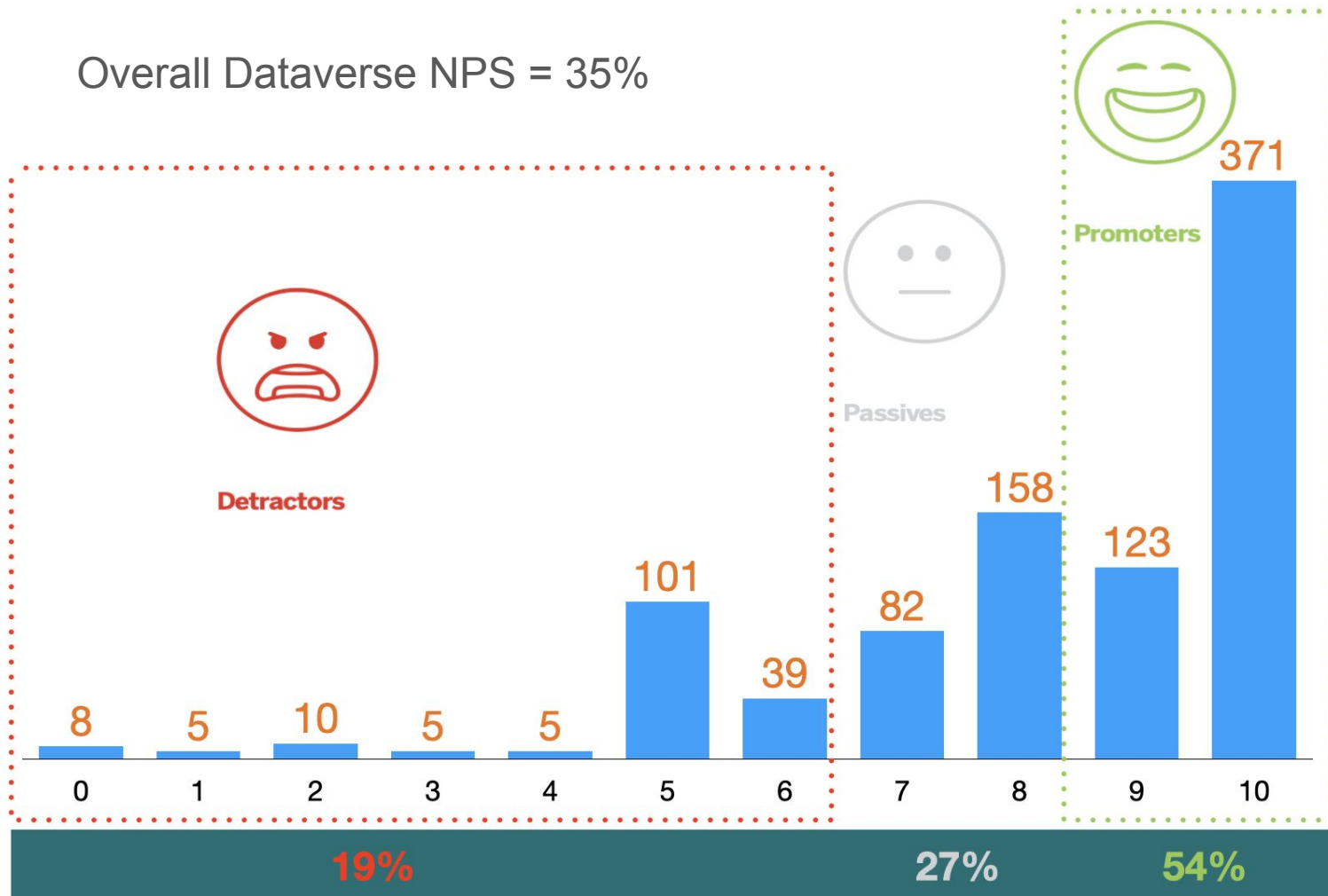


NPS value *per se* is not interesting. We should focus on the 😊 and the 😡 and its value over time.

# NPS



Overall Dataverse NPS = 35%



# Dataverse Project Today & Tomorrow

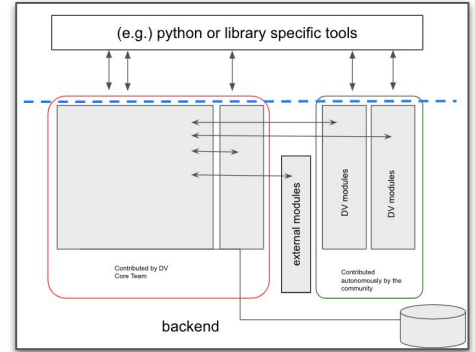
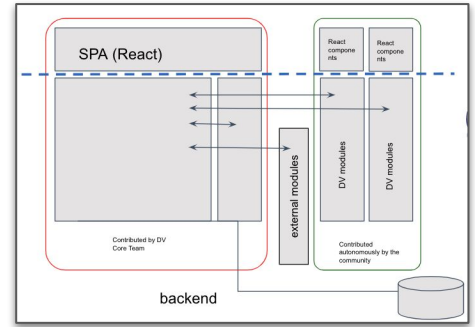
# The re-arch project (last 12 months)

## Goals

- **Modernize** the application
- **Separate backend & frontend**
- Increase **interoperability**
- **Modularize** backend & frontend
- **Speed up development**
- Deploy **new UI/UX ideas**
- **Shorten release time**
- **Empower** the **community**

## Approach

- Single Page Application (**SPA**)
- Native support for **localization**
- Native support for **accessibility**
- **Full automatic testing**
- Improved **modularity**
- **“Plug&play”** module approach
- **Lighter** modules
- Optimization of backend



# Latest Releases (IQSS)

- [6.2](#) New features including search and facet by license, rate limiting, support for using multiple PID providers, and more.
- [6.1](#) Updates include new guestbook at request and collection-level storage quotas features, improved Globus support, support for JSON validation of datasets prior to creation, new or improved API endpoints, new previewers, and improvements to OpenID Connect (OIDC) and Solr.
- [6.0](#) Major platform upgrade release. Payara, Solr, and Java have been upgraded. No features have been added to the Dataverse software itself. Only a handful of bugs were fixed.
- [5.14](#) Improved indexing performance, MDC (Make Data Count) and legacy access count support, Signposting for Dataverse, Permalinks support, Datasets API extended
- [5.13](#) - [5.6](#) Schema.org Improvements, DVWebloader, NetCDF/HDF5 support, Geospatial Search, CodeMeta, Globus, Remote File Storage Configuration, Computational Workflows, Linked Data Notifications (LDN), Terms for Restricted Files, Muting Notifications, Multiple License Support, Embargo feature, External Vocabulary Integration, Support for Double-blind review, Github integration, Migration APIs, other new APIs, bug fixes, system optimizations and upgrades

# Releases since last year

Four releases:

- **5.14** - August 2023
- **6.0** - September 2023
- **6.1** - December 2023
- **6.2** - April 2024

# Dataverse 5.14

- Developed in parallel with 6.0
- Highlights - <https://github.com/IQSS/dataverse/releases/tag/v5.14>
  - **Registering PIDs (DOIs or Handles) for files in select collections**
  - Signposting for Dataverse
  - Permalinks support
  - Mechanism Added for Adding External Exporters
  - Creating datasets with incomplete metadata through API
  - Support for Grouping Dataset Files by Folder and Category Tag
  - Indexing performance improved
  - Changes to PID Provider JVM Settings
    - In preparation for a future feature to use multiple PID providers at the same time
  - New External Tool Type - “query” type tool

# Registering PIDs for files in select collections

- It is now possible to configure registering PIDs for files in individual collections.
  - registration of PIDs for files can be enabled in a specific collection when it is disabled instance-wide
  - Or it can be disabled in specific collections where it is enabled by default
- Available via the Admin API

## :FilePIDsEnabled

Toggles publishing of file-level PIDs for the entire installation. By default this setting is absent and Dataverse Software assumes it to be false. If enabled, the registration will be performed asynchronously (in the background) during publishing of a dataset.

It is possible to override the installation-wide setting for specific collections, see [:AllowEnablingFilePIDsPerCollection](#). For example, registration of PIDs for files can be enabled in a specific collection when it is disabled instance-wide. Or it can be disabled in specific collections where it is enabled by default. See [Change Collection Attributes](#) for details.

To enable file-level PIDs for the entire installation:

```
``curl -X PUT -d 'true' http://localhost:8080/api/admin/settings/:FilePIDsEnabled``
```

If you don't want to register file-based PIDs for your entire installation:

```
``curl -X PUT -d 'false' http://localhost:8080/api/admin/settings/:FilePIDsEnabled``
```

## :AllowEnablingFilePIDsPerCollection

Toggles whether superusers can change the File PIDs policy per collection. By default this setting is absent and Dataverse Software assumes it to be false.

For example, if this setting is true, registration of PIDs for files can be enabled in a specific collection when it is disabled instance-wide. Or it can be disabled in specific collections where it is enabled by default. See [Change Collection Attributes](#) for details.

To enable setting file-level PIDs per collection:

```
``curl -X PUT -d 'true' http://localhost:8080/api/admin/settings/:AllowEnablingFilePIDsPerCollection``
```

When [:AllowEnablingFilePIDsPerCollection](#) is true, setting File PIDs to be enabled/disabled for a given collection can be done via the Native API - see [Change Collection Attributes](#) in the Native API Guide.

# Dataverse 6.0

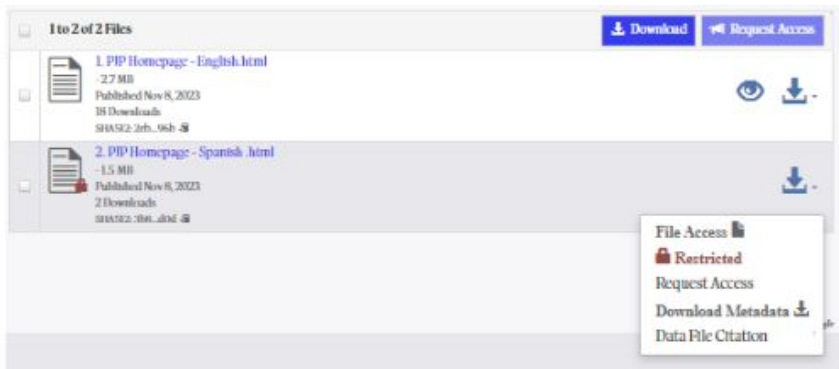
- Functionally the same as 5.14, but featuring major upgrades to core components:
  - The required **Java** version has been increased (from version 11) to **version 17**
  - **Payara** application server has been upgraded to **version 6**
  - **Solr** has been upgraded to **version 9.3.0**
  - **PostgreSQL 13** remains the tested and supported version
- <https://github.com/IQSS/dataverse/releases/tag/v6.0>

# Dataverse 6.1

- First “timed” release
- Highlights - <https://github.com/IQSS/dataverse/releases/tag/v6.1>
  - **Guestbook at request**
  - **Collection-level storage quotas**
  - Globus support (experimental), continued
  - JSON Schema for datasets
  - OpenID Connect (OIDC) improvements
  - Solr improvements
  - New release of Dataverse Previewers (including a Markdown previewer)
  - External tools: configure tools now available at the dataset level

# Guestbook at request

- Dataverse can now be configured to display any configured guestbook to users when they request restricted files
  - The default, showing guestbooks when files are downloaded, remains as it was in prior Dataverse versions
- The global default can be overridden at the collection level and at the individual dataset level



## Data Project Terms

Please confirm and/or complete the information needed below in order to request access to this or other data project.

This data project is made available under the following terms. Please confirm and/or complete the information needed below in order to continue.

I agree to the [License Data Use Agreement](#) on [our Commons page](#) as well as good scientific practices expect that proper credit is given on citation. Please see the data statement down on the data project page.

[QDR](#) [Controlled Access](#)

Name\*

Email\*

Institution\*

Position\*

### REQUIRED Questions

Are you currently a student?\*

If you are a student, where and what are you studying? Please specify your field of scholarship and degree program (i.e., bachelor's, master's, doctoral degree), if applicable.

If you are a student, would this research project be part of a thesis or dissertation project?

PROPOSED RESEARCH TOPIC AND QUESTION (1-3-4 sentences)\*

PROPOSED RESEARCH PLAN: Please outline a brief research plan in the form you intend to submit to your institution's IRB. You can also email a file to [qdr@pij.org](mailto:qdr@pij.org).

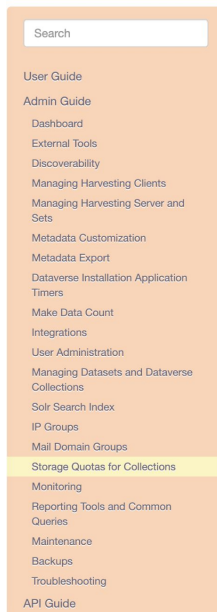
PROPOSED PUBLICATION GOALS: Please indicate your intended research product(s) and potential target journal(s) or other publication venues for your work using PIP data.\*

DATA REQUESTED: Please specify the data to which you are requesting access.\*

If you are requesting access to audio and/or image files, please explain why this will be required for the research you plan to conduct.

# Collection-level storage quotas

- Instance admins can now define storage quota limits for specific collections
  - Storage quotas are inherited by subcollections
  - A storage quota defined on a child sub-collection overrides whatever quota that may be defined on the parent, or inherited from an ancestor
- the users will be informed of the remaining storage allocation on the file upload page
- Still an experimental feature; learn more at <https://guides.dataverse.org/en/6.1/admin/collectionquotas.html>



## Storage Quotas for Collections

Please note that this is a new and still experimental feature (as of Dataverse v6.1 release).

Instance admins can now define storage quota limits for specific collections. These limits can be set, changed and/or deleted via the provided APIs (please see the [Collection Storage Quotas](#) section of the [Native API](#) guide). The Read version of the API is available to the individual collection admins (i.e., a collection owner can check on the quota configured for their collection), but only superusers can set, change or disable storage quotas.

Storage quotas are *inherited* by subcollections. In other words, when storage use limit is set for a specific collection, it applies to all the datasets immediately under it and in its sub-collections, unless different quotas are defined there and so on. Each file added to any dataset in that hierarchy counts for the purposes of the quota limit defined for the top collection. A storage quota defined on a child sub-collection overrides whatever quota that may be defined on the parent, or inherited from an ancestor.

For example, a collection **A** has the storage quota set to 10GB. It has 3 sub-collections, **B**, **C** and **D**. Users can keep uploading files into the datasets anywhere in this hierarchy until the combined size of 10GB is reached between them. However, if an admin has reasons to limit one of the sub-collections, **B** to 3GB only, that quota can be explicitly set there. This both limits the growth of **B** to 3GB, and also *guarantees* that allocation to it. I.e. the contributors to collection **B** will be able to keep adding data until the 3GB limit is reached, even after the parent collection **A** reaches the combined 10GB limit (at which point **A** and all its subcollections except for **B** will become read-only).

We do not yet know whether this is going to be a popular, or needed use case - a child collection quota that is different from the quota it inherits from a parent. It is likely that for many instances it will be sufficient to be able to define quotas for collections and have them apply to all the child objects underneath. We will examine the response to this feature and consider making adjustments to this scheme based on it. We are already considering introducing other types of quotas, such as limits by users or specific storage volumes.

Please note that only the sizes of the main datafiles and the archival tab-delimited format versions, as produced by the ingest process are counted for the purposes of enforcing the limits. Automatically generated "auxiliary" files, such as rescaled image thumbnails and metadata exports for datasets are not.

When quotas are set and enforced, the users will be informed of the remaining storage allocation on the file upload page together with other upload and processing limits.

Part of the new and experimental nature of this feature is that we don't know for the fact yet how well it will function in real life on a very busy production system, despite our best efforts to test it prior to the release. One specific issue is having to update the recorded storage use for every parent collection of the given dataset whenever new files are added. This includes updating the combined size of the root, top collection - which will need to be updated after every file upload. In an unlikely case that this will start causing problems with race conditions and database update conflicts, it is possible to disable these updates (and thus disable the storage quotas feature), by setting the `dataverse.storageuse.disable-storageuse-increments` JVM setting to true.

# Dataverse 6.2

- Highlights - <https://github.com/IQSS/dataverse/releases/tag/v6.2>
  - Search and Facet by License
  - Support for Using Multiple PID Providers
  - **Rate Limiting**
  - Simplified SMTP Configuration
  - Harvesting Handle Missing Controlled Values
  - Ingested Tabular Data Files Can Be Stored Without the Variable Name Header
  - Uningest/Reingest Options Available in the File Page Edit Menu
  - New Accounts Metrics API
  - Several Existing APIs extended

# Rate limiting

- The option to rate limit has been added to prevent users from over taxing the system either deliberately or by runaway automated processes
- Rate limiting can be configured on a tier level
  - tier 0 reserved for guest users
  - tiers 1+ for authenticated users
  - Superuser accounts are exempt from rate limiting
- Still an experimental feature; learn more at <https://guides.dataverse.org/en/6.2/installation/config.html#rate-limiting>



## Rate Limiting

Rate limiting has been added to prevent users from over taxing the system either deliberately or by runaway automated processes. Rate limiting can be configured on a tier level with tier 0 being reserved for guest users and tiers 1-any for authenticated users. Superuser accounts are exempt from rate limiting. Rate limits can be imposed on command APIs by configuring the tier, the command, and the hourly limit in the database. Two database settings configure the rate limiting. Note: If either of these settings exist in the database rate limiting will be enabled (note that a Payara restart is required for the setting to take effect). If neither setting exists rate limiting is disabled.

- `:RateLimitingDefaultCapacityTiers` is the number of calls allowed per hour if the specific command is not configured. The values represent the number of calls per hour per user for tiers 0,1,... A value of -1 can be used to signify no rate limit. Tiers not specified in this setting will default to -1 (No Limit). I.e., -d "10000" is equivalent to -d "10000,-1,..."

```
curl http://localhost:8080/api/admin/settings/:RateLimitingDefaultCapacityTiers -X PUT -d '10000,20000'
```

- `:RateLimitingCapacityByTierAndAction` is a JSON object specifying the rate by tier and a list of actions (commands). This allows for more control over the rate limit of individual API command calls. In the following example, calls made by a guest user (tier 0) for API `GetLatestPublishedDatasetVersionCommand` is further limited to only 10 calls per hour, while an authenticated user (tier 1) will be able to make 30 calls per hour to the same API.

**rate-limit-actions.json** Example JSON for `RateLimitingCapacityByTierAndAction`

```
curl http://localhost:8080/api/admin/settings/:RateLimitingCapacityByTierAndAction -X PUT -d '{"tier": 0, "limitPerHour": 10, "actions": [{"GetLatestPublishedDatasetVersionCommand", "GetPrivateUrlCommand", "GetDatasetCommand", "GetLatestAccessibleDatasetVersionCommand"}], {"tier": 0, "limitPerHour": 1, "actions": [{"CreateGuestbookResponseCommand", "UpdateDatasetVersionCommand", "DestroyDatasetCommand", "DeleteDataFileCommand", "FinalizeDatasetPublicationCommand", "PublishDatasetCommand"}]}, {"tier": 1, "limitPerHour": 30, "actions": [{"CreateGuestbookResponseCommand", "GetLatestPublishedDatasetVersionCommand", "GetPrivateUrlCommand", "GetDatasetCommand", "GetLatestAccessibleDatasetVersionCommand", "UpdateDatasetVersionCommand", "DestroyDatasetCommand", "DeleteDataFileCommand", "FinalizeDatasetPublicationCommand", "PublishDatasetCommand"}]}'
```

# Recent External Tools

- New File Previewers
  - Rich HTML previewer
  - RO-Crate
  - Shape files/geotiff
  - NcML
  - H5 Web
- Updated Globus Transfer Tool
- “Query” type tools for tabular data files
  - Ask the Data
- Dataset level “configure” tools
  - Turbo Curator

# The present future ...



- New Single Page Application (SPA)
- Large Data Support/ Globus Transfer
- Remote Storage & Sensitive Data
- Big Data & Computation
- AI-powered tools
- Harvard and Community research projects and integrations
- Indigenous data support - [TK Labels](#)

# Upcoming Release Plans

- Timed Release Plan; every 3 months
- Next release **6.3** upcoming, **mid-June 2024**
  - Includes Globus and SPA updates
- **Beyond**
  - Continue SPA work for eventual retirement of current UI
  - Marketplace - for easier configuration of modular components - exporters, metadata blocks, etc.
  - More Integrations and External tools!

# Recent and Upcoming Release Plans

- In **December 2023**, we released **6.1**
  - First release under the new Timed Release Plan
- **6.2** was released on **2 April 2024**
  - In conjunction with the **SPA Beta**
  - <https://github.com/IQSS/dataverse/issues?q=milestone:6.2>
- **6.3** upcoming, **mid-June 2024**
  - Includes Globus and SPA updates
- **Beyond**
  - Future releases every 3 months
  - Continue SPA work for eventual retirement of current UI

# Generalist Repository Ecosystem Initiative (GREI)

# NIH Generalist Repository Ecosystem Initiative (GREI)

The mission of GREI is to establish a common set of capabilities, services, metrics, and social infrastructure; raise general awareness and facilitate researchers to adopt FAIR principles to better share and reuse data.

This initiative will further enhance the biomedical data ecosystem and help researchers find and share data from NIH-funded studies in generalist repositories.

## Goals of the Generalist Repository Ecosystem Initiative



1

Make it easier for researchers to **share data**.



2

Enable the improved **discoverability** of NIH-funded data across generalist repositories.



3

Support greater **reproducibility** of NIH-funded research by ensuring data associated with publications is readily available.



4

**Avoid duplication** of the data across repositories.



5

Encourage NIH-funded researchers to be both contributors and consumers to **increase the reuse** of data.



# NIH GREI (2022-ongoing)

## NIH GREI (Generalist Repository Ecosystem Initiative)

- Current programs and activities involved:
  - Remote Large Storage Support
  - Controlled Vocabularies for Biomedical
  - Discovery for Metadata (e.g. Data Documentation Initiative (DDI))
  - Computational Workflows
  - Harvesting and Sharing Metadata Across Repositories
  - Usage Metrics - Make Data Count Support
  - Improvements for Sensitive Data Support
  - Evaluation and Evolution of Architecture
  - NIH Data Management Plans
  - Community Repository Training
  - Support for Research Objects (beyond datasets)

## Participating Repositories



# NIH Desirable Characteristics of Data Repositories

## For Federally Funded Research

- Unique Persistent Identifiers
- Long-Term Sustainability
- Metadata
- Curation and Quality Assurance
- Free and Easy Access
- Broad and Measured Reuse
- Clear User Guidance
- Security and Integrity
- Confidentiality
- Common Format
- Provenance
- Retention Policy

Guidance set forth by NIH  
and by The National Science  
and Technology Council, cited  
in OSTP guidance





## Current Features



**Installations**



**Collections**



**Datasets**



**Files**

## NIH OTA Adds...



### Datasets

- New Workflows for **Large Data Support** ([Globus](#), [demo](#))
- **Remote** Data Support ([TRSA](#))
- Flexible **Biomedical Metadata** Support through External Vocabularies and Data Dictionaries (UMLS, MeSH)
- Investigate CEDAR integration



### Files

- Additional Metadata for **Code files** ([Codemeta](#))
  - Support for **Replication Packages** through Workflows and Containers
  - **Sensitive Data Support** through Differential Privacy ([OpenDP](#), [DataTags](#), [PSIprivacy](#))
  - Encryption
- |                      |   |
|----------------------|---|
| • Usage Metrics      | • Metadata Harvesting                   |
| • UX/UI Enhancements | • Curation Services                     |
| • New APIs           | • <a href="#">Training and Outreach</a> |
| • Interoperability   |   |




# Large Data Support and Dataverse

# Deluge of Large Data

2004 - 2023: < 70TB

Gen - Apr '24: + 35 TB (+50%)

Prospect by June '24: +420 TB

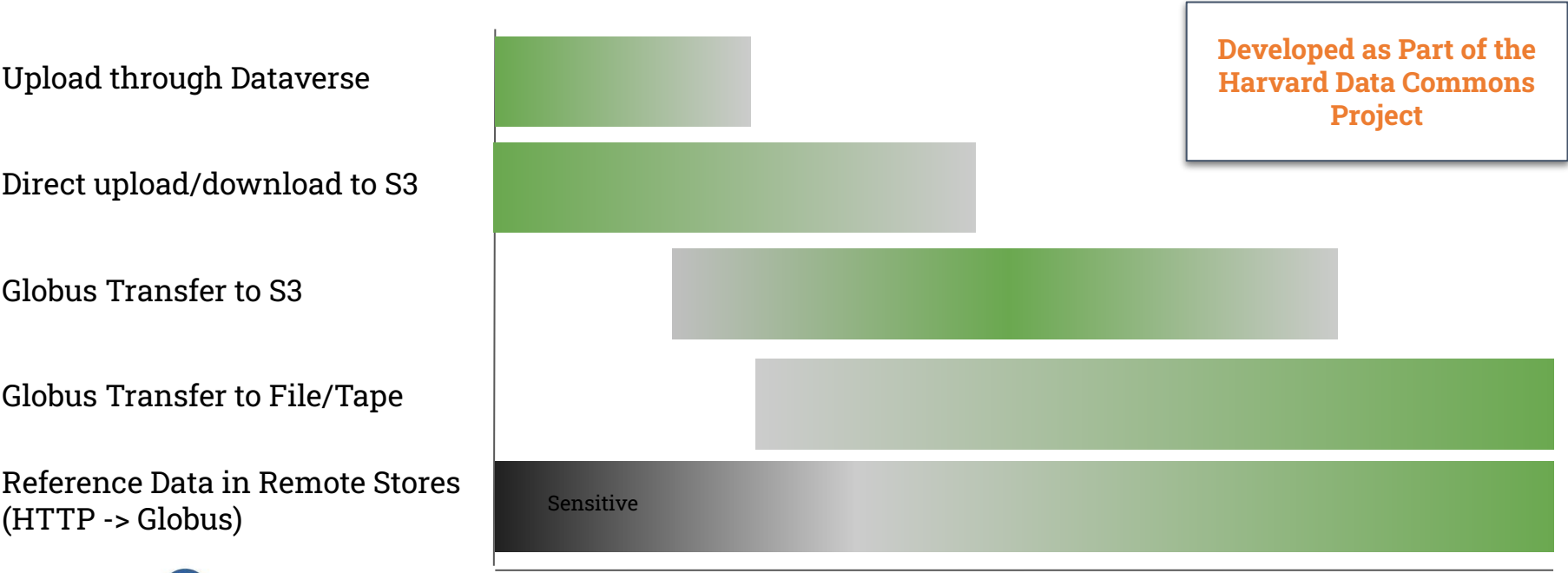
**Sustainability model:**   
data owner contributes to storage costs

Moving to **MOC** (Mass Open Cloud):

- no egress costs
- 30% (storage/computing) cost of AWS



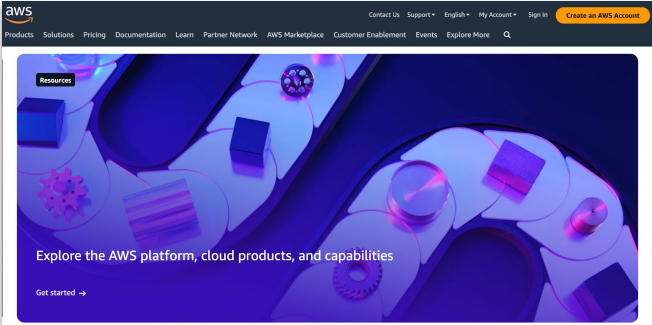
# Harvard Dataverse large data support



GB TB PB

# Large data support infrastructure

Amazon Web Services (AWS) HUIT



**MASSACHUSETTS GREEN HIGH PERFORMANCE COMPUTING CENTER**



**New England Research Cloud**

**Northeast Storage Exchange (NESE)**

Draft dataset of OMAMA project on HDV, large data support on NESE

**Dataverse** Add Data


Collection for testing Globus/NESE storage

(Dataverse.org)

Demo Dataverse > Collection for testing Globus/NESE storage >

## Dataset, NESE Tape Storage Demo


Version 1.0

 Admin, Dataverse, 2024, "Dataset, NESE Tape Storage Demo", <https://doi.org/10.70122/> Dataverse, V1

Cite Dataset - Learn about [Data Citation Standards](#).


**Description** test test test

**Subject** Social Sciences

**License/Data Use Agreement**  CC0 1.0


Files Metadata Terms Versions

1 File

 v7-04-12-geos4-2001-Run0.tar.gz  
Gzip Archive - 1.0 GB  
Published May 17, 2024  
0 Downloads  
MD5: not...rse

## 3D Tomosynthesis + DeepSight Cancer Annotations

**Draft** **Unpublished**

 Haehn, Daniel; Zurrin, Ryan; Goyal, Neha; Bendiksen, Pablo; Manocha, Muskaan; Simovici, Dan; Haspel, Nurit; Pomplun, Marc; Lotter, Bill; Sorensen, Greg, 2024, "3D Tomosynthesis + DeepSight Cancer Annotations", <https://doi.org/10.7910/DVN/E7GHGE>, Harvard Dataverse, DRAFT VERSION

Cite Dataset - Learn about [Data Citation Standards](#).

**Description**

Each image in the dataset is accompanied by a metadata file in JSON format, providing detailed information about the image. Below is an outline of the metadata content:

**PatientID:** A unique identifier for the patient associated with the image.  
**View:** The mammographic view (e.g., crano-caudal) that the image represents.  
**WindowCenter:** An array indicating the center pixel values used for windowing the image, which affects image brightness. Multiple values represent different window settings.  
**WindowWidth:** An array indicating the width of the windowing range, which affects the contrast of the image. Multiple values correspond to the center values for different contrast settings.

WindowCenterWidthExplanation: Descriptions corresponding to the "WindowCenter" and "WindowWidth" values.


[Read full Description \[+\]](#)

**Subject** Medicine, Health and Life Sciences

**Keyword** anomaly detection, outlier detection, mammograms, unsupervised learning, 3D data

**Related Publication** Ryan Zurrin and Neha Goyal and Pablo Bendiksen and Muskaan Manocha and Dan Simovici and Nurit Haspel and Marc Pomplun and Daniel Haehn, 2023. "Outlier Detection for Mammograms." International Conference on Medical Imaging with Deep Learning.

**Notes** For more detailed descriptions and dataset information, please visit: <https://mpsyg.org/omama/>

**License/Data Use Agreement**  CC0 1.0

Files Metadata Terms Versions

**Access Dataset -**

**Publish Dataset**

**Edit Dataset -**

Contact Owner Share

**Dataset Metrics**

0 Downloads

1 File Edit Files -

 **3d.tar**  
TAR Archive - 7.9 TB  
Published Apr 26, 2024  
0 Downloads  
MD5: not...rse

3D Tomosynthesis Dataset: images/ \*.npz - images of various shapes; metadata/ \*.json (includes lesion bounding box and other metadata)

**NESEtape**

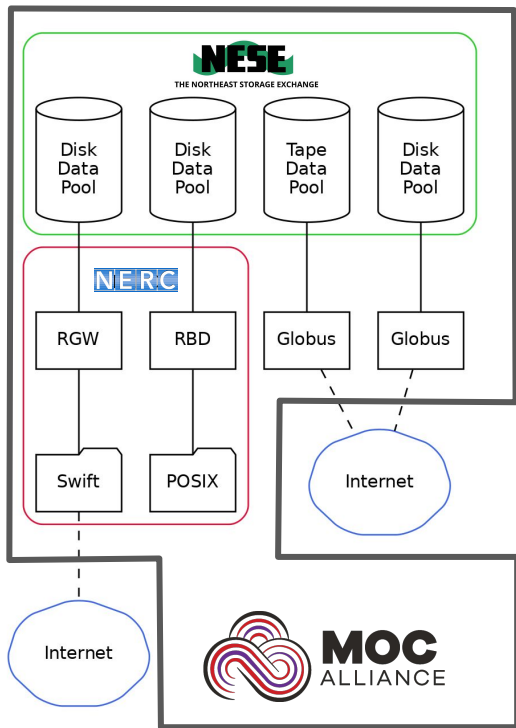
 

# Harvard Dataverse large data services

Category	Storage Options	Description
<b>Consultations</b>	-	Basic and extended consultations to choose best service offering
<b>Data Curation</b>	-	Curation consultations and other services
<b>Basic Service</b>	NESE Tape NESE Disk	Dynamic file access for some files, delayed access for others
<b>On-Demand Service</b>	NESE Disk w/S3 AWS S3	Dynamic file access option using AWS S3 or NESE disk w/S3 mounted containers
<b>Cold Storage</b>	NESE Tape AWS Glacier	Cold storage for rarely accessed data
<b>Custom</b>	Combination	Custom solutions using multiple storage and access options

# Computing on the Data

# Connecting Dataverse to HPC infrastructure: MOC

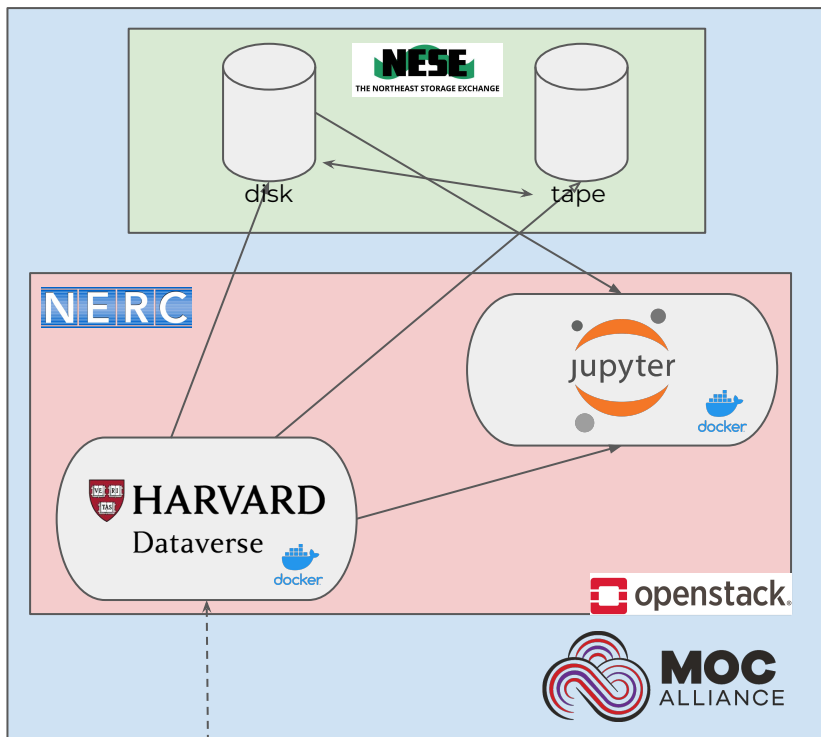


NESE — NERC Data Storage Stack

The screenshot shows the Dataverse web interface for a collection titled "Dataverse MOC Proof of Concept Demo".

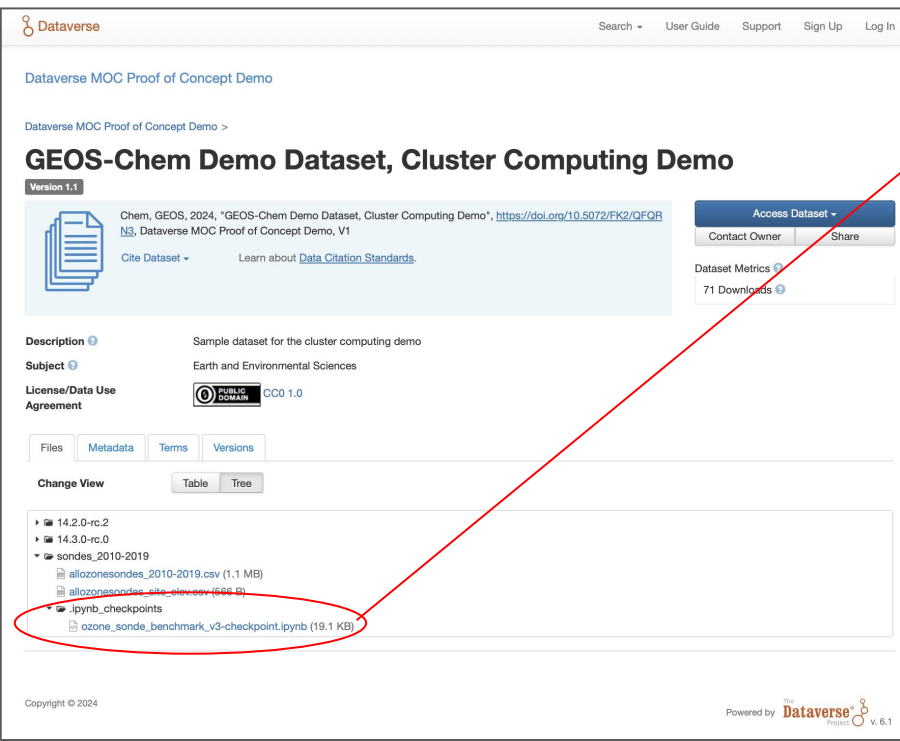
- Header:** Dataverse logo and a hamburger menu icon.
- Collection Title:** Dataverse MOC Proof of Concept Demo
- Metrics:** 71 Downloads. Includes "Contact" and "Share" buttons.
- Description:** This collection is for the proof-of-concept demo of the MOC-Dataverse collaboration on the NERC cluster.
- Search:** A search bar with the text "Search this dataverse..." and an "Advanced Search" link.
- Filters:** Checkmarks for "Dataverses (0)", "Datasets (1)", and "Files (27)".
- Results:** A single result is displayed: "GEOS-Chem Demo Dataset, Cluster Computing Demo" (Feb 12, 2024). The description includes the title, author (Chem, GEOS), and a DOI link: <https://doi.org/10.5072/FK2/QFQRN3>. A note below the result states: "Sample dataset for the cluster computing demo".
- Metadata:** Publication Year: 2024 (1); Author Name: Chem, GEOS (1); Subject: Earth and Environmental Sciences (1); Deposit Date: 2024 (1).

# Connecting Dataverse to HPC infrastructure: MOC

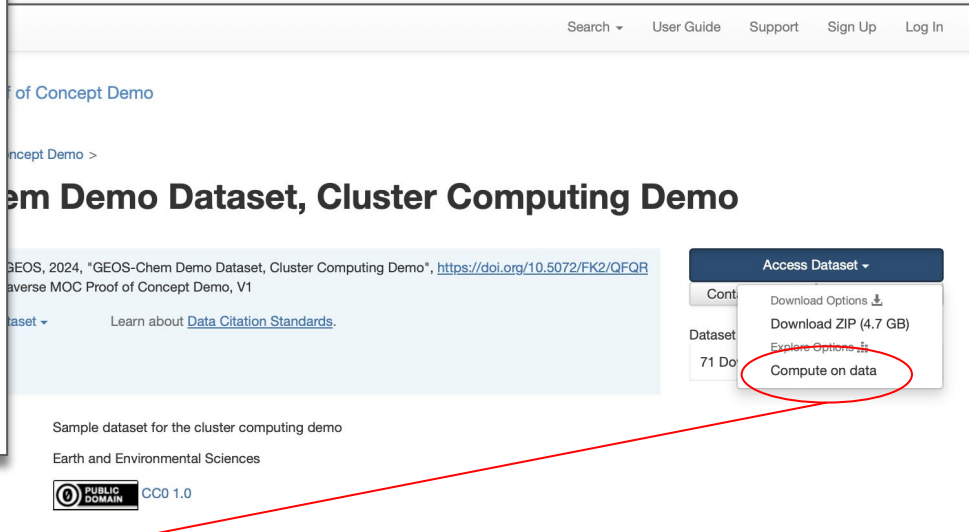


happy Harvard Dataverse user  
(only NERC PI's will be able to run compute)

The screenshot shows the Dataverse interface for the 'Dataverse MOC Proof of Concept Demo'. The page includes a search bar, a metrics section showing '72 Downloads', and a list of two results. The first result is 'GEOS-Chem Dataset, Cluster Computing Demo' by Stefano Iacus, dated Feb 23, 2024. The second result is 'GEOS-Chem Dataset, NESE Tape Storage Demo' by Admin, dated Feb 23, 2024. The page footer includes 'Copyright © 2024' and 'Powered by The Dataverse Project v. 6.1'.



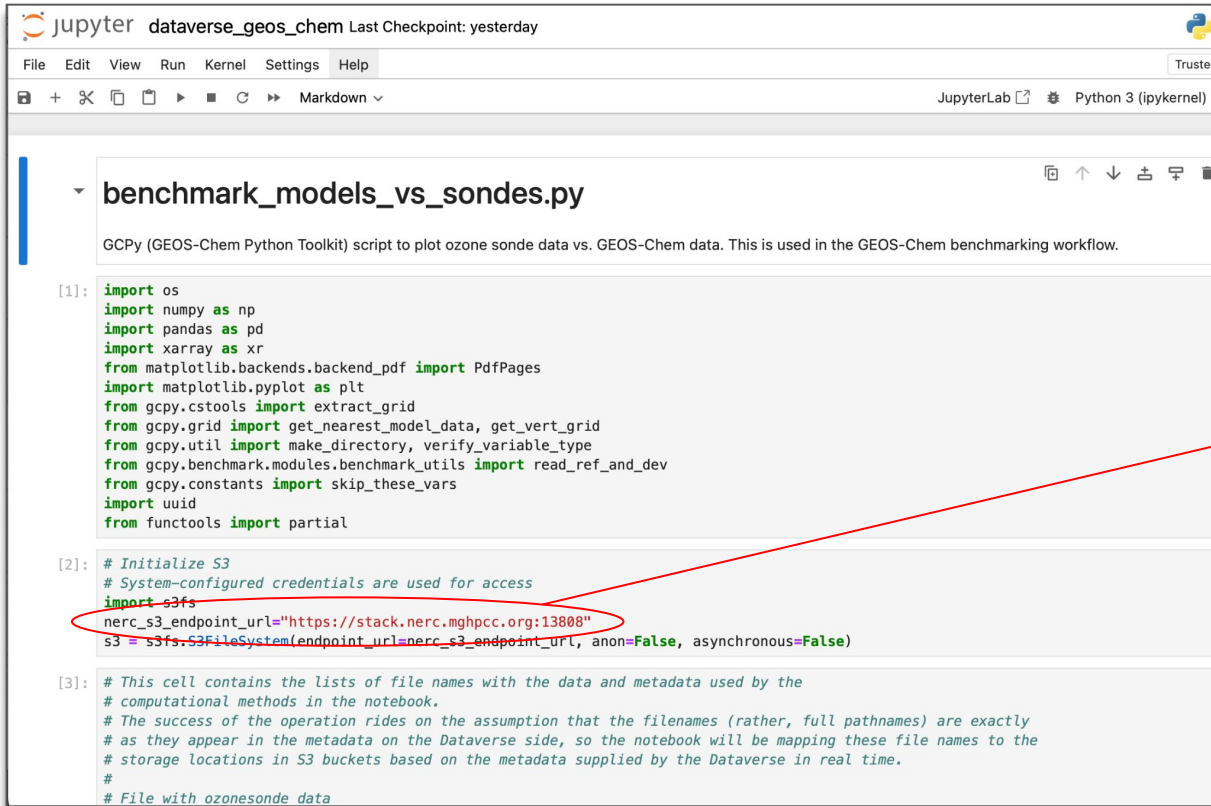
If Dataverse sees a (python) **notebook**, the new menu item “**Compute on data**” appears in the *Access Dataset* drop down menu



This **menu will launch** the JupyterLab VM with the pre-loaded notebook taken from the dataset.

All files in this collection are seen as **local** to the Jupyter instance. Python will simply load them into memory for computing purposes.

# Computing on the data



```

Jupyter dataverse_geos_chem Last Checkpoint: yesterday
File Edit View Run Kernel Settings Help Trusted
+ ✂ 📄 📄 ▶ ■ ↻ ▶▶ Markdown v JupyterLab Python 3 (ipykernel)

benchmark_models_vs_sondes.py
GCPy (GEOS-Chem Python Toolkit) script to plot ozone sonde data vs. GEOS-Chem data. This is used in the GEOS-Chem benchmarking workflow.

[1]: import os
import numpy as np
import pandas as pd
import xarray as xr
from matplotlib.backends.backend_pdf import PdfPages
import matplotlib.pyplot as plt
from gcpy.cstools import extract_grid
from gcpy.grid import get_nearest_model_data, get_vert_grid
from gcpy.util import make_directory, verify_variable_type
from gcpy.benchmark.modules.benchmark_utils import read_ref_and_dev
from gcpy.constants import skip_these_vars
import uuid
from functools import partial

[2]: # Initialize S3
# System-configured credentials are used for access
import s3fs
nerc_s3_endpoint_url="https://stack.nerc.mghpcc.org:13808"
s3 = s3fs.S3FileSystem(endpoint_url=nerc_s3_endpoint_url, anon=False, asynchronous=False)

[3]: # This cell contains the lists of file names with the data and metadata used by the
# computational methods in the notebook.
# The success of the operation rides on the assumption that the filenames (rather, full pathnames) are exactly
# as they appear in the metadata on the Dataverse side, so the notebook will be mapping these file names to the
# storage locations in S3 buckets based on the metadata supplied by the Dataverse in real time.
#
# File with ozonesonde data

```

NERC S3 endpoint for the containerized storage (which exists on NESE)

# Computing on the data

Automatic mapping of local file names (local to the python notebook) to Harvard Dataverse file pointers on NESE

```
mapping file pathnames as listed in the dataset metadata to the direct storage locations:  
sondes_2010-2019/allozonesondes_2010-2019.csv -> s3://mopcpcdemocontainer/10.5072/FK2/QFQRN3/18d8fa95075-22edad4355c8  
sondes_2010-2019/allozonesondes_site_elev.csv -> s3://mopcpcdemocontainer/10.5072/FK2/QFQRN3/18d8fa950f3-ed27dd9b2f0a  
14.3.0-rc.0/GCCClassic/FullChem/OutputDir/GEOSChem.SpeciesConc.20190101_0000z.nc4 -> s3://mopcpcdemocontainer/10.5072/FK2/QFQRN3/18d8  
fa81861-a21e0a10e8a5  
14.2.0-rc.2/GCCClassic/FullChem/OutputDir/GEOSChem.SpeciesConc.20190101_0000z.nc4 -> s3://mopcpcdemocontainer/10.5072/FK2/QFQRN3/18d8  
fa4ee6e-12fc2fa05bf8  
14.2.0-rc.2/GCCClassic/FullChem/OutputDir/GEOSChem.SpeciesConc.20190201_0000z.nc4 -> s3://mopcpcdemocontainer/10.5072/FK2/QFQRN3/18d8  
fa4cd02-4a684a55b7dc  
14.3.0-rc.0/GCCClassic/FullChem/OutputDir/GEOSChem.SpeciesConc.20190201_0000z.nc4 -> s3://mopcpcdemocontainer/10.5072/FK2/QFQRN3/18d8  
fa7ee17-5242cdf4a74  
14.2.0-rc.2/GCCClassic/FullChem/OutputDir/GEOSChem.SpeciesConc.20190301_0000z.nc4 -> s3://mopcpcdemocontainer/10.5072/FK2/QFQRN3/18d8  
fa53f9b-7d6cae9694b7  
14.3.0-rc.0/GCCClassic/FullChem/OutputDir/GEOSChem.SpeciesConc.20190301_0000z.nc4 -> s3://mopcpcdemocontainer/10.5072/FK2/QFQRN3/18d8  
fa85f90-048bd360915f  
14.2.0-rc.2/GCCClassic/FullChem/OutputDir/GEOSChem.SpeciesConc.20190401_0000z.nc4 -> s3://mopcpcdemocontainer/10.5072/FK2/QFQRN3/18d8  
fa4f6fe-b811828db8aa  
14.3.0-rc.0/GCCClassic/FullChem/OutputDir/GEOSChem.SpeciesConc.20190401_0000z.nc4 -> s3://mopcpcdemocontainer/10.5072/FK2/QFQRN3/18d8,
```

An upcoming version of **pyDataverse** will automatically generate these mappings for general cloud systems.

# Computing on the data

jupyter dataverse\_geos\_chem Last Checkpoint: yesterday

File Edit View Run Kernel Settings Help Trusted

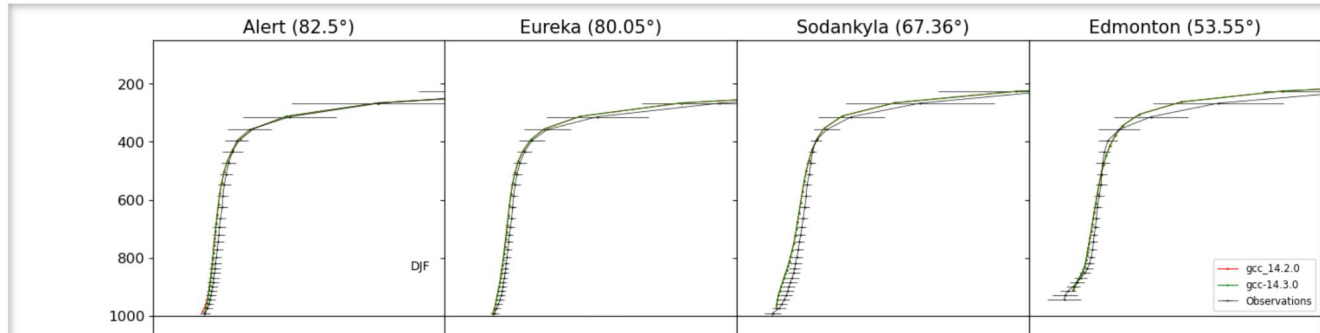
Markdown

```
#pdf_pages.close()
```

## Create plots of GEOS-Chem model output vs. observational data

A new figure will be generated showing GEOS-Chem data vs. ozonesonde sites for each meteorological season (winter = DJF, spring = MAM, summer = JJA, fall = SON). Close each plot window to move to the next plot.

```
[6]: # Finally, generate plots of GEOS-Chem data vs. ozonesondes
# These are used for GEOS-Chem benchmarking and model evaluation
make_benchmark_models_vs_sondes_plots(
    obs_data_file,
    obs_site_file,
    ref_filepaths,
    ref_label,
    dev_filepaths,
    dev_label,
)
```



Then some nice computation happens

# AI and Dataverse

# TurboCurator (ICPSR)

# AI & Dataverse

From across the community...



## TurboCurator

- Reads Dataset metadata fields entered by users in Dataverse
  - Suggestions for Title, Abstract (Description), Keywords
  - Uses ICPSR metadata best practices
  - Metadata curation tasks:
    - Metadata creation, submit for review workflows, improve metadata collections
    - Other kinds of metadata work potentially, possibilities for file level analysis, etc.

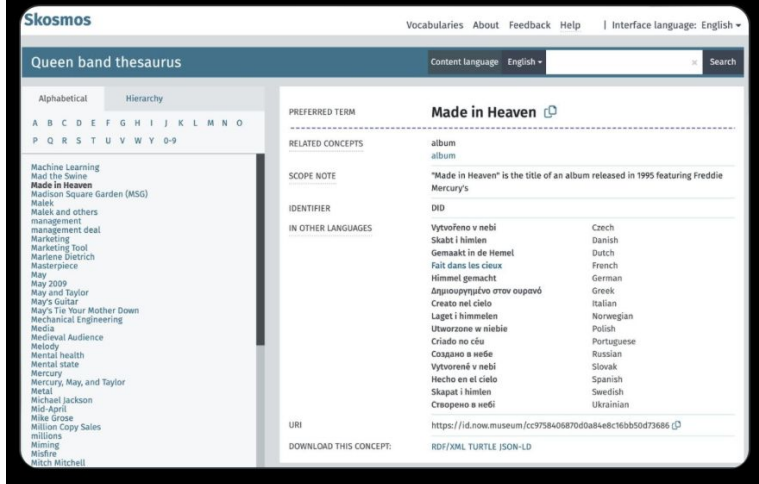
The screenshot shows the TurboCurator web interface. At the top, there is a navigation bar with the ICPSR logo and the text 'TurboCurator'. Below the navigation bar, there is a section titled 'Tell us more about your deposit' with a link 'Why are we requesting more information?'. A yellow warning box states: 'Do not enter any sensitive information that may contain direct or indirect identifiers in your data. Information entered here will be sent to Open AI's ChatGPT. ICPSR's data confidentiality policy (new tab)'. Below this, there is a section titled 'Enter additional information that might offer improved suggestions' with a text area. A blue box asks 'What kind of information are we looking for?' and provides instructions: 'Abstract or summary about your research. This could come from a published paper or a press release on your study. If you don't have an abstract or summary on hand, give us a two-sentence description of your study.' Below this, there is a 'Title' section with a text area containing 'Labour Force Survey: January 2024 [Canada]'. A blue button 'Copy Title Recommendation' is next to it. Below the 'Title' section, there is a 'Your Entry' section with a text area containing 'Testing Curation Features'. A 'Hide' button is at the bottom. A purple text overlay 'Title Abstract Keywords' is positioned to the right of the form, with an arrow pointing to the 'Title' section.

# AI & Dataverse

From across the community... [Slava Tykhonov \(DANS-KNAW\)](#)

 **Slava Tykhonov**   
@4tykhonov

My latest experiments on the Open Source #LLM models' "decompression" show great potential in transforming all "packaged" content into a structured format and ingesting it back into the knowledge graph. Even in the current state of art, Generative AI is capable of saving significant human resources and constructing a 'good enough' thesaurus on any topic in a limited time. Still don't believe me? Check out this thesaurus on the @QueenWillRock, generated by #AI without any human involvement and automatically translated into other languages! More is coming in the @MuselT\_EU.




**Skosmos** Vocabularies About Feedback Help | Interface language: English ▾

Queen band thesaurus Content language: English ▾ Search

Alphabetical Hierarchy  
A B C D E F G H I J K L M N O  
P Q R S T U V W X Y 0-9

Machine Learning  
Mad the Swine  
**Made in Heaven**  
Madison Square Garden (MSG)  
Malek  
Malek and others management management deel  
Marketing Tool  
Marlene Dietrich  
Masterpiece  
May  
May 2009  
May and Taylor  
May's Guitar  
May's Te Your Mother Down  
Mechanical Engineering  
Media  
Medival Audience  
Melody  
Mental health  
Mental state  
Mercury  
Mercury, May, and Taylor  
Metal  
Michael Jackson  
Mid-April  
Mike Grose  
Million Copy Sales  
millions  
Mining  
Nofire  
Nutch Mitchell

**PREFERRED TERM**  
**Made in Heaven** 


**RELATED CONCEPTS**  
album  
album

**SCOPE NOTE**  
"Made in Heaven" is the title of an album released in 1995 featuring Freddie Mercury's

**IDENTIFIER**  
DID

**IN OTHER LANGUAGES**

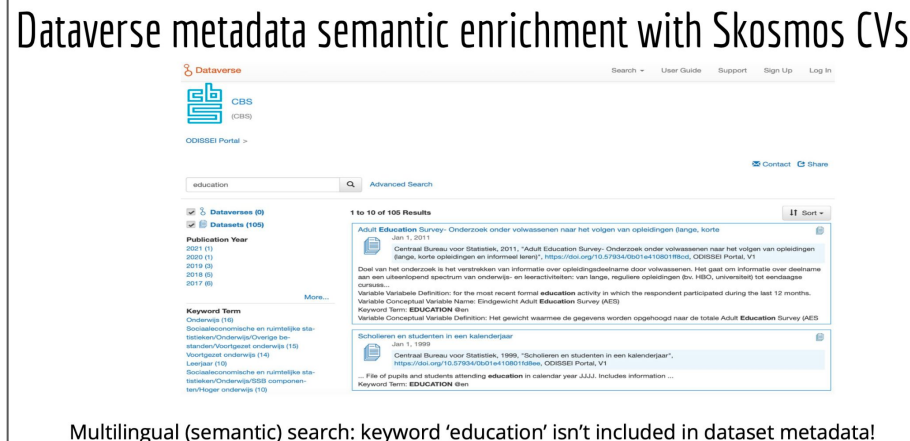
Выворочено з неби	Czech
Skabt i himlen	Danish
Gemaakt in de Hemel	Dutch
Fait dans les cieux	French
Himmel gemacht	German
Άγιοσπηγιάβιο στρον ουσάβν	Greek
Creato nel cielo	Italian
Laget i himmelen	Norwegian
Utaorzone w niebie	Polish
Criado no céu	Portuguese
Cozagan a nebe	Russian
Vytvoreni v nebi	Slovak
Hecho en el cielo	Spanish
Skapat i himlen	Swedish
Сропено а неби	Ukrainian

**URI**  
<https://id.now.museum/cc975840870d0a84e4ec16bb50d73686> 


**DOWNLOAD THIS CONCEPT:** RDF/XML Turtle JSON-LD

## Knowledge Graph Extraction via LLM

Applications to Dataverse include:  
automatic **metadata semantic enrichment**,  
**keyword translation**, etc.



**Dataverse** Search ▾ User Guide Support Sign Up Log In

 CBS (CBS)  
ODDSSEI Portal ▾

education  Advanced Search

Datasets (0)  
 Datasets (105)

1 to 10 of 105 Results 11 Sort ▾

**Adult Education Survey- Onderzoek onder volwassenen naar het volgen van opleidingen (lange, korte**  
Jan 1, 2011  
Centraal Bureau voor de Statistiek, 2011. "Adult Education Survey- Onderzoek onder volwassenen naar het volgen van opleidingen (lange, korte opleidingen en informeel leren)". <https://doi.org/10.27334/doi/10001800>. ODISSEI Portal, V1

Doel van het onderzoek is het verstrekken van informatie over opleidingsdeelname door volwassenen. Het gaat om informatie over deelname aan een uiteenlopend spectrum van onderwijs- en leeractiviteiten: van lange, reguliere opleidingen (bv. HBO, universiteit) tot eenzijdige cursussen.  
Variable Labels: Definition: In the most recent formal education activity in which the respondent participated during the last 12 months.  
Variable Conceptual Variable Name: Eenvolgewicht Adult Education Survey (AES)  
Keyword Term: EDUCATION Item  
Variable Conceptual Variable Definition: Het gewicht waarmee de gegevens worden opgehoogd naar de totale Adult Education Survey (AES)

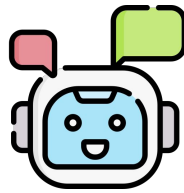
**Scholieren en studenten in een kalenderjaar**  
Jan 1, 1999  
Centraal Bureau voor de Statistiek, 1999. "Scholieren en studenten in een kalenderjaar". <https://doi.org/10.27334/doi/1010801800a>. ODISSEI Portal, V1

... File of pupils and students attending education in calendar year JJJJ. Includes information ...  
Keyword Term: EDUCATION Item

**Multilingual (semantic) search: keyword 'education' isn't included in dataset metadata!**

# AI & Dataverse

From across the community...



## “Ask the Data” (IQSS, Harvard)

- Reads tabular data file variable names, labels, codes, data values
- Performs data analysis of two or more variables and crosstabulates, statistical regression, topic-modelling
- Connected to LLM AI for AI-generated responses
- Natural language instructions/questions processed as SQL query in Dataverse
- Only 181 lines of Python code with Shiny app and OpenAI’s ChatGPT
- Supports multiple languages

Variable names & labels

What is this data about?

Tell me what you want to know

What is this data about?

Answer please

The data in the dataframe "df" appears to be related to a survey or study involving various demographic and employment-related variables. Here are some of the columns present in the data: SURVYEAR: Survey year - SURVMNTH: Survey month - LFSSTAT: Labor force status - PROV: Province - CMA: Census Metropolitan Area - AGE\_12: Age in 12-year groups - AGE\_6: Age in 6-year groups - SEX: Gender - MARSTAT: Marital status - EDUC: Education level - MJH: Major field of study - EVERWORK: Ever worked - FTPTLAST: Full-time or part-time in last job - COWMAIN: Class of worker in main job - IMMIG: Immigration status - NAICS\_21: Industry sector - NOC\_10: National Occupational Classification - NOC\_43: National Occupational Classification - and many more... The data seems to capture information about individuals' demographics, employment status, education, and other related factors.

REC_NUM	SURVYEAR	SURVMNTH	LFSSTAT	PROV	CMA	AGE_12	AGE_6	SEX	MARSTAT	EDUC	MJH	EVERWORK	FTPTLAST	COWMAIN	IMMIG	
1	2024	1	4	24	0	12	2	1	2	2					3	
2	2024	1	4	47	0	4		1	6	2	3				3	
3	2024	1	1	10	0	1	2	1	6	2	1			2	3	
4	2024	1	1	35	4	12		1	2	4	1			2	3	
5	2024	1	1	35	0	6		1	1	4	1			2	3	
6	2024	1	4	59	0	10		2	2	4	2				3	
7	2024	1													2	3
8	2024	1													2	3
9	2024	1													2	3

Ask the Data

Multivariate statistical analysis and modelling

Tell me what you want to know

What is the relationship between education and income?

Answer please

The average hourly earnings (HRLYEARN) for each level of education are as follows: Level 0: \$23.58 - Level 1: \$22.69 - Level 2: \$26.3 - Level 3: \$25.83 - Level 4: \$33.01 - Level 5: \$40.58 - Level 6: \$48.12. From the data, we can see that there is a positive relationship between education level and income. Generally, as the level of education increases, the average income also tends to increase.

REC_NUM	SURVYEAR	SURVMNTH	LFSSTAT	PROV	CMA	AGE_12	AGE_6	SEX	MARSTAT	EDUC	MJH	EVERWORK	FTPTLAST	COWMAIN	IMMIG
1	2024	1	4	24	0	12		2	1	2		2			3
2	2024	1	4	47	0	4		1	6	2	3				3
3	2024	1	1	10	0	1	2	1	6	2	1			2	3
4	2024	1	1	35	4	12		1	2	4	1			2	3
5	2024	1	1	35	0	6		1	1	4	1			2	3
6	2024	1	4	59	0	10		2	2	4	2				3
7	2024	1	1	24	2	10		1	5	6	1			2	3
8	2024	1	1	59	0	3	6	1	2	4	1			2	3
9	2024	1	1	59	0	11		1	1	2	1			2	3
10	2024	1	3	48	0	5		2	1	4		1	2	2	3

Viewing rows 1 through 10 of 109278

What's Next...

# Panel & Audience Q & A



**Stefano Iacus**  
*Director of Data Science &  
Product Research, Managing  
Director Dataverse Project*



**Sonia Barbosa**  
*Associate Director, Dataverse, Data  
Curation, and The Murray Archive*



**Amber Leahey**  
*Service Director, Borealis (Canada)*



**Gustavo Durand**  
*Dataverse Technical Lead &  
Architect*



**Ceilyn Boyd**  
*Dataverse Development Project  
Manager*