



GREI

Facilitating use of Generalist Repositories to Share and Discover Data:

**A Workshop by the NIH Generalist Repository
Ecosystem Initiative repositories**

March 27, 2023

RDAP 2023 Summit Workshop





GREI

Workshop Etiquette

- Please be mindful of the time so we can keep to the schedule
- Slides are publicly available (<https://doi.org/10.5281/zenodo.7774200>)
- Participants are muted during presentations, but active participation is encouraged
- Ask questions via Zoom chat or raise your hand for a facilitator to call on you
- We value the diversity of views, expertise, opinions, and experiences of all participants
- Please treat your fellow workshop participants with respect and consideration





GREI

RDAP Code of Conduct

The Research Data Access and Preservation (RDAP) Association is committed to providing an inclusive environment where all people can participate fully in all activities without fear of harassment or discriminatory behavior. The [RDAP Code of Conduct](#) is in effect during the RDAP Summit 2023.

To report an incident, please reach out to one of the Code of Conduct Helpers through Whova, send an email to codeofconduct@rdapassociation.org, or anonymously through the incident report form.



Dave

Meet our Speakers



David Scherer
*Customer Consultant,
Elsevier*



Ana Van Gulick, PhD
*Government and Funder
Lead, Head of Data Review,
Figshare*



Andrew Mckenna-Foster
*Product Specialist,
Figshare*



Meet our Speakers



Julie Goldman
*Research Data Services
Librarian, Harvard Library*

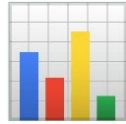


Sarah Lippincott
*Head of Community
Engagement, Dryad*



Gretchen Gueguen
*Product Owner,
Center for Open Science*





Poll Question

Tell us why you are here!



Workshop Outline

- Welcome and introductions (5 min)
- About generalist repositories (10 min)
- Scenario 1: Depositing Data (20 min)
- Scenario 2: Finding and Reusing Data (20 min)
- Facilitated breakout sessions (25 min)
- Break! (10 min)
- Scenario 3: Data Sharing in Multiple Repositories (20 min)
- Scenario 4: Budgeting for Generalist Repositories (20 min)
- Facilitated breakout sessions (25 min)
- Wrap-up and close (10 min)




Introduction to Generalist Data Repositories





Research Data Repository Ecosystem


Different trees in the same forest



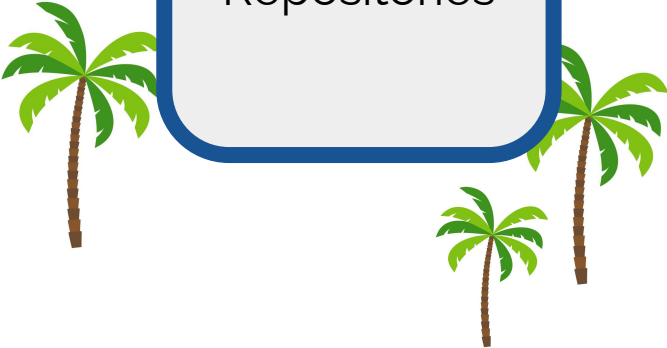
Domain-specific
Repositories



Generalist
Repositories



Institutional
Repositories





Generalist Repositories



<https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/generalist-repositories>



Desirable Characteristics of Data Repositories

When choosing a repository to manage and share data resulting from Federally funded research, look for:

- 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



Generalist Repository Comparison Chart

doi:10.5281/zenodo.3946720

This chart is designed to assist researchers in finding a generalist repository should no domain repository be available to preserve their research data. Generalist repositories accept data regardless of data type, format, content, or disciplinary focus. For this chart, we included a repository available to all researchers specific to clinical trials (Vivli) to bring awareness to those in this field.

<https://fairsharing.org/collection/GeneralRepositoryComparison>

TOPIC	HARVARD DATAVERSE	DRYAD	FIGSHARE	MENDELEY DATA	OSF	VIVLI	ZENODO
Brief Description	Harvard Dataverse is a free data repository open to all researchers from any discipline, both inside and outside of the Harvard community, where you can share, archive, cite, access, and explore research data.	Open-source, community-led data curation, publishing, and preservation platform for CC0 publicly available research data Dryad is an independent non-profit that works directly with: <ul style="list-style-type: none"> researchers to publish datasets utilizing best practices for discovery and reuse publishers to support the integration of data availability statements and data citations into their workflows institutions to enable scalable campus support for research data management best practices at low cost 	A free, open access, data repository where users can make all outputs of their research available in a discoverable, reusable, and citable manner. Users can upload files of any type and are able to share diverse research products including datasets, code, multimedia files, workflows, posters, presentations, and more. With discoverable metadata supporting FAIR principles, file visualizations, and integrations, researchers can make their work more impactful and move research further faster.	Mendeley Data is a free repository specialized for research data. Search more than 20+ million datasets indexed from 1000s of data repositories and collect and share datasets with the research community following the FAIR data principles.	OSF is a free and open source project management tool that supports researchers throughout their entire project lifecycle in open science best practices.	Vivli is an independent, non-profit organization that has developed a global data-sharing and analytics platform. Our focus is on sharing individual participant-level data from completed clinical trials to serve the international research community.	Powering Open Science, built on Open Source. Built by researchers for researchers. Run from the CERN data centre, whose purpose is long term preservation for the High Energy Physics discipline, one of the largest scientific datasets in the world
Size limits	No byte size limit per dataset. Harvard Dataverse currently sets a file size limit of 2.5GB.	300GB/dataset	Soft limit of 20GB/file for free accounts. System limit of 5000GB/file. Unlimited storage of public data but 20GB storage for private data for free accounts. Email info@figshare.com to have upload and storage limits raised.	10GB per dataset	Projects currently have not storage limit. There is a 5GB/file upload limit for native OSF Storage. There is no limit imposed by OSF for the amount of storage used across add-ons connected to a given project.	If more than 10GB per study data, reach out to us	50GB per dataset, contact us via https://zenodo.org/support for higher limits
Storage space per researcher	1 TB per researcher	No limit	No limit	No limit	No limit	No limit	No limit
Persistent, Unique Identifier Support	DOI, Handle	DOI	DOI	DOI	DOI	DOI	DOI

<https://doi.org/10.5281/zenodo.3946719>

Common and Unique Repository Features:

Common:

Core Metadata
 Persistent Identifiers
 Discoverable
 Flexibility
 Open access, FAIR
 Metrics

Unique:

Output types
 Storage, size limits
 Licenses
 Review
 Controlled Access
 Visualization
 Costs



Introduction to the Generalist Repository Ecosystem Initiative (GREI)



NIH Generalist Repository Ecosystem Initiative

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

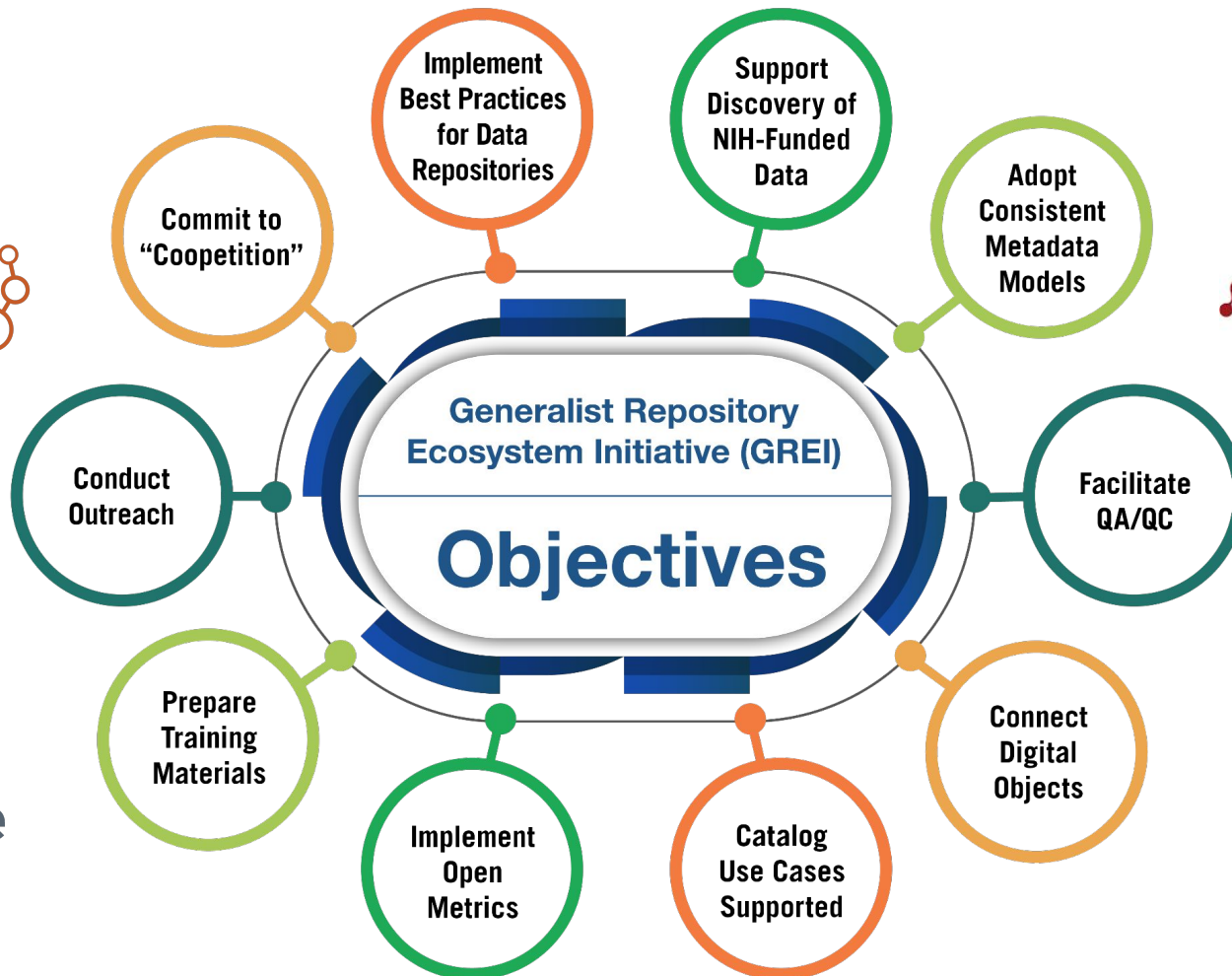
Avoid **duplication** of data across repositories



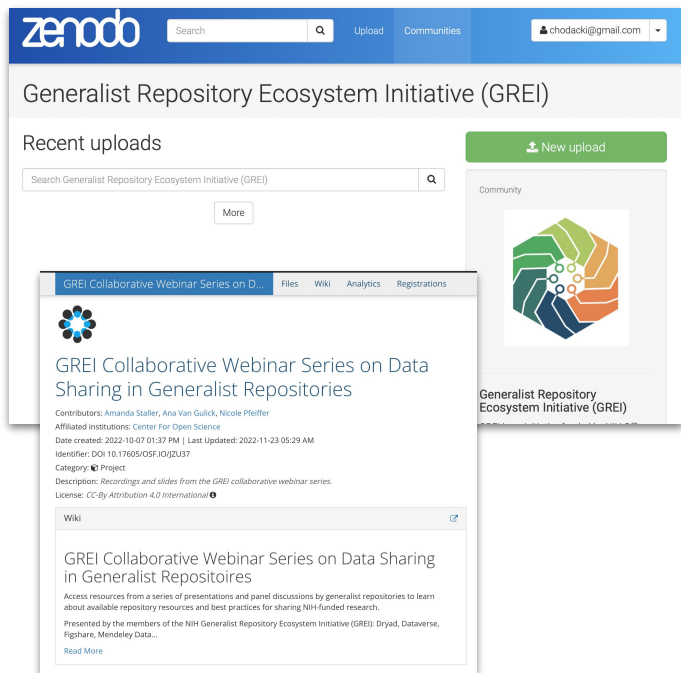
5

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





Connect with GREI



The screenshot displays the Zenodo interface for the Generalist Repository Ecosystem Initiative (GREI). At the top, the Zenodo logo is on the left, and a search bar, 'Upload' button, 'Communities' link, and user profile 'chodacki@gmail.com' are on the right. Below the header, the page title is 'Generalist Repository Ecosystem Initiative (GREI)'. A 'Recent uploads' section includes a search bar for 'Generalist Repository Ecosystem Initiative (GREI)' and a 'More' button. To the right is a 'New upload' button and a 'Community' section featuring the GREI logo (a colorful hexagon with arrows) and the text 'Generalist Repository Ecosystem Initiative (GREI)'. An inset window shows a 'GREI Collaborative Webinar Series on Data Sharing in Generalist Repositories' page. This page includes the GREI logo, title, contributors (Amanda Staller, Ana Van Gulick, Nicole Pfeiffer), affiliated institutions (Center For Open Science), dates (created: 2022-10-07 01:37 PM, last updated: 2022-11-23 05:29 AM), identifier (DOI: 10.17605/OSF.IO/JZU37), category (Project), description (Recordings and slides from the GREI collaborative webinar series), license (CC-BY Attribution 4.0 International), and a 'Wiki' section with a brief overview of the webinar series and its purpose.

- [GREI Training & Outreach Calendar](#): Individual and collaborative webinars and workshops
- [GREI Forum](#): Program updates and questions. Join at <https://groups.google.com/g/contactgrei>
- GREI Community on Zenodo: <https://zenodo.org/communities/grei>
- GREI Webinar Series Recordings and Slides: <https://doi.org/10.17605/OSF.IO/JZU37>
- GREI Email: contactgrei@googlegroups.com
- NIH Program: grei@nih.gov



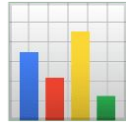
Scenario 1: Depositing Data



Librarians are key to helping researchers navigate data deposit in generalist repositories and employ data sharing best practices

Flexibility of generalist repositories can actually hinder adoption of best practices





Poll Question

Have you shared data or supported a researcher in sharing data in a generalist repository?



Best practices for Data Sharing

1. Gather all data needed for **reanalysis**
2. Verify files **can be shared** publicly
3. Choose **open** file formats
4. **Organize** files logically
5. Describe the dataset in a detailed **README** file
6. Choose a suitable repository to **share** the data
7. Create high quality, complete **metadata** with PIDs

Data Management



Data Sharing



Best practices for Data Sharing

1. Gather all data needed for **reanalysis**
2. Verify files **can be shared** publicly
3. Choose **open** file formats
4. **Organize** files logically
5. Describe the dataset in a detailed **README** file
6. Choose a suitable repository to **share** the data
7. Create high quality, complete **metadata** with PIDs

Data Management



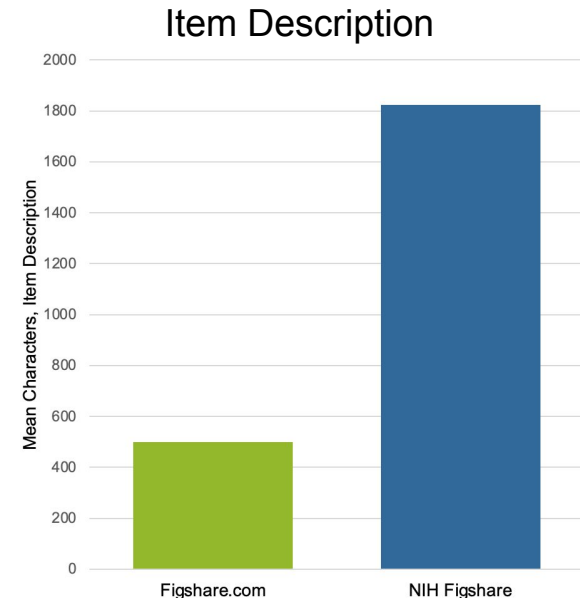
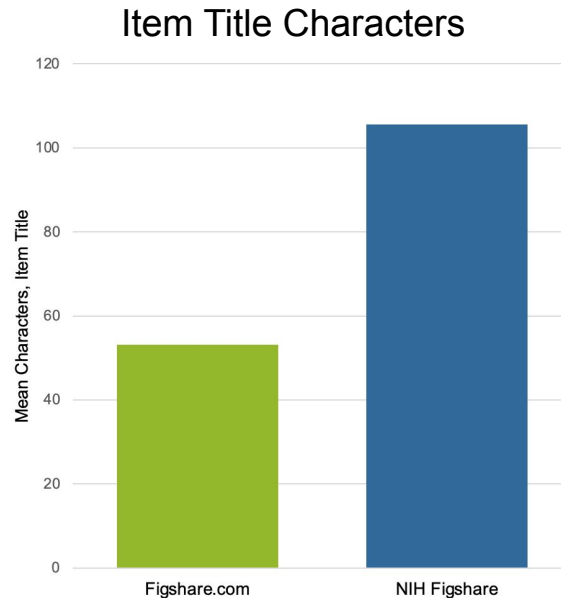
Data Sharing



We know without guidance or curation researchers often don't create rich metadata:

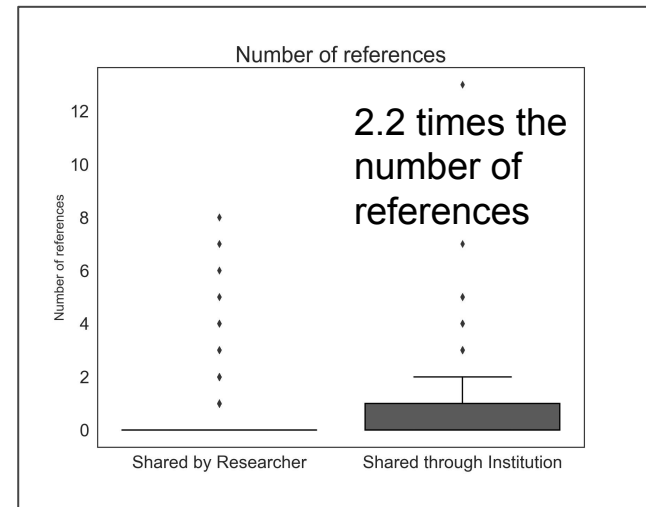
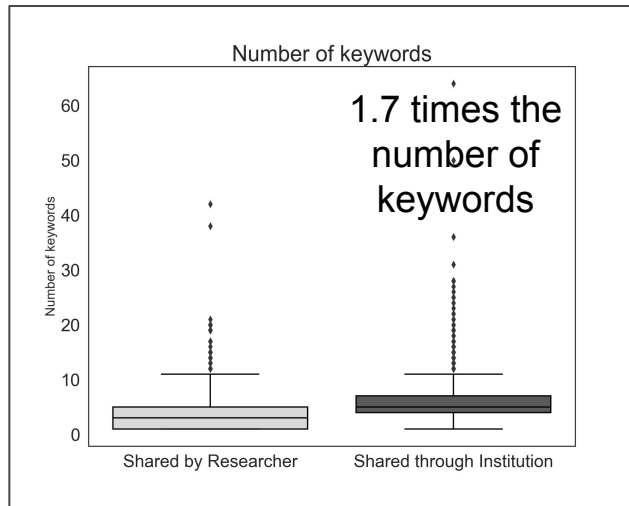
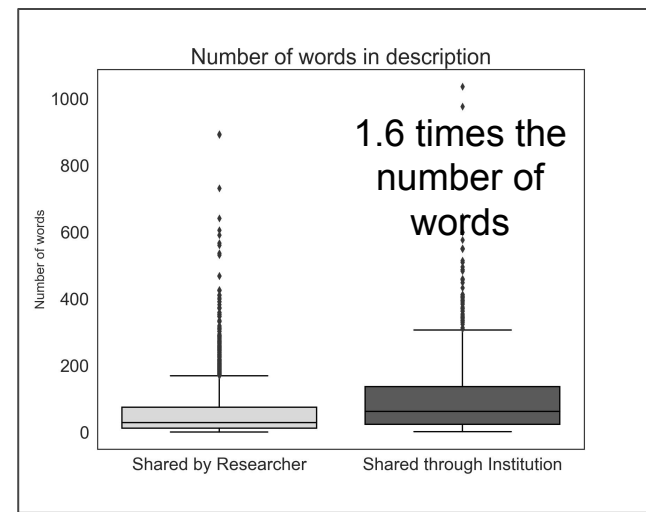
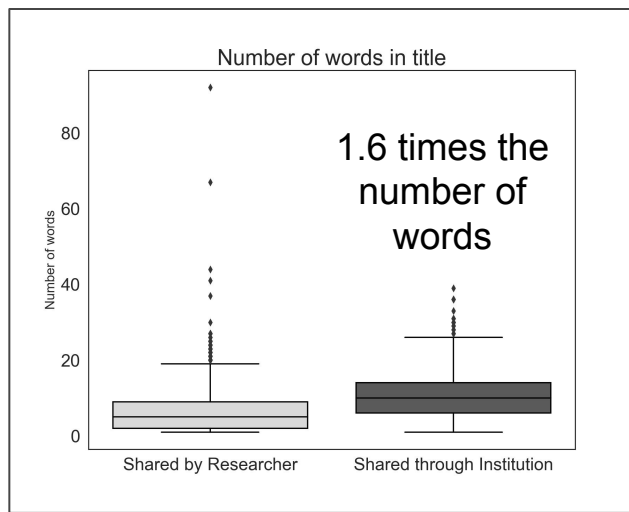
- Shorter titles, descriptions, fewer keywords, fewer related material links
- These are easy things for a librarian or curator to help with and significantly increase the findability and interoperability of a dataset

NIH Figshare items had on average 2x and 3.5x more characters in the item title and description respectively



**Records from
researchers
n = 2,470 records**

**Records from
institutions
n = 1,509 records**



You can help make good records great by offering suggestions on files and metadata in generalist repositories

Metadata and documentation support data discovery and reuse without adding significant barriers to data sharing



Librarians can also provide researchers with the *'why'* of specific repository best practices

Why are high quality metadata and documentation important?

Why is using a trusted repository with specific features important?

Discoverability and context of the dataset

- The dataset should stand on its own from the paper
- The dataset should be reusable on its own
- The dataset should be linked to related materials via PIDs - link to papers, funding sources, author IDs, institution IDs

Tracking impact and getting credit

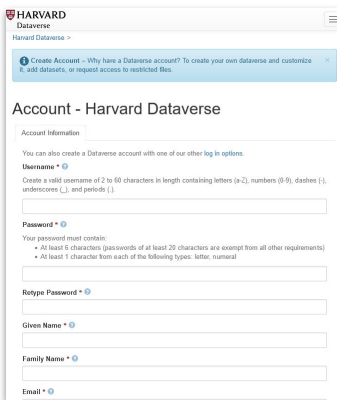
- Have more impact with open data that can be found and reused
- Meet funder requirements, report to NIH
- Track citations, metrics
- Reflected on researcher profiles (e.g. ORCID)



Getting started to support data deposit



Sign up for an account



HARVARD Dataverse

Harvard Dataverse >

Create Account - Why have a Dataverse account? To create your own dataverse and customize it, add datasets, or request access to restricted files.

Account - Harvard Dataverse

Account Information

You can also create a Dataverse account with one of our other log in options.

Username *

Create a valid username of 2 to 60 characters in length containing letters (a-z), numbers (0-9), dashes (-), underscores (_), and periods (.).

Password *

Your password must contain:

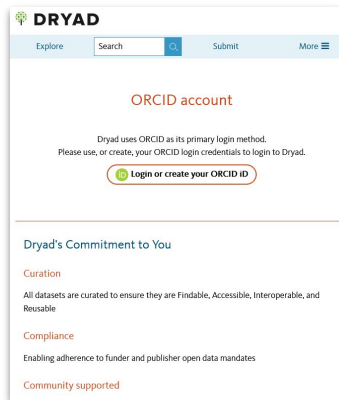
- At least 8 characters (passwords of at least 20 characters are exempt from all other requirements)
- At least 1 character from each of the following types: letter, numeral

Repeat Password *

Given Name *

Family Name *

Email *



DRYAD

Explore Search Submit More

ORCID account

Dryad uses ORCID as its primary login method. Please use, or create, your ORCID login credentials to login to Dryad.

[Login or create your ORCID ID](#)

Dryad's Commitment to You

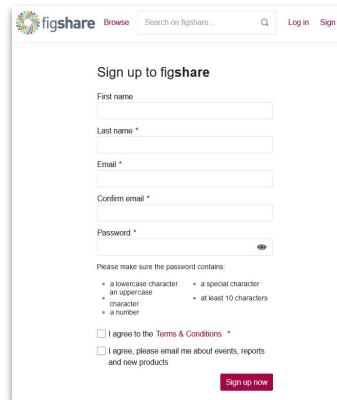
Curation

All datasets are curated to ensure they are Findable, Accessible, Interoperable, and Reusable

Compliance

Enabling adherence to funder and publisher open data mandates

Community supported



figshare Browse Search on figshare... Log in Sign up

Sign up to figshare

First name

Last name *

Email *

Confirm email *

Password *

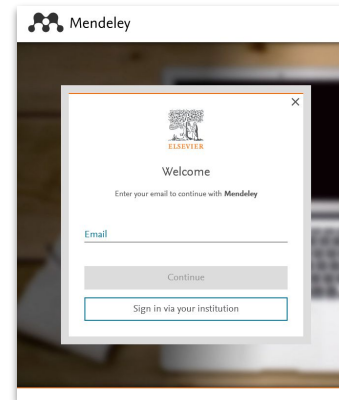
Please make sure the password contains:

- a lowercase character
- a special character
- a character
- a number
- at least 10 characters

I agree to the [Terms & Conditions](#) *

I agree, please email me about events, reports and new products

[Sign up now](#)



Mendeley

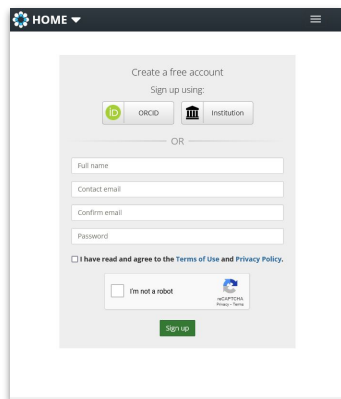
Welcome

Enter your email to continue with Mendeley

Email

[Continue](#)

[Sign in via your institution](#)



HOME

Create a free account

Sign up using:

[ORCID](#) [Institution](#)

OR

Full name

Contact email

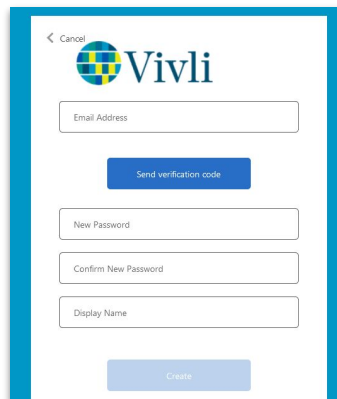
Confirm email

Password

I have read and agree to the [Terms of Use and Privacy Policy](#).

I'm not a robot

[Sign up](#)



Vivli

Email Address

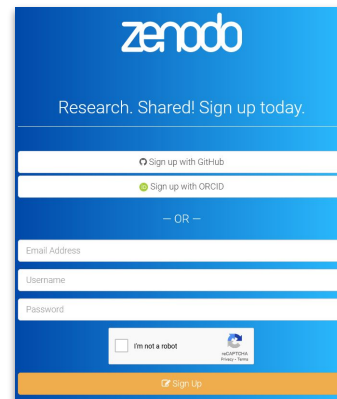
[Send verification code](#)

New Password

Confirm New Password

Display Name

[Create](#)



zenodo

Research. Shared! Sign up today.

[Sign up with GitHub](#)

[Sign up with ORCID](#)

— OR —

Email Address

Username

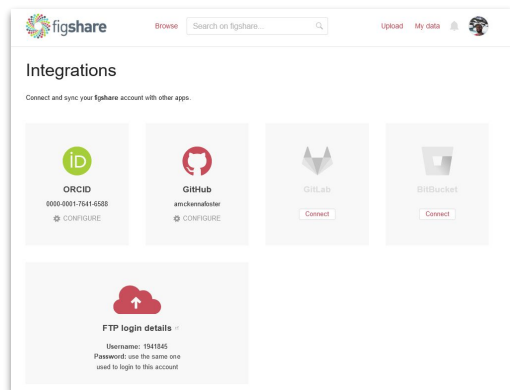
Password

I'm not a robot

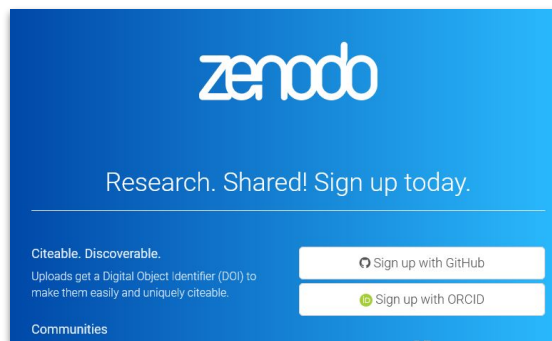
[Sign up](#)



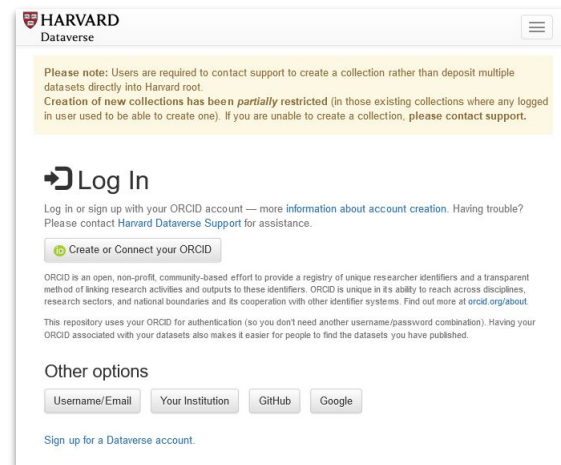
Connect account to ORCID



The screenshot shows the Figshare Integrations page. At the top, there is a search bar and navigation links for 'Upload' and 'My data'. The main heading is 'Integrations' with a sub-heading 'Connect and sync your Figshare account with other apps.' Below this, there are four integration options: ORCID (with ID 0000-0001-7641-6588), GitHub (username amckemwister), GitHub, and Bitbucket. Each option has a 'CONFIGURE' link and a 'Connect' button. At the bottom, there is a section for 'FTP login details' with a username of 1941945 and a note to use the same password as the account.



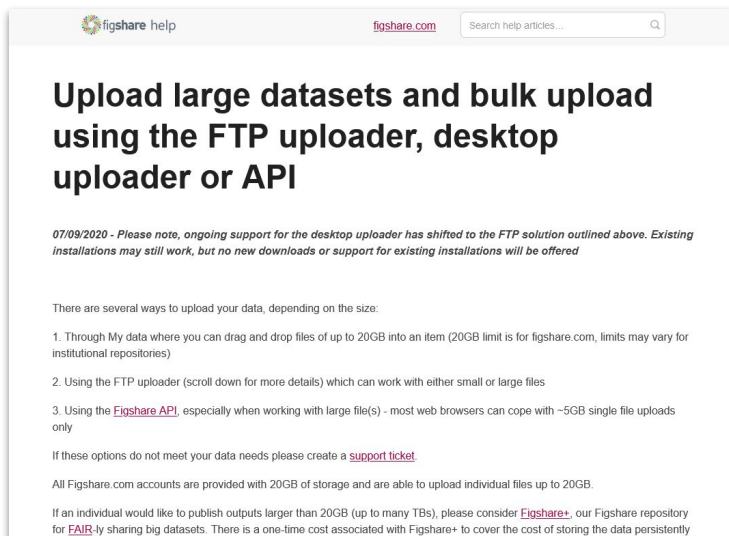
The screenshot shows the Zenodo sign-up page. The background is blue with the Zenodo logo at the top. The main text reads 'Research. Shared! Sign up today.' Below this, there are two buttons: 'Sign up with GitHub' and 'Sign up with ORCID'. At the bottom, there is a 'Communities' link.



The screenshot shows the Harvard Dataverse login page. At the top, there is a navigation bar with the Harvard Dataverse logo and a search bar. Below this, there is a yellow warning box with the text: 'Please note: Users are required to contact support to create a collection rather than deposit multiple datasets directly into Harvard root. Creation of new collections has been partially restricted (in those existing collections where any logged in user used to be able to create one). If you are unable to create a collection, please contact support.' Below the warning box, there is a 'Log In' section with a sub-heading 'Log in or sign up with your ORCID account — more information about account creation. Having trouble? Please contact Harvard Dataverse Support for assistance.' There is a button for 'Create or Connect your ORCID'. Below this, there is a paragraph of text explaining ORCID and its use for authentication. At the bottom, there is a section for 'Other options' with buttons for 'Username/Email', 'Your Institution', 'GitHub', and 'Google'. There is also a link for 'Sign up for a Dataverse account.'



Know about alternative upload options



figshare help figshare.com Search help articles...

Upload large datasets and bulk upload using the FTP uploader, desktop uploader or API

07/09/2020 - Please note, ongoing support for the desktop uploader has shifted to the FTP solution outlined above. Existing installations may still work, but no new downloads or support for existing installations will be offered

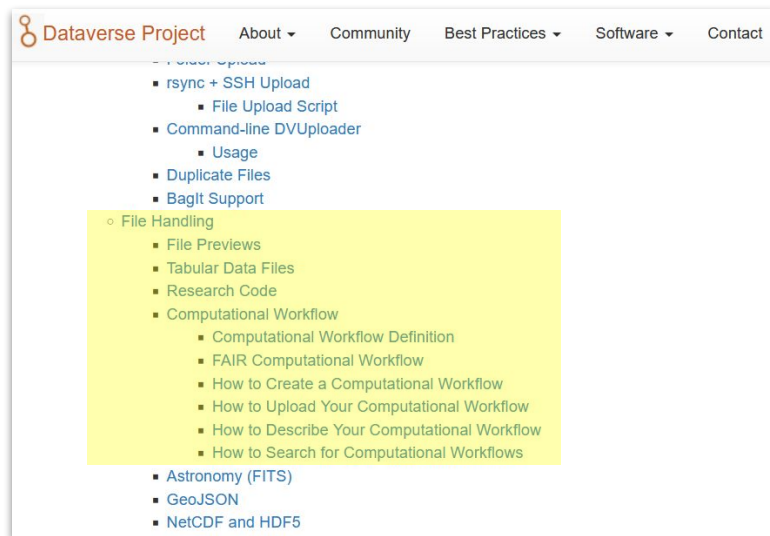
There are several ways to upload your data, depending on the size:

1. Through My data where you can drag and drop files of up to 20GB into an item (20GB limit is for figshare.com, limits may vary for institutional repositories)
2. Using the FTP uploader (scroll down for more details) which can work with either small or large files
3. Using the [Figshare API](#), especially when working with large file(s) - most web browsers can cope with ~5GB single file uploads only

If these options do not meet your data needs please create a [support ticket](#).

All Figshare.com accounts are provided with 20GB of storage and are able to upload individual files up to 20GB.

If an individual would like to publish outputs larger than 20GB (up to many TBs), please consider [Figshare+](#), our Figshare repository for [FAIR](#)-ly sharing big datasets. There is a one-time cost associated with Figshare+ to cover the cost of storing the data persistently



Dataverse Project About Community Best Practices Software Contact

- File Handling
 - rsync + SSH Upload
 - File Upload Script
 - Command-line DVUploader
 - Usage
 - Duplicate Files
 - BagIt Support
- File Handling
 - File Previews
 - Tabular Data Files
 - Research Code
 - Computational Workflow
 - Computational Workflow Definition
 - FAIR Computational Workflow
 - How to Create a Computational Workflow
 - How to Upload Your Computational Workflow
 - How to Describe Your Computational Workflow
 - How to Search for Computational Workflows
- Astronomy (FITS)
- GeoJSON
- NetCDF and HDF5




Policies

What are the size limits in Zenodo?

We currently accept up to 50GB per dataset (you can have multiple datasets); there is no size limit on communities. However, we don't want to turn away larger use cases. If you would like to upload larger files, please [contact us](#), and we will do our best to help you. Please be aware that we cannot offer infinite space for free, so donations towards sustainability are encouraged.



Try uploading something (and maybe publish)

Browse Upload My data  

Draft





My Example Dataset

Add Files Link to external files Set as metadata record

Drag and drop files to upload or

Browse for files

Item actions

-  Add embargo
-  Share with private link
-  Reserve DOI
-  Edit timeline

Preview item

Delete item

Save changes

Publish item

* mandatory fields required to publish the item

Item title *

Group *

Selecting a different group will result in a different set of metadata fields. The information on the fields that don't match the new group will be lost.

Current group: **figshare**

Change group

Item Type *

[Learn more about item types](#)

Authors *

Andrew Mckenna-Foster x

Search for authors by name, full email or ORCID.

Categories *

Biomedical imaging x

Browse or search for categories







Keywords *

fMRI adaptation x

Add keywords for easy discovery. Hit enter after each...

Description *

Add as much context as possible so that others can interpret your research and reproduce it. Make sure you include methodology, techniques used and if relevant information about approval for data collection to confirm adherence to legal or ethical requirements. The description should have at least four characters.

H2 H3 H4 P B I U S    **A, A²**   

C



Try uploading something (and maybe publish)

Funding ⓘ

Biomedical Simulations Resource (BMSR) ×

My University Research Award Grant ×

+ [Add another](#)

References ⓘ

Add a URL that links to references or related content

+ [Add another](#)

Resource Title and DOI ⓘ

Resource Title

Add the title of the resource you want to link to...


Resource DOI


Add the DOI you want to link to...


Licence* ⓘ


CC0 ▾

Item actions

 [Add embargo](#)

 [Share with private link](#)

 [Reserve DOI](#)

 [Edit timeline](#)

Preview item

Delete item

[Save changes](#) [Publish item](#)



Provide help for researchers



Browse

Search on figshare...



Upload

My data



TF

My data

Projects

Collections

Activity

← Curation Help Private Project

MANAGE ⚙

Created on 17 Mar 2023 - 14:34 by [Team Figshare](#)
Last edited on 17 Mar 2023 - 14:35

Members (2)



Drop your files here and I can provide help with adding metadata and formatting files before you publish. Before you publish your record, add it to this project. I will be able to see the metadata and files and will leave comments for you. You can remove it from this project at anytime.

⌵ Hide project details

+ Add new content

Sort ▾

Filter ▾

search items



just now



Data for: An extremely fascinating and groundbreaking experiment

...



add your comment here

Comment

just now



A few things to make the record more discoverable and reusable:

just now



Add a README file that describes the processing, analysis, and variables. You can use the template in the library LibGuide: <https://libguides.mylib.edu/data-services>

just now



The description could use a concise summary of the methods and tools used to gather the data.

just now



Add a few more keywords- really good to have at least 5. Think of other terms people might use to search for this type of data.

Edit Remove



Browse

Search on figshare...



Upload

My data



TF

Draft My Example Dataset

edited on 2023-03-17, 14:01 by [Team Figshare](#)

● Add Files

Private Link

<https://figshare.com/s/d452c6df6825daecebd>

Copy link

Do not reference this link in papers. For referencing, use the public DOI.

Currently, the private link will be disabled on the date shown below. You can select a new one by using the calendar. Please note that Figshare uses UTC time!

2037-12-30

If you want to immediately disable the private link, use the button below.

[Disable private link](#)

Close

Useful metadata tips

5	872	Het	F		100	0	7.300435	65.816952	68.491092	
6	875	KO	F		100	0	10.346308	45.088171	47.231379	
7	874	WT	F		100	0	3.077627	19.545515	18.54744	
8	1013	KO	F		100	0	41.626739	58.744869	64.618665	
9	1001	WT	F		100	0	0	4.954823	16.468346	
10	1009	Het	F		100	0	4.344541	34.201454	38.664873	
11	1024	KO	F		100	0	3.335691	11.917454	18.937212	
12	1022	WT	F		100	0	0	2.307316	4.750816	
13	1023	Het	F		100	0	30.119366	57.731174	82.34355	
14	1031	KO	F		100	0	1.801409	25.59414	30.11289	
15	1035	Het	F		100	0	10.418207	52.96203	61.571854	
16	1037	WT	F		100	0	2.094009	11.018232	22.754934	

IR RPP (Fig 1) IR Infact (Fig 1) Oligomycin RPP (Fig 2) Oligomycin Infact (Fig 2) Doxycycline RPP (Fig 3) Doxycycline Infact (Fig 3)

Complete Data Set for ATF5 UPRmt pa...xlsx (92.84 kB)

Complete Data Set for ATF5 UPRmt paper

[Cite](#) [Download \(92.84 kB\)](#) [Share](#) [Embed](#) [+ Collect](#) [...](#)

Dataset posted on 2019-08-08, 11:10 authored by [Paul Brookes](#)

USAGE METRICS

435 views 56 downloads 2 citations

CATEGORIES

- Biochemistry and cell biology not elsewhere classified

KEYWORDS

Mitochondria

Biochemistry and Cell Biology not ...

LICENCE

CC BY 4.0

Complete Data Set for ATF5 UPRmt paper

FUNDING

NIH R01-HL127891, NIH R01-HL071158

HISTORY

- 2019-08-08 - First online date, Posted date

REFERENCES

Descriptive file name

Descriptive title

Add detailed description

List and link to funding

Suggest linking to related records

Suggest more categories

Suggest 5 keywords

License is as open as possible



Useful metadata tips

5	872	Het	F		100	0	7.300435	65.816952
6	875	KO	F		100	0	10.346308	45.098171
7	874	WT	F		100	0	3.077627	19.545515
8	1013	KO	F		100	0	41.626739	18.744869
9	1001	WT	F		100	0	0	4.954823
10	1009	Het	F		100	0	4.344541	34.201454
11	1024	KO	F		100	0	3.335691	11.917454
12	1022	WT	F		100	0	0	2.307316
13	1023	Het	F		100	0	30.119366	57.731174
14	1031	KO	F		100	0	1.801409	25.59414
15	1035	Het	F		100	0	10.418207	52.96203
16	1037	WT	F		100	0	2.094099	11.018232

IR RPP (Fig 1) IR Infarct (Fig 1) Oligomycin RPP (Fig 2) Oligomycin Infarct (Fig 2) Doxycycline RPP (Fig 3) 0

Complete Data Set for ATF5 UPRmt pa...xlsx (92.84 kB)

Complete Data Set for ATF5 UPRmt paper

[Cite](#) [Download \(92.84 kB\)](#) [Share](#) [Embed](#) [Collect](#) [...](#)

Dataset posted on 2019-08-08, 11:10 authored by [Paul Brookes](#)

Complete Data Set for ATF5 UPRmt paper

FUNDING

NIH R01-HL127891, NIH R01-HL071158

HISTORY

2019-08-08 - First online date, Posted date

REFERENCES

figshare Browse Search on figshare...

Paul Brookes
0000-0002-8639-8413

5592 item views | 1202 item downloads | 12 citations

[Follow](#)

Author(s) linked to ORCID

Paul Brookes's public data

10 posts

DATASET

Original data for Milliken et al. Acidosis metabolomics paper
Dataset posted on 2022-12-02
Paul Brookes

DATASET

Dog Metabolomics Data Set 2021
Dataset posted on 2021-05-19
Paul Brookes

DATASET

Complete data set for Cx11-Cx1 RET ROS paper
Dataset posted on 2020-09-11
Paul Brookes

DATASET

Complete data set for Alkbh7 paper
Dataset posted on 2020-08-14
Paul Brookes

Descriptive file name

Descriptive title

Add detailed description

List and link to funding

Suggest linking to related records

Suggest more categories

Suggest 5 keywords

License is as open as possible



Scenario 2: Finding and Reusing Data



Scenario

You are asked by a faculty member for help in identifying datasets that can be combined for an interdisciplinary metastudy. They need:

- ? Relevant data that may exist in different disciplinary and generalist repositories
- ? Good quality, complete data with metadata or codebooks for interpretation so that the data can be combined
- ? Data that is available and licensed for reuse



Breakdown of the problem

Find relevant data

Finding relevant datasets should be the first step. Use methods like:

- Data papers
- Citation chaining
- Targeted searching
- Looking for key characteristics

Evaluate data viability

Once datasets are identified, they can be evaluated for quality and fit with the research question. Evaluate the following:

- Metadata quality
- Indications of Reuse
- Metrics

Select for reuse

Viable datasets can be selected for reuse. Look for:

- Completeness of the data
- Presence of documentation
- Licensing terms and other restrictions



Finding datasets

Disciplinary repositories

- Targeted disciplines
- Rich discipline metadata



Generalist repositories

- Broad and interdisciplinary
- Commonly used metadata



Institutional Repositories

- Interdisciplinary, but limited to an institution
- Commonly used metadata



Meta-search

- Broad coverage
- Minimal metadata



Data Papers

- Describes a dataset in detail
- Peer reviewed
- Includes reuse information

cell

1,361,132 W

[CIL:48831](#)
Dataset publish
To determine th
where only one
other mouse ex
injected on the
and analyzed e

No citations v

<https://doi.org/>

[CIL:42161](#)
Dataset publish
Representative
mammary fragm
of movies that i

No citations v

<https://doi.org/>

[CIL:12608](#)
Dataset publish
Movie showing

Home About Contact Content Research Integrity Search... Account

Journal of Open Humanities Data

Start Submission

Reading: MultiHATHI: A Complete Collection of Multilingual Prose Fiction in the HathiTrust Digital Library [Download](#) A- A+ Share: [f](#) [t](#) [in](#) [e](#)

Data Papers

MultiHATHI: A Complete Collection of Multilingual Prose Fiction in the HathiTrust Digital Library

Authors: Sil Hamilton , Andrew Piper

Abstract

This dataset provides detailed metadata on ca. 10.2 million works of fiction and non-fiction written after 1799 in 521 different languages available in the HathiTrust Digital Library. The dataset bolsters the May 2022 Hathfile by supplying missing predicted fiction tags with a bespoke BERT-based multilingual classifier. Our classifier completes the catalogue with an additional 400,000 non-English volumes predicted to be works of fiction, capturing 95% of all works presently provided by HathiTrust. We provide each work with metadata including the work's genre at the level of fiction or non-fiction, length in pages, original language, and the year the work was published. With a total page count of ca. 1.4 billion pages, our dataset provides researchers with a substantial source of non-English modern literature. We also present insight into how multilingual classifiers can be trained with monolingual data, itself a discovery with implications for the study of lower resource languages. We hope our provisions will accelerate empirical research into non-English prose and literature.

Keywords: [fiction](#), [multilingual fiction](#), [non-English prose](#), [world literature](#)

Year: 2023 Volume: 9 Page/Article: 3 DOI: [10.5334/johd.95](#)

Published on 8 Feb 2023

Peer Reviewed CC BY 4.0

Contents

Annotations & Comments

Related Articles

Abstract

(1) Overview

(2) Method

(3) Dataset description

Object name

Format names and versions

Creation dates

Dataset Creators

Language

License

Repository name

Publication date

(4) Reuse potential

Repository location

Funding Information

Competing interests

Author Contributions

References

Citation Chaining

Cites

DRYAD Search []

Explore data | About | Help | Login

Data from: Evaluation of electronically supported nursing transfers between hospital and nursing home based on a test health telematics infrastructure: a case analysis

Schulte, Georg
Hübner, Ursula
Rienhoff, Otto
Quade, Matthias
Rottmann, Thorsten
Fenske, Matthias
Egbert, Nicole
Kuhlisch, Raik
Sellemann, Björn
Publication date: October 16, 2018
Publisher: Dryad
<https://doi.org/10.5061/dryad.9f2d8>

Citation

Schulte, Georg et al. (2018), Data from: Evaluation of electronically supported nursing transfers between hospital and nursing home based on a test health telematics infrastructure: a case analysis, Dryad, Dataset, <https://doi.org/10.5061/dryad.9f2d8>

Works referencing this dataset

Schulte, Georg et al. (2017), Evaluation einer elektronisch unterstützten pflegerischen Überleitung zwischen Krankenhaus und Pflegeheim unter Nutzung einer Test-Telematikinfrastruktur: eine Fallanalyse, GMS Medizinische Informatik, Article-Journal, <https://doi.org/10.3205/mibe000172>

Data files

Download dataset

October 16, 2018

Related Works

Article doi.org/10.3205/mibe000172

224 views

51 downloads

1 citations

GMS German Medical Science

Deutsch | MBE | About MBE | For authors | Contact | Imprint | GMS Meetings

Portal | Journals | Meetings | Reports | Guidelines | Handbooks

gmds GMS Medizinische Informatik, Biometrie und Epidemiologie
Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMS) eV
ISSN 1860-9171

Article | Current Volume | Archive | Search in MBE | Newsletter

Originalarbeit

Evaluation einer elektronisch unterstützten pflegerischen Überleitung zwischen Krankenhaus und Pflegeheim unter Nutzung einer Test-Telematikinfrastruktur: eine Fallanalyse

Evaluation of electronically supported nursing transfers between hospital and nursing home based on a test health telematics infrastructure: a case analysis

Georg Schulte - Forschungsgruppe Informatik im Gesundheitswesen, Hochschule Osnabrück, Osnabrück, Deutschland; Klinikum Osnabrück GmbH, Osnabrück, Deutschland

Ursula Hübner - Forschungsgruppe Informatik im Gesundheitswesen, Hochschule Osnabrück, Osnabrück, Deutschland

Otto Rienhoff - Institut für Medizinische Informatik, Universitätsmedizin Göttingen, Göttingen, Deutschland

Matthias Quade - Institut für Medizinische Informatik, Universitätsmedizin Göttingen, Göttingen, Deutschland

Thorsten Rottmann - Institut für Medizinische Informatik, Universitätsmedizin Göttingen, Göttingen, Deutschland

Matthias Fenske - Diakoniewerk Osnabrück gGmbH, Osnabrück, Deutschland

Nicole Egbert - Forschungsgruppe Informatik im Gesundheitswesen, Hochschule Osnabrück, Osnabrück, Deutschland

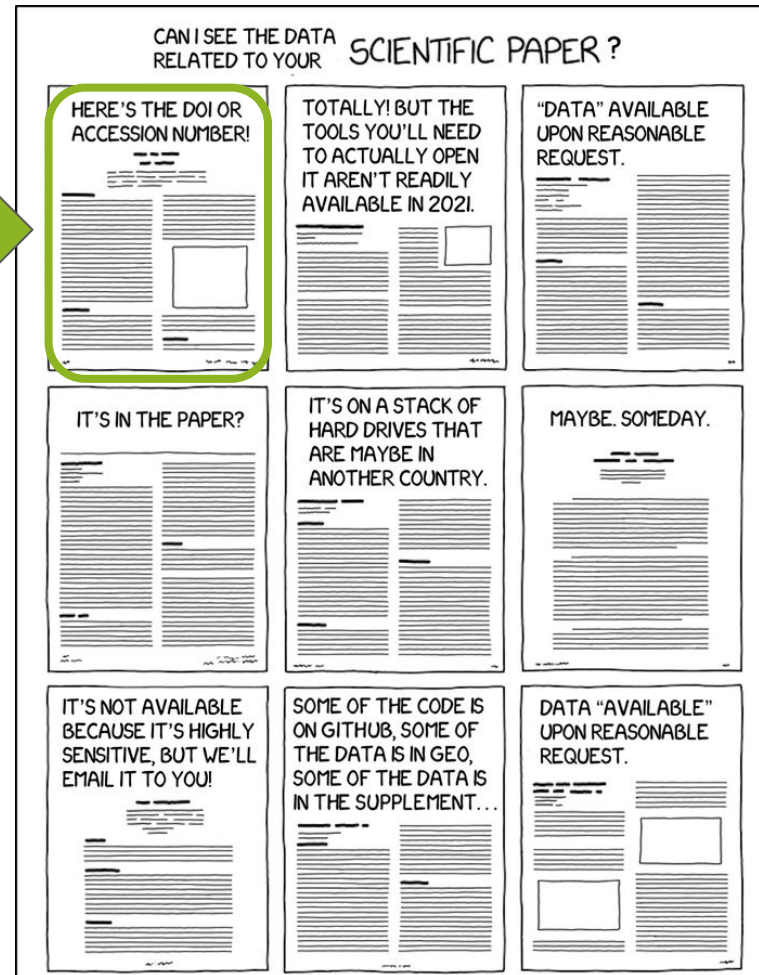
Raik Kuhlisch - Fraunhofer FOKUS, Berlin, Deutschland

Björn Sellemann - Institut für Medizinische Informatik, Universitätsmedizin Göttingen, Göttingen, Deutschland; Interdisziplinäre Notaufnahme, Universitätsmedizin Göttingen, Göttingen, Deutschland

GMS Med Inform Biom Epidemiol 2017;13(1):Doc05
[doi:10.3205/mibe000172](https://doi.org/10.3205/mibe000172) e7, www.ncbi.nlm.nih.gov/pmc/articles/PMC5400017/ e7

Is Cited By

Data Citation



John Borghi [@JohnBorghi]. (2021, May 5). As seen today at @GSVConference: Can I see the data related to your scientific paper? Another riff on that @XKCD comic. Original here: <https://xkcd.com/2456/> #csvconf #commallama [Tweet]. Twitter. <https://twitter.com/JohnBorghi/status/1390033561579835393>



Generalist Repository Identification

re3data | FAIRsharing.org

The screenshot shows the re3data.org search interface. On the left is a 'Filter' sidebar with categories: Subjects, Content Types (Archived data (1), Audiovisual data (2), Databases (3), Images (3), Raw data (3), Scientific and statistical data formats (4), Software applications (1), Standard office documents (4), Structured graphics (2), Structured text (1), other (1)), Countries, AID systems, API, Certificates, Data access, Data access restrictions, Database access, Database licenses, Data licenses, Data upload, Data upload restrictions, Enhanced publication, Institution responsibility type, and Institution type. The main search area has a search bar with 'animal behavior' and a 'Filter' button. Below the search bar are navigation arrows and 'Found 5 result(s)'. The first result is 'Movebank', with filters for Subject(s) (Life Sciences, Biology, Zoology, Animal Ecology, Biodiversity and Conservation), Content type(s) (Raw data, Scientific and statistical data formats, Standard office documents), and Country (Germany). A description of Movebank is provided. Below this is another result for 'Movebank Data Repository' with filters for Subject(s) (Evolution, Anthropology, Zoology, Biology, Life Sciences), Content type(s) (Raw data, Scientific and statistical data formats, Standard office documents, Images, Databases), and Country (Germany). A description of this repository is also shown.

The screenshot shows the 'Databases' registry interface. At the top, it says 'Databases' and 'A registry of knowledgebases and repositories of data and other digital assets.' Below this is a search bar with 'animal behavior' and buttons for 'Clear All', 'Query string: animal behavior', 'Registry: database', and 'Record Type: repository'. There are also navigation arrows and 'Displaying 1 to 11 of 11.' The main content area shows a result for 'ModelDB'. The result includes a 'ModelDB' logo, a description: 'ModelDB provides an accessible location for storing and efficiently retrieving computational neuroscience models. A ModelDB entry contains a model's source code.', and a row of tags: 'Neurobiolo...', 'Computati...', 'Mathemati...', 'Network M...', 'All', and '+6 more tags'. Below the tags are sections for 'Standards Implemented' (1) and 'Related Databases' (2).



Desirable Characteristics

When choosing a repository to manage and share data resulting from Federally funded research, look for:

- 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

Generalist Repository Discovery

Required fields

Identifier

Best Practice: Document Search Strategy

Creators

Title

PublicationYear

Publisher

ResourceType

Datacite Metadata

Search Strategy2.docx

Page: 1 of 1 Automatic Zoom

Refine Search

Provider

- ClinicalTrials
- Research Papers
- Open Access

Real World
Youth Studies

Creative arts and writing

Environmental sciences

Indigenous studies

Law and legal studies

M

Category 1: Searched with OR

1. Cold water immersion
2. Water immersion
3. Cold immersion
4. Ice-water immersion
5. Water submersion
6. Cryotherapy
7. Cold water baths
8. Ice baths
9. Hand immersion
10. Arm immersion
11. Forearm immersion
12. Lower body immersion
13. Whole body immersion

Category 2: Searched with OR

14. Exercis*
15. Sprint*
16. Run*
17. Activity
18. Effort
19. Swim*
20. Resistance training
21. Strength training
22. Athlet*
23. Cycl*
24. Sport*
25. Competition
26. Repeated exercise
27. Repeated sprint
28. Repeated cycl*
29. Repeated bouts
30. Team

Category 3: Searched with OR

31. Performance
32. Fatigue
33. Muscle strength
34. Time trial
35. Time to completion
36. Time to exhaustion

42. Voluntary activation

43. EMG activity
44. Core temperature
45. Body temperature
46. Rectal temperature
47. Gastrointestinal temperature
48. Skin temperature
49. Cutaneous temperature
50. Shell temperature
51. Muscle temperature
52. Heart rate
53. Neuromuscular
54. Torque

Category 4: Searched with NOT

- Firefighters
- Fire fighters
- Infants
- Children
- Patients
- Animal
- Systematic Review
- Meta Analysis
- Occupation* (Title)
- Protective clothing
- Military
- NBC

All 4 Categories Searched with AND

Limits where applicable:

- Title, Abstract, Keyword search
- English
- Human
- Adult
- Academic Journal Article
- January 1990 – April 2021

Databases:

- CINHAL



Generalist Repository

Discovery

- Replication studies
- Meta-analysis
- Systematic reviews

The screenshot displays the OSF Registries interface for a project registration. The top navigation bar includes 'OSF REGISTRIES', 'Add New', 'My Registrations', 'Help', and 'Donate'. The main header shows the project title 'Investigating variation in replicability: A "Many Labs" Replication Project' and its status as a 'Public registration'. A left sidebar contains navigation links for Overview, Metadata, Files, Resources, Wiki, and Comments. The main content area is divided into sections: 'OSF-Standard Pre-Data Collection Registration', 'Data collection status' (with a note that no data collection has begun), 'Data access status' (set to 'No'), 'Contributors' (listing Richard A. Klein, Kate Ratliff, Brian A. Nosek, Michelangelo Vianello, Ronaldo Pilati, Zeynep Cemalcilar, Jesse J. Chandler, Thierry Devos, Elisa Maria Galliani, Mark Brandt, and 34 more), 'Description' (detailing the replication of 12 effects across labs), 'Registration type' (OSF-Standard Pre-Data Collection Registration), 'Date registered' (September 15, 2013), 'Date created' (June 14, 2013), and 'Associated project' (osf.io/wx7ck). A bottom sidebar titled 'Open practice resources' lists categories like Data, Analytic code, Materials, Papers, and Supplements, with the 'Data' category highlighted by a red circle.



Evaluate Datasets

● Metadata Quality

○ FAIR

- Findable
- Accessible
- Interoperable
- Reusable

● Documentation Quality

- Who, what, when, where, why, how

● Reuse indicators

- Metrics
- Citations

zenodo Search Upload Communities Log in Sign up

July 16, 2021 Software Open Access

biowdl/structural-variantcalling: Release 1.2.0

Cedrick Agaser; Ruben Vorderman; Davy Cats; Jasper

- Remove GRIDSS from the pipeline.
- Exclude GRIDSS from SURVIVOR merging; SVs were only defined as BNDs in GRIDSS.
- Optional filtering of missing and hom-ref genotypes.
- Optional filtering of FP deletions and duplications.
- Add DUPHOLD: annotate SVs with depth values.
- Make bcftools indexing optional.
- Structural-variantcalling pipeline: add sorting and change id.
- Added GRIDSS sv caller

37 views 2 downloads See more details...

Available in **GitHub** Indexed in **OpenAIRE**

Publication date: July 16, 2021
DOI: [10.5281/zenodo.5109918](https://doi.org/10.5281/zenodo.5109918)
Related identifiers: Supplement to <https://github.com/biowdl/structural-variantcalling/tree/v1.2.0>
License (for files): [Other \(Open\)](#)

Versions

Version	Created
Version v1.2.0 10.5281/zenodo.5109918	Jul 16, 2021
Version v1.1.0 10.5281/zenodo.3974344	Aug 6, 2020
Version v1.0.0 10.5281/zenodo.3734347	Mar 31, 2020

Cite all versions? You can cite all versions by using the DOI [10.5281/zenodo.3734346](https://doi.org/10.5281/zenodo.3734346). This DOI represents all versions.

Preview

structural-variantcalling-v1.2.0.zip

- biowdl-structural-variantcalling-494463a
 - .github
 - PULL_REQUEST_TEMPLATE.md 115 Bytes
 - workflows
 - ci.yml 1.8 kB
 - .gitignore 97 Bytes
 - .gitmodules 160 Bytes
 - CHANGELOG.md 1.1 kB
 - LICENSE 1.1 kB
 - README.md 849 Bytes
 - VERSION 6 Bytes
 - docs
 - index.md 3.5 kB
 - inputs.md 20.2 kB
 - requirements-test.txt 393 Bytes
 - scripts
 - structural-variantcalling.wdl 9.3 kB
 - tasks

Files (4.5 MB)

Name	Size	Preview	Download
biowdl/structural-variantcalling-v1.2.0.zip	4.5 MB	<input type="checkbox"/>	<input type="checkbox"/>
md5:02a039ef54ba7069cde4bb33f8c38d1			

Citations

Show only: Literature (0) Dataset (0) Software (0) Unknown (0)
 Citations to this version

Search

No citations.



Select Data

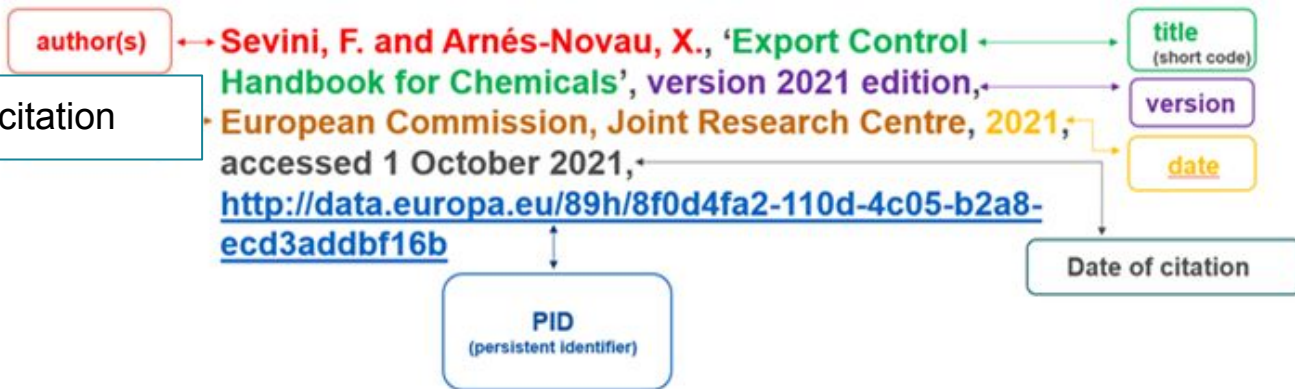
```
.  
|-- CITATION  
|-- README  
|-- LICENSE  
|-- requirements.txt
```

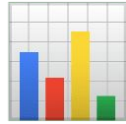
Best Practice: Data citation

```
-- doc  
|  -- notebook.md  
|  -- manuscript.m  
|  -- changelog.tx  
|-- results  
|  -- summarized_r  
|-- src  
|  -- sightings_an  
|  -- runall.py
```

Completeness: data|docur

Data citation





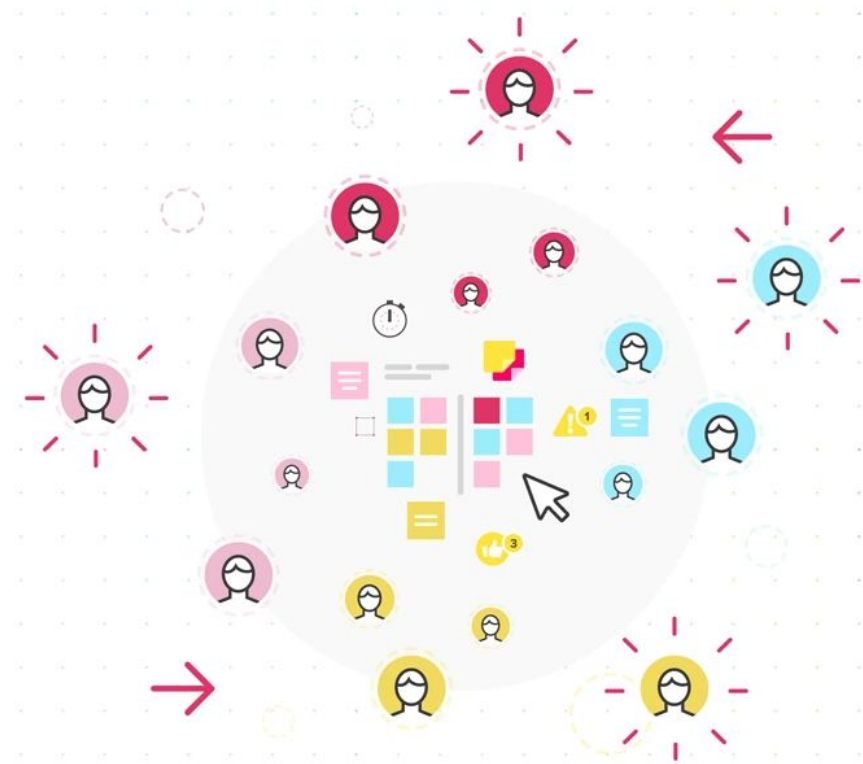
Poll Question

What are your biggest challenges in finding, or helping others to find data for reuse?



Breakout Sessions Part 1

- Everyone will randomly be assigned a 'room' to join.
- Facilitators will ask questions for your feedback and/or you can pose your own questions for discussion.
- We will use mural boards to capture thoughts/ideas.
- After 25 minutes, we will regroup in the main meeting room.



We welcome you to participate if you feel comfortable. But also come to just listen and learn too.



Breakout Sessions Part 1

Breakout session questions:

1. What are your on-the-ground experiences with repository support?
2. What challenges do you have with finding data?



10 Minute Break!



Scenario 3: Data Sharing in Multiple Repositories



Whenever possible, researchers should deposit data in **discipline-specific** or **data-type specific** repositories.

Otherwise, they should use a trusted generalist repository.

What happens when a project has both?



Specialist repositories

Optimized for disciplinary needs

Enhance discovery and reuse by using discipline-specific metadata

Support niche file formats

May support embedded visualization or analysis

Serve as “community hubs”

Generalist repositories

Accommodate heterogeneous data

Provide a home for data that might not have a place elsewhere

Facilitate broad and serendipitous discovery

Expands potential audience for data

May support custom metadata and linking to related content managed elsewhere

May be free to use (within limits)



When multiple repositories may be appropriate

- **When producing heterogeneous data types or formats**
 - Are there subsets of data that belong in an accepted specialist repository due to their file type or subject matter?
 - Are there funder requirements governing where data should be deposited?
- **When there are security and confidentiality considerations**
 - If raw data is too sensitive for broad sharing, are there subsets or processed datasets that could be openly distributed?



As a researcher, my study uses **genomic data** in combination with **landscape, dispersal, and occupancy data**, to inform [conservation unit] CU delineation in Nevada populations of the Great Basin Distinct Population Segment of the Columbia spotted frog (*Rana luteiventris*).

Forester, Brenna et al. (2022), Genomics-informed delineation of conservation units in a desert amphibian, Dryad, Dataset, <https://doi.org/10.5061/dryad.w6m905qqn>



Good practices

- Identify appropriate repositories
- Indicate use of multiple repositories in your data management plan
- Build connections and provide context



Identify appropriate repositories

[Supplemental Information to the NIH Policy for Data Management and Sharing: Selecting a Repository for Data Resulting from NIH-Supported Research](#)

[DataCite Repository Finder](#)

[Re3data](#)

[FAIRsharing](#)



Raw sequencing data



National Library of Medicine
National Center for Biotechnology Information

BioProject

A BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data types generated for that project.

Filtered data

(Variant Call Format
[VCF]) and metadata
(text, TSV)



DRYAD



Indicate use of multiple repositories in your DMP

Element 4: Data Preservation, Access, and Associated Timelines

A. Repository where scientific data and metadata will be archived:

Raw, demultiplexed sequencing data will be made available on the NCBI Sequence Read Archive.

Filtered VCFs and metadata will be made available via Dryad.

Here's an example of an [NIH Data Management plan](#) that includes multiple repositories.



Build connections and provide context

- Add persistent identifiers (PIDs) for related datasets (and other outputs)
- Describe how and why data has been divided between repositories
- Strategically apply metadata



Genomics-informed delineation of conservation units in a desert amphibian

Forester, Brenna, Colorado State University,  <https://orcid.org/0000-0002-1608-1904>

Murphy, Melanie, University of Wyoming

Mellison, Chad, United States Fish and Wildlife Service

Petersen, Jeffrey, Nevada Department of Wildlife

Pilliod, David, United States Geological Survey

Van Horne, Rachel, US Forest Service

Harvey, Jim, US Forest Service

Funk, W. Chris, Colorado State University,  <https://orcid.org/0000-0002-6466-3618>

brenna.forester@colostate.edu, chris.funk@colostate.edu

Publication date: August 25, 2022

Publisher: Dryad

<https://doi.org/10.5061/dryad.w6m905qqn>

Data files

 Download dataset

> August 25, 2022

** changes not displayed to the public*

Related works

Article

<https://doi.org/10.1111/mec.16660>

Dataset

<https://www.ncbi.nlm....oproject/PRJNA869693>



ORIGINAL ARTICLE |  Open Access |  

Genomics-informed delineation of conservation units in a desert amphibian

Brenna R. Forester  Melanie Murphy, Chad Mellison, Jeffrey Petersen, David S. Pilliod, Rachel Van Horne, Jim Harvey, W. Chris Funk

First published: 17 August 2022 | <https://doi.org/10.1111/mec.16660> | Citations: 1

Open Research

DATA AVAILABILITY STATEMENT

Raw, demultiplexed sequencing data are available on the NCBI Sequence Read Archive under BioProject PRJNA869693: <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA869693> . Filtered VCFs and metadata are available on Dryad (Forester et al., [2022](#)): <https://doi.org/10.5061/dryad.w6m905qqn> .



Keywords

Biological sciences
Adaptive differentiation
conservation genomics
conservation units
double digest RADseq (ddRADseq)
evolutionarily significant units
genetic rescue
management units
Rana luteiventris

Keywords

adaptive differentiation
conservation genomics
evolutionarily significant units
genetic rescue management units
Rana luteiventris

Organism

Rana luteiventris [Taxonomy ID: 58176]

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Amphibia; Batrachia;
Anura; Neobatrachia; Ranoidea; Ranidae; Rana; Rana; Rana luteiventris



Scenario 4: Budgeting for Generalist Repositories



The Dataverse Project



Researchers

Enjoy full control over your data. Receive *web visibility, academic credit, and increased citation counts*. A personal Dataverse collection 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 collection?*



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 a Dataverse collection as part of your journal data policy or list of repository recommendations. *Want to find out more about journal Dataverse collections?*



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 the Dataverse Project. Want to contribute?*

95 Installations around the world



Open source research data
repository software

<https://dataverse.org>



Harvard Dataverse

Use Cases Supported

All disciplines, journals, organizations, institutions, labs, teaching courses, research projects, researchers

What types of outputs can be shared?

- **Computationally usable data** (*required*)
- **File size limits:**
 - 2.5GB per file
 - 1T overall project size
- **File types:**
 - Data output required
 - Other types of research outputs related to deposited data
- **License type:**
 - Default CC0
 - Custom license
 - Support for multiple licenses
- **Human subjects data:**
 - Only deidentified data accepted
 - Planned: support via OpenDP, DataTags



Harvard Dataverse

Deposit Workflow

- **Self Curated repository** with guided support
 - Self deposit
 - Self publish
 - Controlled workflow and publishing for collections

Key Deposit Requirements

- >1 dataset required for “collection creation”
- **Required metadata:** Title, Author, Description, Subject, Contact
- **Data files**
- **Accessibility compliance** for restricted content

Services Offered

- Free consultation and support (3 hours) before and after deposit
- [Paid Curation Services](#)



Budgeting Considerations: NIH Perspective

- Curating data
- Developing supporting documentation
- Formatting data according to community standards or repository
- De-identifying data
- Preparing metadata to foster discoverability, interpretation, and reuse
- Local data management considerations
- Preserving and sharing data through established repositories





NIH Data Management and Sharing Plan as an Indicator of Potential Costs

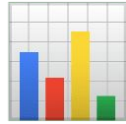
1. **Data Types:** Costs for data collection, acquisition, storage
2. **Related Tools, Software, Code:** Costs for data access, security
3. **Common Data Standards:** Costs for data documentation
4. **Data Preservation, Access, Timelines:** Costs for data preservation
5. **Access, Distribution, Reuse:** Costs for data availability, reuse
6. **Oversight of Data Management:** Costs for operationalizing



Questions to Ask & What to Include

- Is there a deposit fee for any of the repositories you plan to use?
- Will you need dedicated personnel time or will you engage the services of a core or vendor?
 - Data curation?
 - Developing supporting documentation?
 - Formatting data according to accepted community standards or for transmission and storage at a selected repository?
 - Preparing metadata?
 - De-identifying data?
- Do you have subrecipients?
 - Who is responsible and need funds for data management and sharing activities?
 - Will the data be sent to the prime and combined into a single data set?
 - Will the prime deposit the data or the subrecipient?





Poll Question

Do you plan to leverage generalist repositories as part of supporting data management and sharing plans?



Repository Costs

HARVARD
Database Support

Curator Dashboard | Curator Guides | Events | Profiles & Governance | Data Harvest/Export

Curator and Data Management Services

Overview | Frequently asked questions | Support | Contact Us

Overview

The Harvard Databases Repository provides Research Curators with a broad range of support services including data ingest, data curation, data management, data visualization, data analysis, data sharing, and data archiving. The Harvard Databases Repository also provides support for the Harvard Databases Repository. Our curators are available to assist you with data ingest, data curation, data management, data visualization, data analysis, data sharing, and data archiving. We are committed to providing you with the highest quality of service and support.

Request Help Please contact a Harvard Databases Repository Curator, your affiliation, and provide a description of the service you need to your manager email address.

Harvard Databases Repository administration and service services

Service Area	Services	Service Description	Cost	Availability
Database Administration	Database Setup & Configuration Database Migration Database Backup & Recovery Database Performance Tuning Database Security Audits Database Patching & Updates	Installation and configuration of various database systems (e.g., PostgreSQL, MySQL, Oracle, Microsoft SQL Server, etc.) Migration of data from one database system to another Regular backups and recovery testing to ensure data integrity Performance monitoring and optimization Security audits and patching to maintain system security	\$1000 - \$5000 (one-time fee)	On-site or Remote
Database Support	Database Troubleshooting Database Query Optimization Database Error Resolution Database Configuration Assistance	Diagnosis and resolution of database-related issues Optimization of database queries for better performance Resolution of database errors and warnings Assistance with database configuration and settings	\$500 - \$2000 (hourly rate)	On-site or Remote
Database Migration	Database Migration Planning Database Migration Execution Database Migration Testing Database Migration Documentation	Planning and execution of database migration projects Data transfer and validation during migration Testing of migrated data and system functionality Documentation of migration process and results	\$1000 - \$3000 (one-time fee)	On-site or Remote
Database Security	Database Security Audits Database Security Patching Database Security Configuration Database Security Monitoring	Regular security audits and vulnerability assessments Application of security patches and updates Configuration of database security settings Monitoring of database activity for suspicious behavior	\$500 - \$2000 (monthly fee)	On-site or Remote

DRYAD

Home | About | Contact | Help

Frequently asked questions

Why type of data does Dryad accept?

Department of Energy, Environmental Sciences and Earth Sciences (EES&E) researchers can submit data to Dryad. Dryad is a public repository for research data. It is a non-profit organization that provides a secure and reliable way to store and share your research data. Dryad is a public repository for research data. It is a non-profit organization that provides a secure and reliable way to store and share your research data. Dryad is a public repository for research data. It is a non-profit organization that provides a secure and reliable way to store and share your research data.

What can the data be used for?

There is no restriction on the use of data deposited in Dryad. The data can be used for any purpose, including research, teaching, and public access. The data is stored in a secure and reliable way, and is available to anyone who wants to access it. The data is also available in a format that is easy to use and understand. The data is also available in a format that is easy to use and understand. The data is also available in a format that is easy to use and understand.

figshare

Home | About | Contact | Help

Using Figshare+ for easily sharing large image file databases and tracking impact

Figshare+ is a leading research data repository for sharing large image file databases and tracking impact. It is a non-profit organization that provides a secure and reliable way to store and share your research data. Figshare+ is a leading research data repository for sharing large image file databases and tracking impact. It is a non-profit organization that provides a secure and reliable way to store and share your research data. Figshare+ is a leading research data repository for sharing large image file databases and tracking impact. It is a non-profit organization that provides a secure and reliable way to store and share your research data.

Country	Downloads	Views	Downloads	Views	Downloads
USA	1000	2000	1000	2000	1000
UK	500	1000	500	1000	500
Canada	200	400	200	400	200
Germany	100	200	100	200	100
France	50	100	50	100	50
Spain	20	40	20	40	20
Italy	10	20	10	20	10
Japan	5	10	5	10	5
China	2	4	2	4	2
India	1	2	1	2	1
Australia	1	2	1	2	1
South Africa	1	2	1	2	1
Other	1	2	1	2	1
Total	1870	3740	1870	3740	1870

digitalcommons

Home | About | Contact | Help

Digital Commons Data: for your institution's RDM Journey

Digital Commons Data is a leading research data repository for sharing large image file databases and tracking impact. It is a non-profit organization that provides a secure and reliable way to store and share your research data. Digital Commons Data is a leading research data repository for sharing large image file databases and tracking impact. It is a non-profit organization that provides a secure and reliable way to store and share your research data. Digital Commons Data is a leading research data repository for sharing large image file databases and tracking impact. It is a non-profit organization that provides a secure and reliable way to store and share your research data.

CDS-DO

Home | About | Contact | Help

How to use the OSF usage tiers

Overview | Frequently asked questions | Support | Contact Us

Green and purple budgets

Use OSF usage tiers to track and manage your research data storage and processing costs. The Green tier is for researchers who are just starting to use OSF, and the Purple tier is for researchers who are using OSF extensively. The Green tier is for researchers who are just starting to use OSF, and the Purple tier is for researchers who are using OSF extensively. The Green tier is for researchers who are just starting to use OSF, and the Purple tier is for researchers who are using OSF extensively.

Additional OSF Storage

Research projects that require more than 100 GB of OSF storage can purchase additional storage capacity. The additional storage is available in a format that is easy to use and understand. The additional storage is available in a format that is easy to use and understand. The additional storage is available in a format that is easy to use and understand.

Include OSF in your repository in funding proposals and budget requests

1. Identify the OSF usage tier that best fits your research project.
2. Estimate the storage and processing costs for your research project.
3. Include the OSF usage tier and associated costs in your funding proposal and budget request.
4. Monitor your research data storage and processing costs as you use OSF.

Usage Tier	Storage	Processing	Cost
Green	100 GB	100 GB	\$1000
Purple	1000 GB	1000 GB	\$10000
Gold	10000 GB	10000 GB	\$100000
Platinum	100000 GB	100000 GB	\$1000000

Vivli

Home | About | Contact | Help

Share Your Data on the Vivli Platform

Vivli is a leading research data repository for sharing large image file databases and tracking impact. It is a non-profit organization that provides a secure and reliable way to store and share your research data. Vivli is a leading research data repository for sharing large image file databases and tracking impact. It is a non-profit organization that provides a secure and reliable way to store and share your research data. Vivli is a leading research data repository for sharing large image file databases and tracking impact. It is a non-profit organization that provides a secure and reliable way to store and share your research data.

Why share your data using Vivli?

- It is a non-profit organization that provides a secure and reliable way to store and share your research data.
- It is a leading research data repository for sharing large image file databases and tracking impact.
- It is a non-profit organization that provides a secure and reliable way to store and share your research data.

Steps to sharing your research data on Vivli

1. Create a Vivli account and verify your email address.
2. Upload your research data to the Vivli platform.
3. Share your research data with your colleagues and the public.
4. Track the impact of your research data on the Vivli platform.

How much does it cost to store my data on Vivli and have it available for research for others to use (FAIR)?

There is no cost to store your research data on the Vivli platform. The cost of storing your research data on the Vivli platform is covered by the Vivli organization. The cost of storing your research data on the Vivli platform is covered by the Vivli organization. The cost of storing your research data on the Vivli platform is covered by the Vivli organization.

zenodo

Home | About | Contact | Help

Terms of Use v1.2

Zenodo is a leading research data repository for sharing large image file databases and tracking impact. It is a non-profit organization that provides a secure and reliable way to store and share your research data. Zenodo is a leading research data repository for sharing large image file databases and tracking impact. It is a non-profit organization that provides a secure and reliable way to store and share your research data. Zenodo is a leading research data repository for sharing large image file databases and tracking impact. It is a non-profit organization that provides a secure and reliable way to store and share your research data.

1. Zenodo is a non-profit organization that provides a secure and reliable way to store and share your research data.
2. Zenodo is a leading research data repository for sharing large image file databases and tracking impact.
3. Zenodo is a non-profit organization that provides a secure and reliable way to store and share your research data.
4. Zenodo is a leading research data repository for sharing large image file databases and tracking impact.
5. Zenodo is a non-profit organization that provides a secure and reliable way to store and share your research data.
6. Zenodo is a leading research data repository for sharing large image file databases and tracking impact.
7. Zenodo is a non-profit organization that provides a secure and reliable way to store and share your research data.
8. Zenodo is a leading research data repository for sharing large image file databases and tracking impact.
9. Zenodo is a non-profit organization that provides a secure and reliable way to store and share your research data.
10. Zenodo is a leading research data repository for sharing large image file databases and tracking impact.



Harvard Dataverse Data Curation Services

Service Name	Description	Service Components	Costs	
<i>Data Curation Services</i>			Harvard	Non-Harvard
Free consultation and assessment	Triage level + Office hours Follow-up with recipients of support letters and consultation clients	Demonstrations of software Consult on projects size and scope and fit for Harvard Dataverse	Up to 3 hours	Up to 2 hours
Extended consultation services		<ul style="list-style-type: none"> • Demonstrations of software • Deidentification consult • Dataverse organization consult • File organization consult • Replication verification consult 	\$100/hour	\$200/hour
Dataverse set-up services and dataset and data file ingest	Establish data curation infrastructure and organization enabling data owners to self-curate their data.	<ul style="list-style-type: none"> • Extended Consultation • Dataverse creation and customization • Dataset creation • Data file ingest • Metadata enrichment of dataset and data files • Documentation for use of final product 	\$2000 per "collection" plus \$100 minimum per dataset	\$4000 per "collection" plus \$200 minimum per dataset
Ongoing dataverse administration and curation services	Ongoing maintenance of Dataverse and datasets Continuous updates of dataverse and datasets as needed to reflect software upgrades and sharing standards	<ul style="list-style-type: none"> • Extended consultation services • Dataverse set-up services and dataset and data file ingest • Maintenance of dataverses and datasets within Harvard Dataverse 	\$10,000/year plus \$100 minimum per dataset	\$20,000/year plus \$200 minimum per datasets
Custom services for existing dataverse	One time curation services This includes: <ul style="list-style-type: none"> • file-upload • metadata enhancement (bi-directional linking), • linking affiliated datasets • Documentation • Digitization of audio/paper/content for sharing purposes 		\$100/hour	\$200/hour

<https://support.dataverse.harvard.edu/curation-services>



Sharing in Harvard Dataverse Scenario

Example Project

- Calcium imaging experiments in mice and drosophila
- Sequential captures at ~30 frames per second for 4-5 hours
- Resulting raw data that spans 40-300 GB

Collection & Data sets

- Time series imaging data
- 1 collection
- 5 data sets
- 2+ TB data
- Metadata based on Minimum Information about Tissue Imaging
- Analytic data

Project Goals & Needs

- Share data from an imaging centered publication with multiple terabytes of data that need to become publicly available
- Create custom metadata blocks for calcium time series images
- Maintain the collection for the time length of the grant period, potentially migrating files to new formats as they are developed



Budgeting in Harvard Dataverse

Category: Data Collection

Activity: Identifying data for deposit which requires additional storage space beyond 1TB.

Costs: \$100 **

- 1TB no cost
- Estimating \$500 per additional TB

*** This is representing a rough estimate that will depend on multiple factors like repository platform, vendor terms, data types, and ...*

Category: Data Documentation

Activity: Create enhanced metadata templates and plan for upload of files in user friendly formats with enhanced file metadata.

Costs: \$500

- \$100 per 5 hours extended consultation



Budgeting in Harvard Dataverse

Category: Data Sharing & Reuse

Activity: Establish data curation infrastructure and organization enabling data owners to self-curate their data.

Costs: \$2,500

- \$2000 per collection
- \$100 per 5 datasets

Category: Data Curation

Activity: Continuous updates of dataverse and datasets as needed to reflect software upgrades and sharing standards.

Costs: \$50,500

- \$10,000/year for 5 years
- \$100 per 5 dataset

Hypothetical 5 Year Project Budget: \$54,000



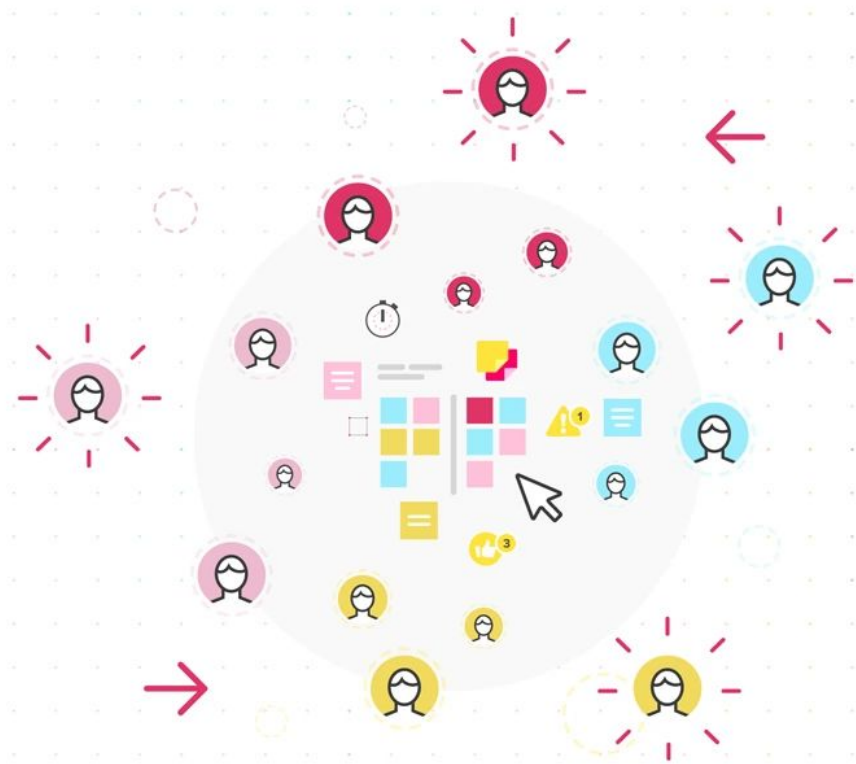
Additional Resources

- NIH [Allowable Costs for Data Management and Sharing](#)
- COGR [Project-Based and Institutional Cost Considerations: Budgeting & Costing](#)
- National Academies' report on [Life-Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs](#)
- OpenAIRE (EU) [Estimating Costs RDM Tool](#)
- Mons, Barend. 2020. "[Invest 5% of research funds in ensuring data are reusable](#)." *Nature* 578(7796): 491-491.
- Perry, Anja, and Sebastian Netscher. 2022. "[Measuring the time spent on data curation](#)." *Journal of Documentation* 78(7).



Breakout Sessions Part 2

- Everyone will randomly be assigned a 'room' to join.
- Facilitators will ask questions for your feedback and/or you can pose your own questions for discussion.
- We will use mural boards to capture thoughts/ideas.
- After 25 minutes, we will regroup in the main meeting room.



We welcome you to participate if you feel comfortable. But also come to just listen and learn too.



Breakout Sessions Part 2

Breakout session questions:

1. How do you talk to researchers about budgeting for the use of repositories?
2. Do you think you (or researchers you support) are more likely to share your data, knowing that you can see that your data is being used and cited?
3. How can GREI better support data librarians in using generalist repositories?



Wrap-up and Close





Poll Questions

Do you feel more prepared to support the use of generalist repositories?

Overall, was this workshop helpful?

Did it meet your expectations and reasons for attending?



Final Reminders

- Thank you for your participation!
- Workshop slides available at: <https://doi.org/10.5281/zenodo.7774200>
- Generalist Repository Comparison Chart - version 3 coming soon!
<https://doi.org/10.5281/zenodo.3946719>
- Stay in touch with the GREI repositories - ask questions, provide feedback, get updates, learn about future events:
 - Join the GREI Forum: <https://groups.google.com/g/contactgrei>
 - GREI Email: contactgrei@googlegroups.com
 - Public outputs in GREI Zenodo Community: <https://zenodo.org/communities/grei>





GREI

Thank you for participating!

