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

MENDELEY DATA

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

figshare

Andrew Mckenna-Foster
Product Specialist,
Figshare

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

The Dataverse® Project

DRYAD

OSF

Poll Question

Tell us why you are here!

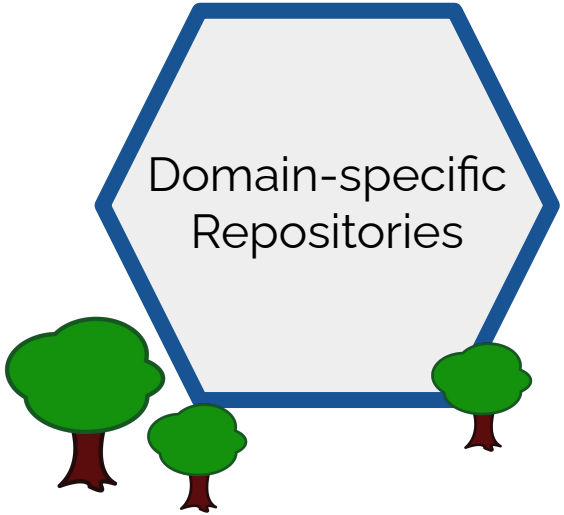
Workshop Outline

- 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



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

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.

doi: 10.5281/zenodo.3946720

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: • researchers to publish datasets utlizing 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 managment 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 reserachers 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

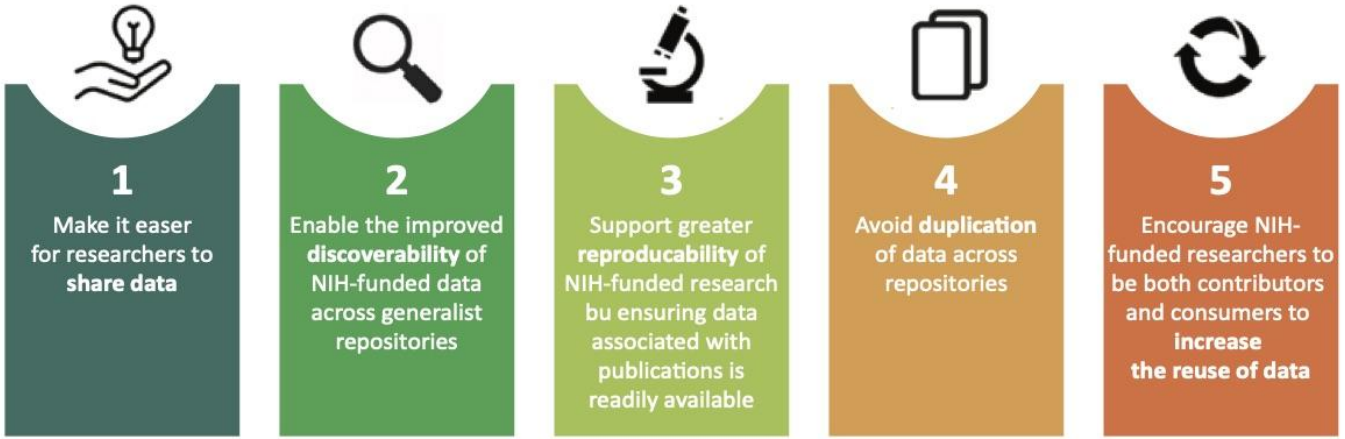
Introduction to the Generalist Repository Ecosystem Initiative (GREI)

NIH Generalist Repository Ecosystem Initiative

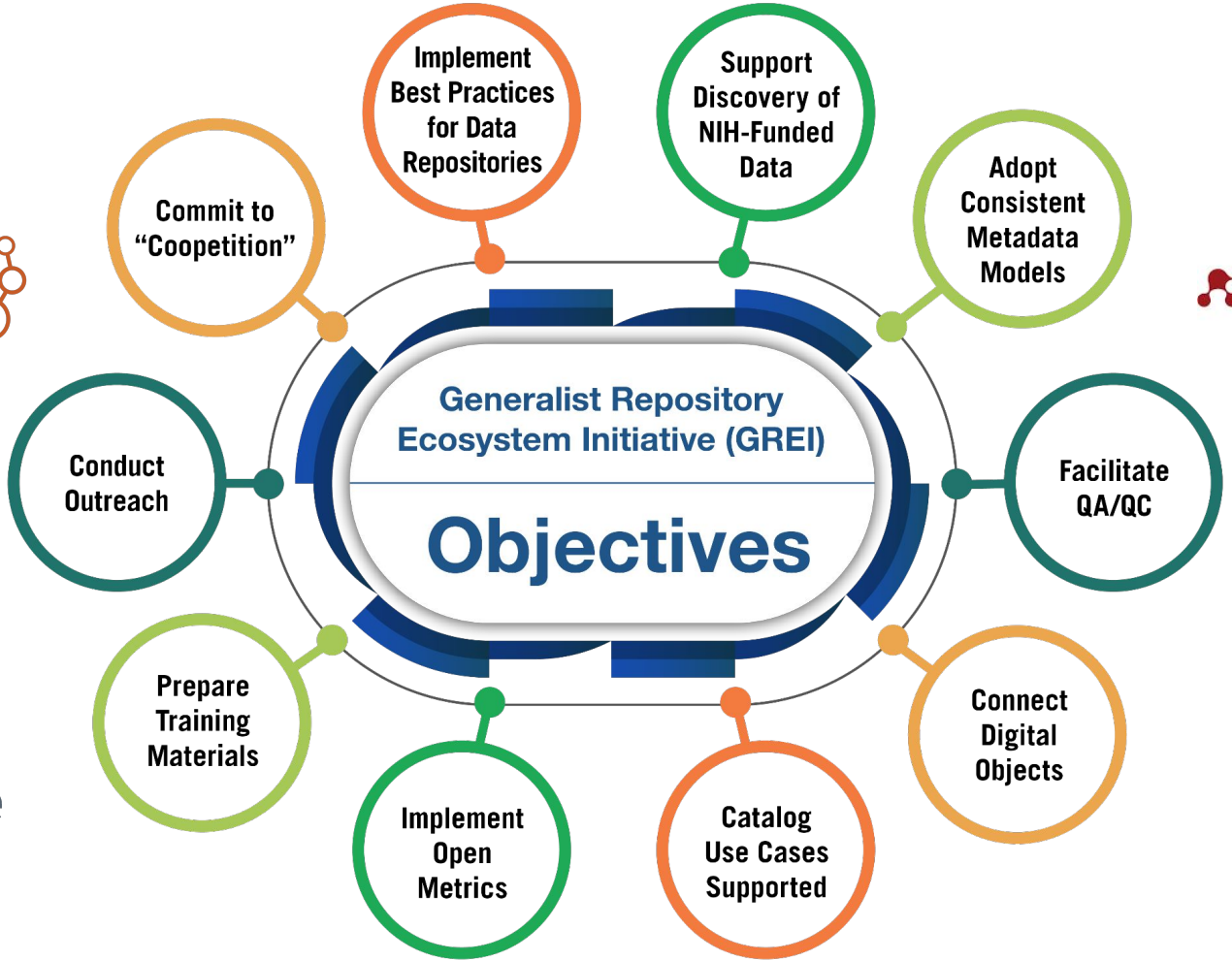
The mission of GREI is to establish a common set of capabilities, services, metrics, and social infrastructure; raise general awareness and facilitate researchers to adopt FAIR principles to better share and reuse data.

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 Inititative

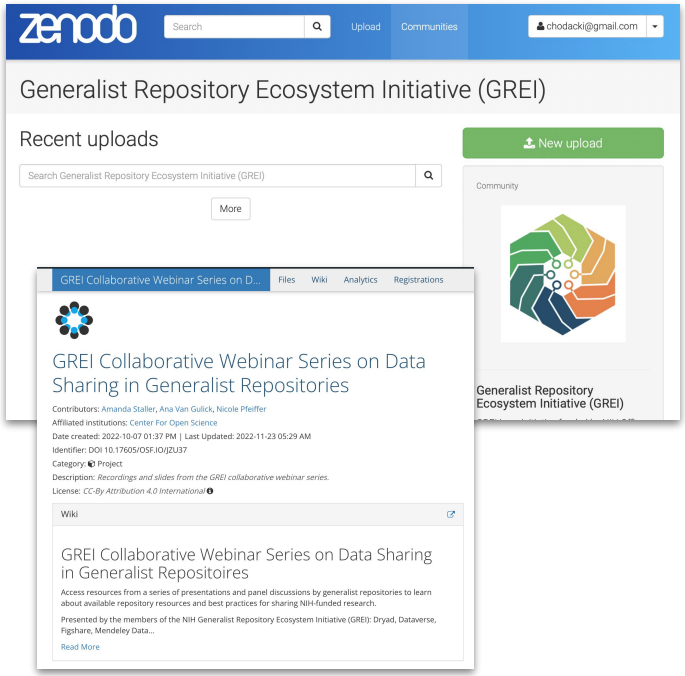


https://doi.org/10.6084/m9.figshare.21318270



https://datascience.nih.gov/data-ecosystem/exploring-a-generalist-repository-for-nih-funded-data

Connect with GREI



- 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

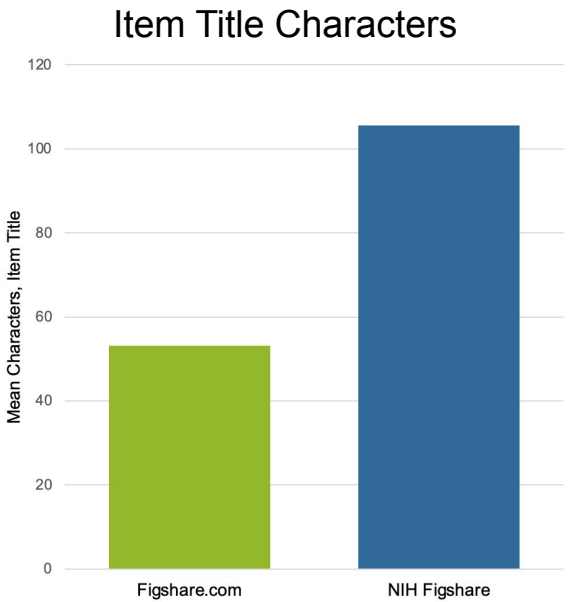
Data Management

Data Sharing

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

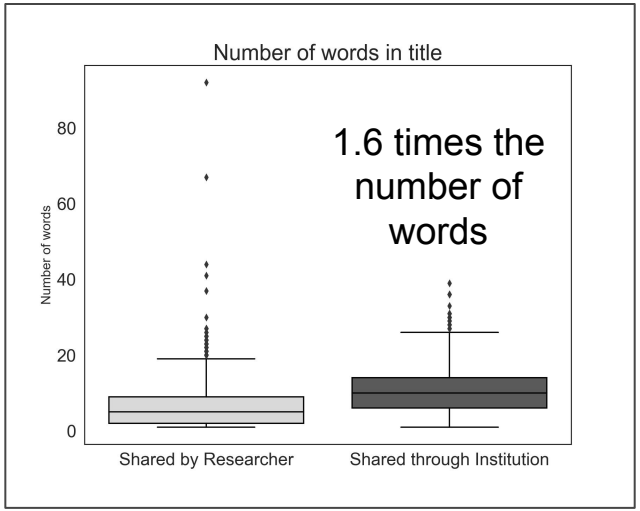
- Shorter titles, descriptions, fewer keywords, fewer related material links
- These are easy things for a librarian or curator to help with and significantly increase the findability and interoperability of a dataset

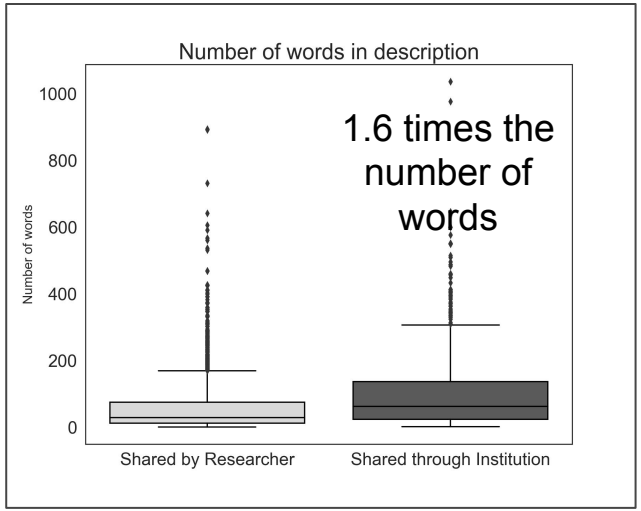
NIH Figshare items had on average 2x and 3.5x more characters in the item title and description respectively

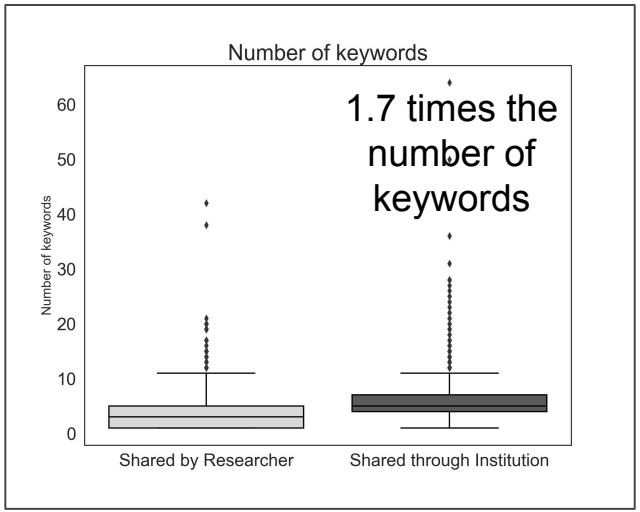


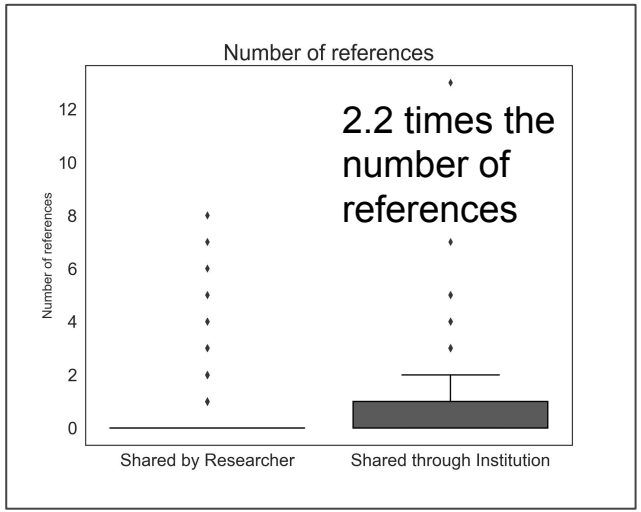
Records from researchers
n = 2,470 records

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

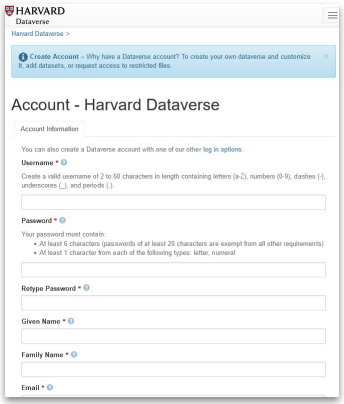
- The dataset should stand on its own from the paper
- The dataset should be reusable on its own
- The dataset should be linked to related materials via PIDs - link to papers, funding sources, author IDs, institution IDs

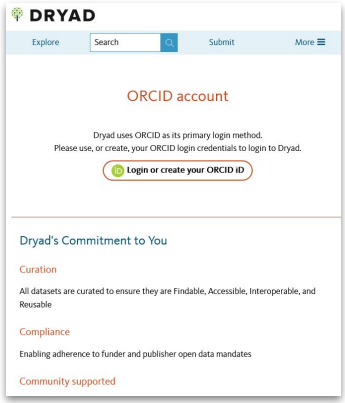
Tracking impact and getting credit

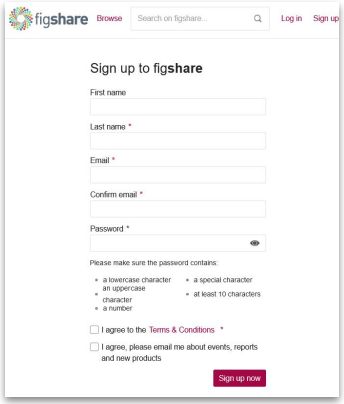
- Have more impact with open data that can be found and reused
- Meet funder requirements, report to NIH
- Track citations, metrics
- Reflected on researcher profiles (e.g. ORCID)

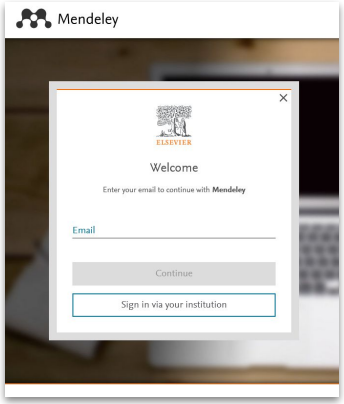
Getting started to support data deposit

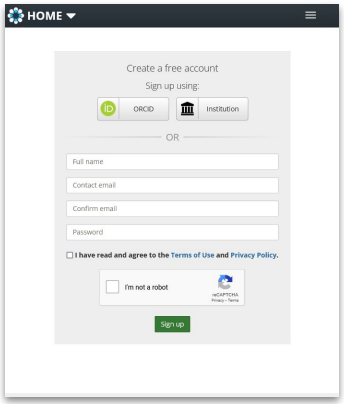
Sign up for an account

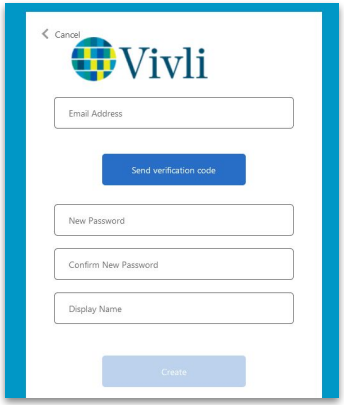


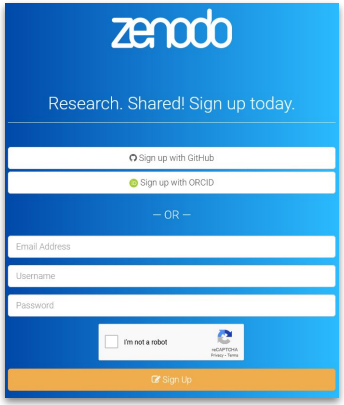




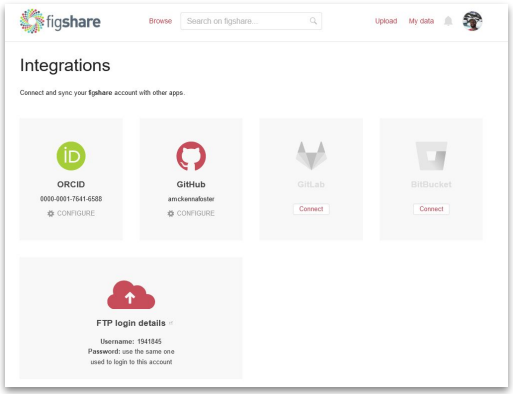


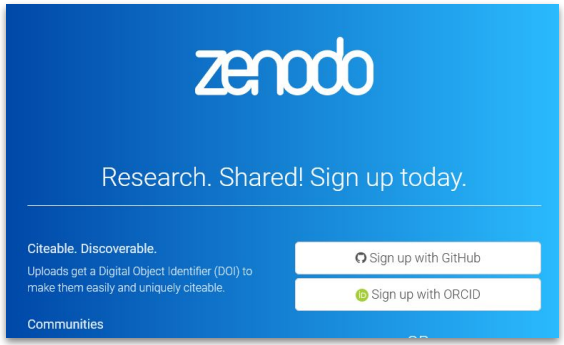


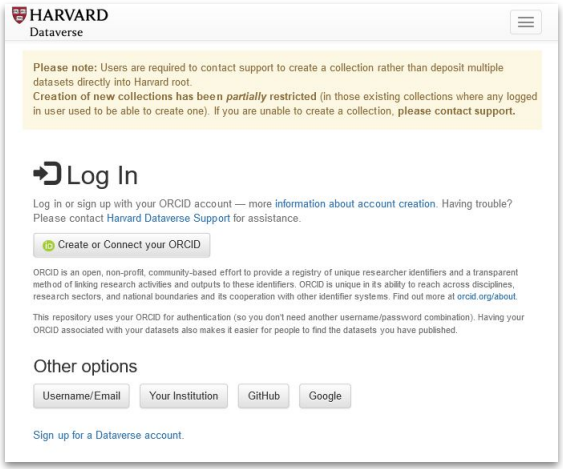




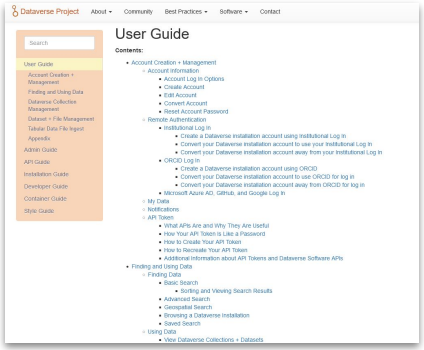
Connect account to ORCID

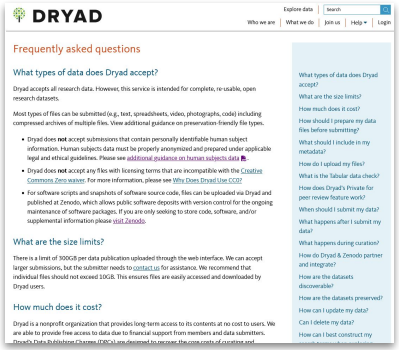


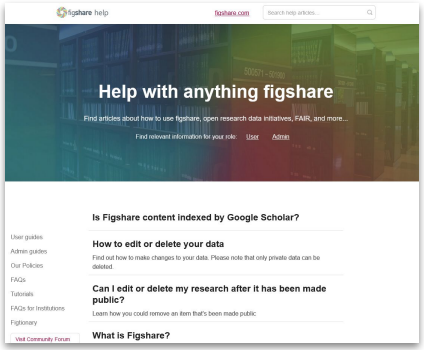


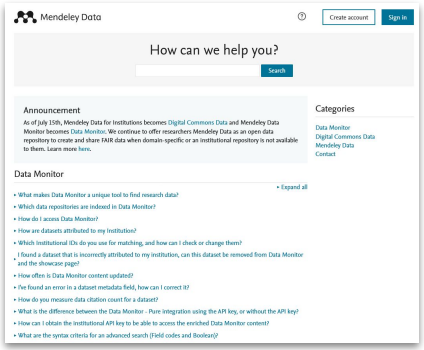


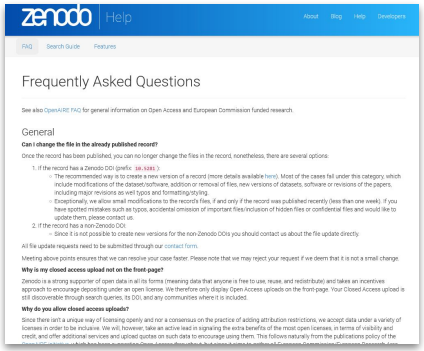
Know about the help resources



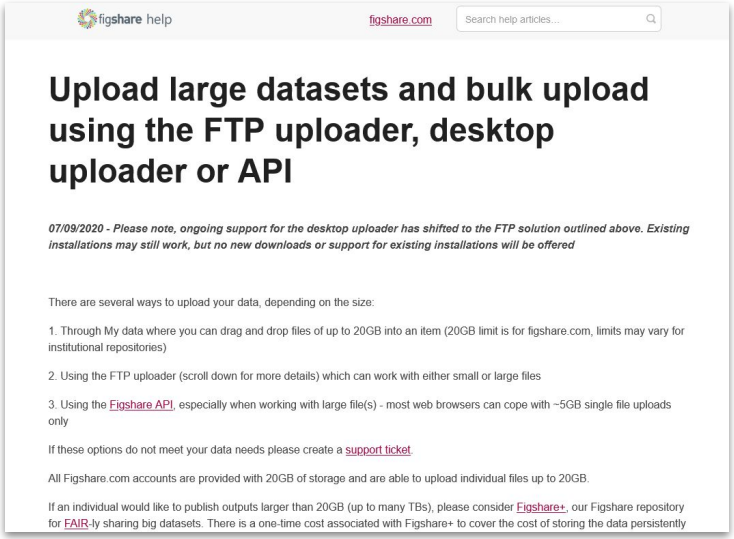


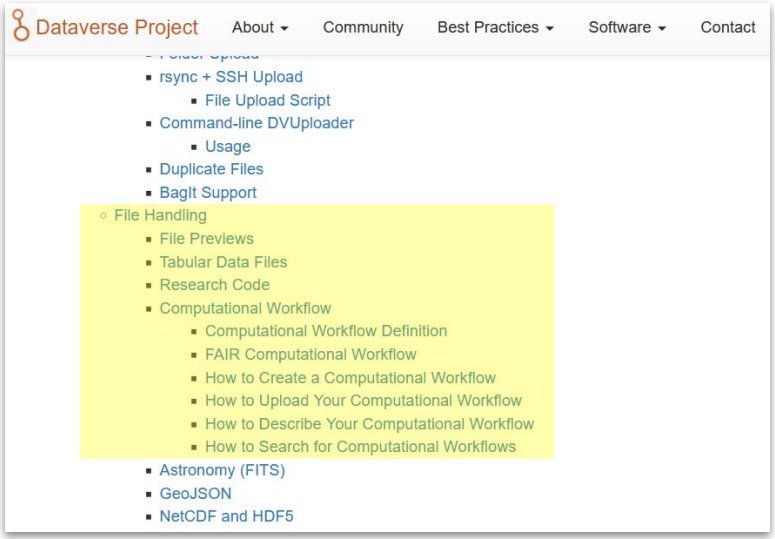


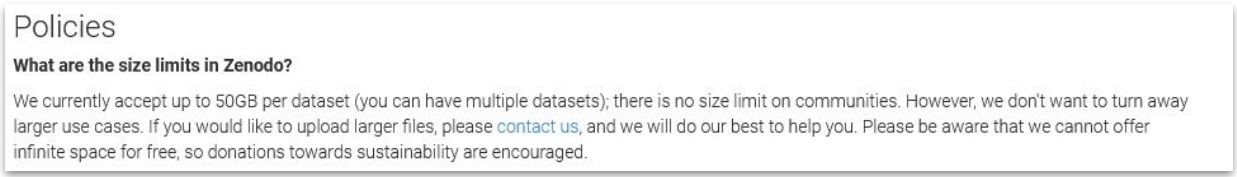




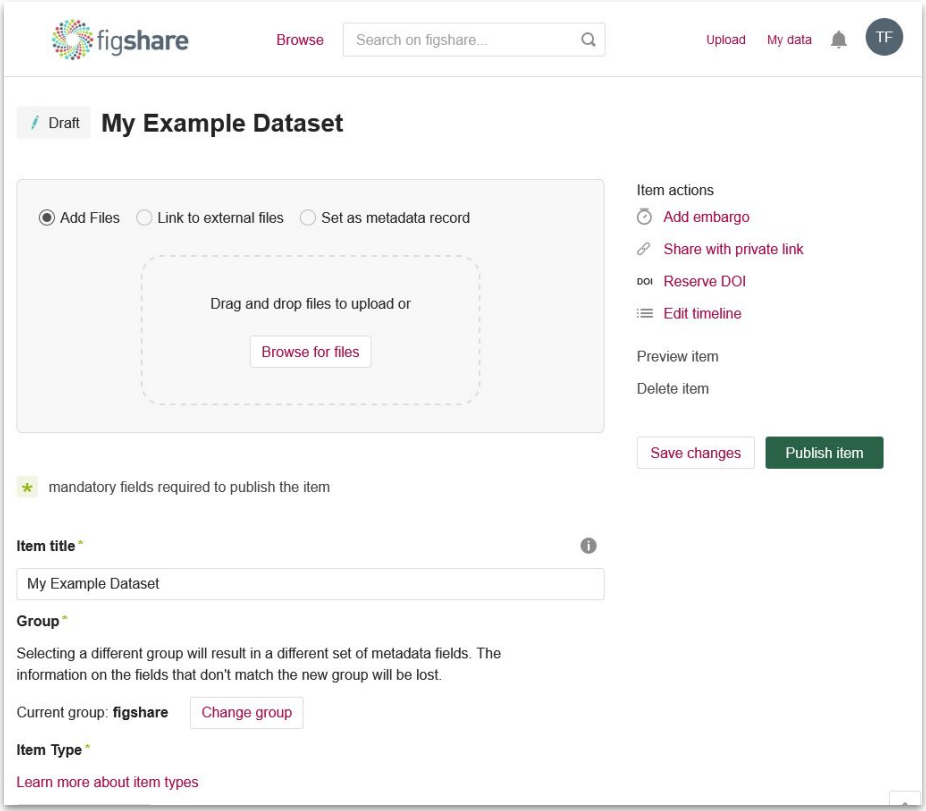
Know about alternative upload options

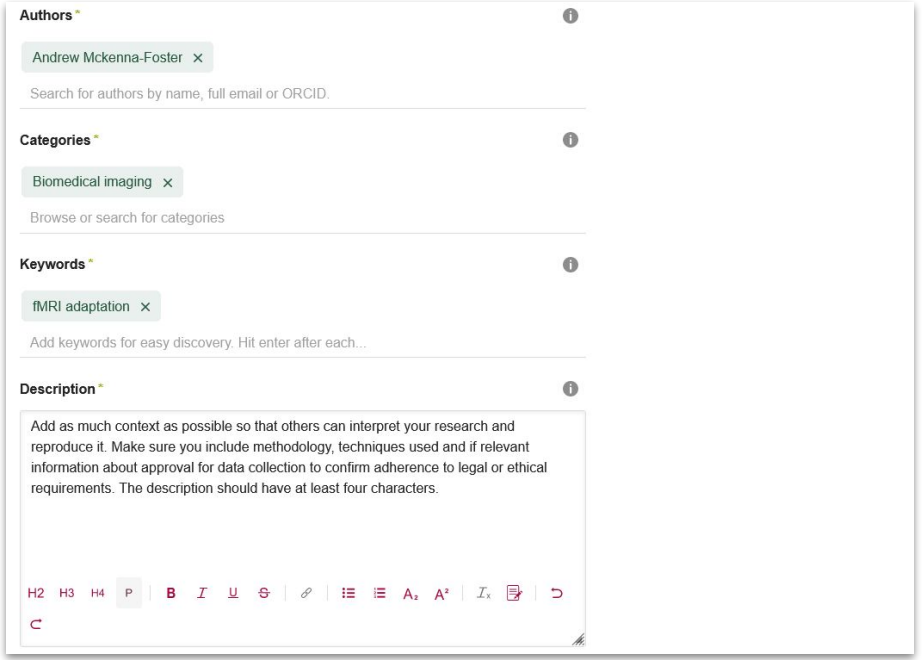




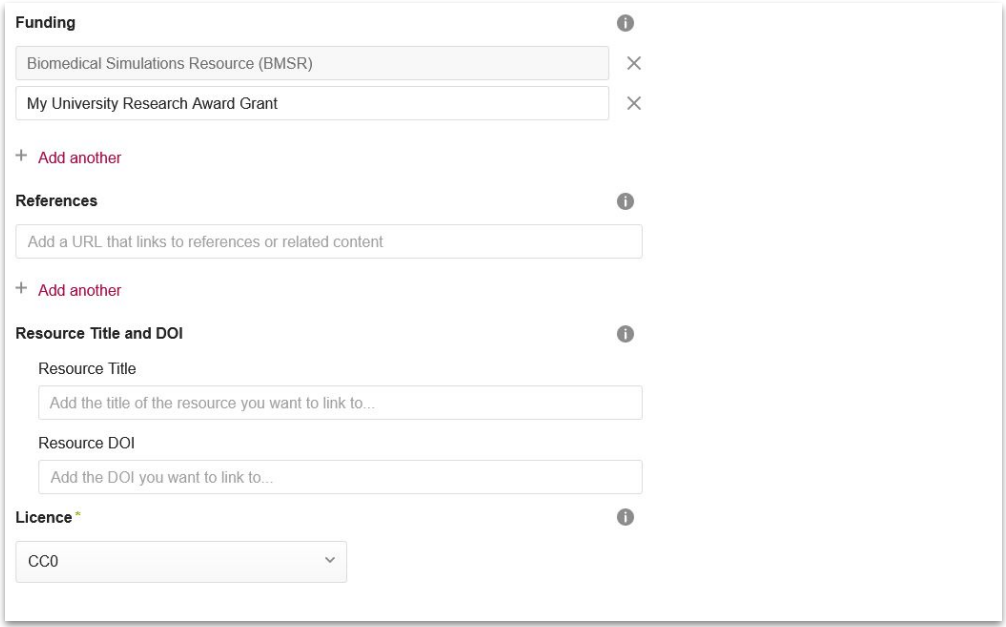


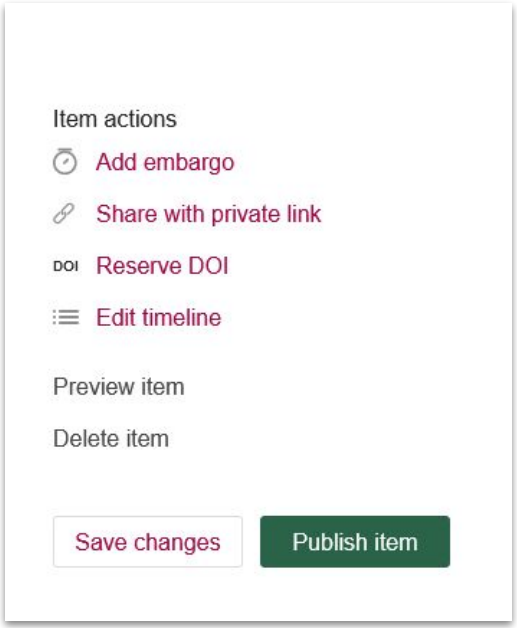
Try uploading something (and maybe publish)



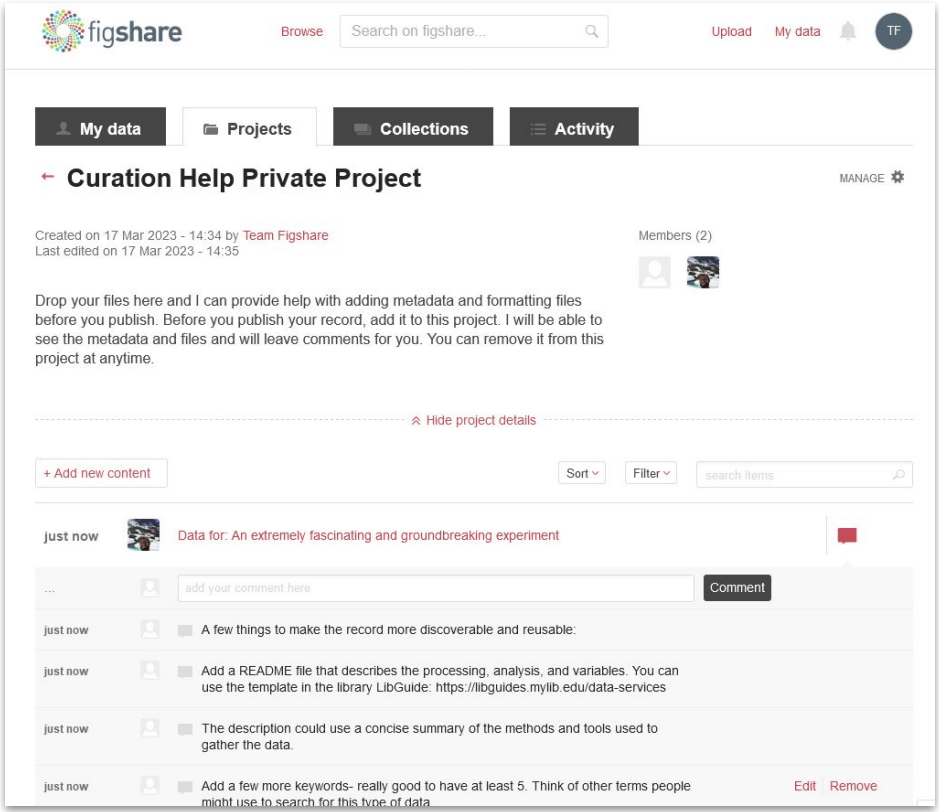


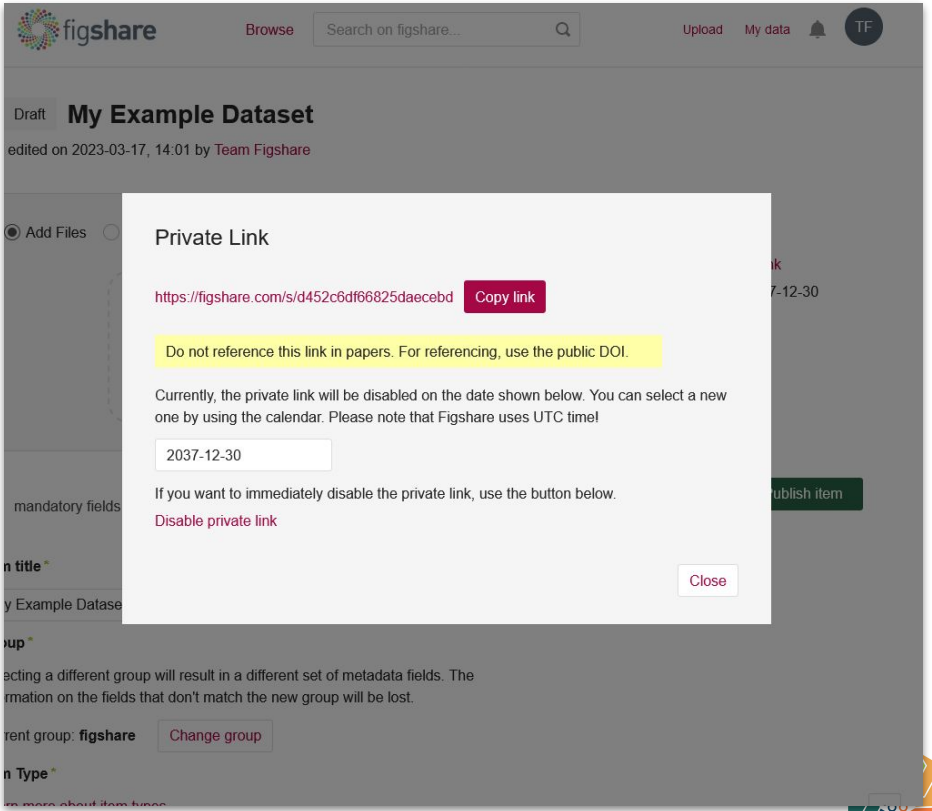
Try uploading something (and maybe publish)





Provide help for researchers





Useful metadata tips

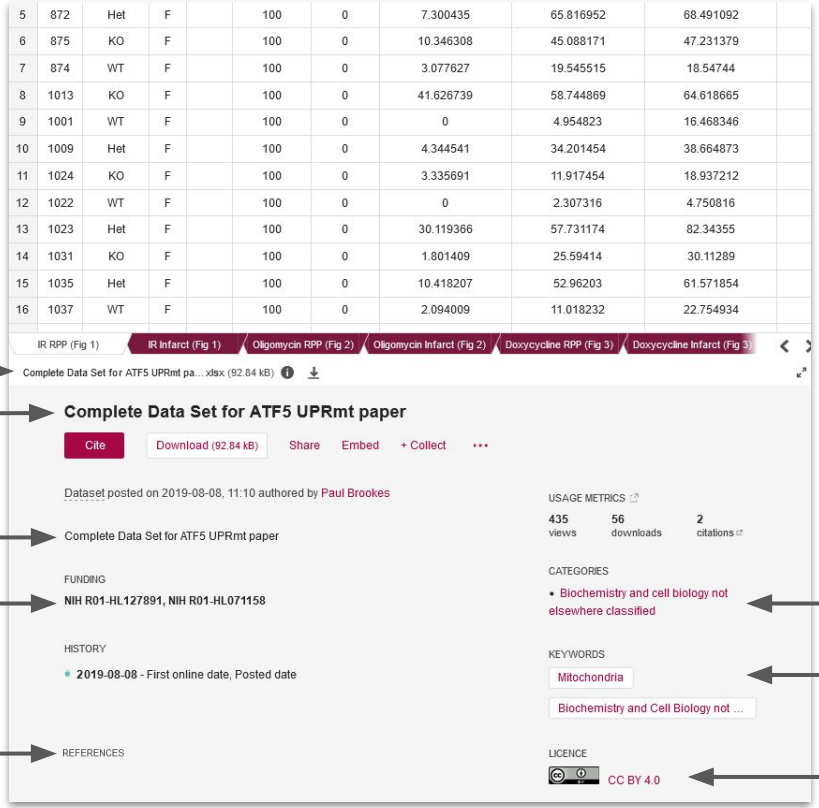
Descriptive file name

Descriptive title

Add detailed description

List and link to funding

Suggest linking to related records



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 Infarct (Fig 1) | Oligomycin RPP (Fig 2) | Oligomycin Infarct (Fig 2) | Doxycycline RPP (Fig 3) | Doxycycline Infarct (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

Complete Data Set for ATF5 UPRmt paper

FUNDING

NIH R01-HL127891, NIH R01-HL071158

HISTORY

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

REFERENCES

USAGE METRICS

435 views | 56 downloads | 2 citations

CATEGORIES

• Biochemistry and cell biology not elsewhere classified

KEYWORDS

Mitochondria

Biochemistry and Cell Biology not …

LICENCE

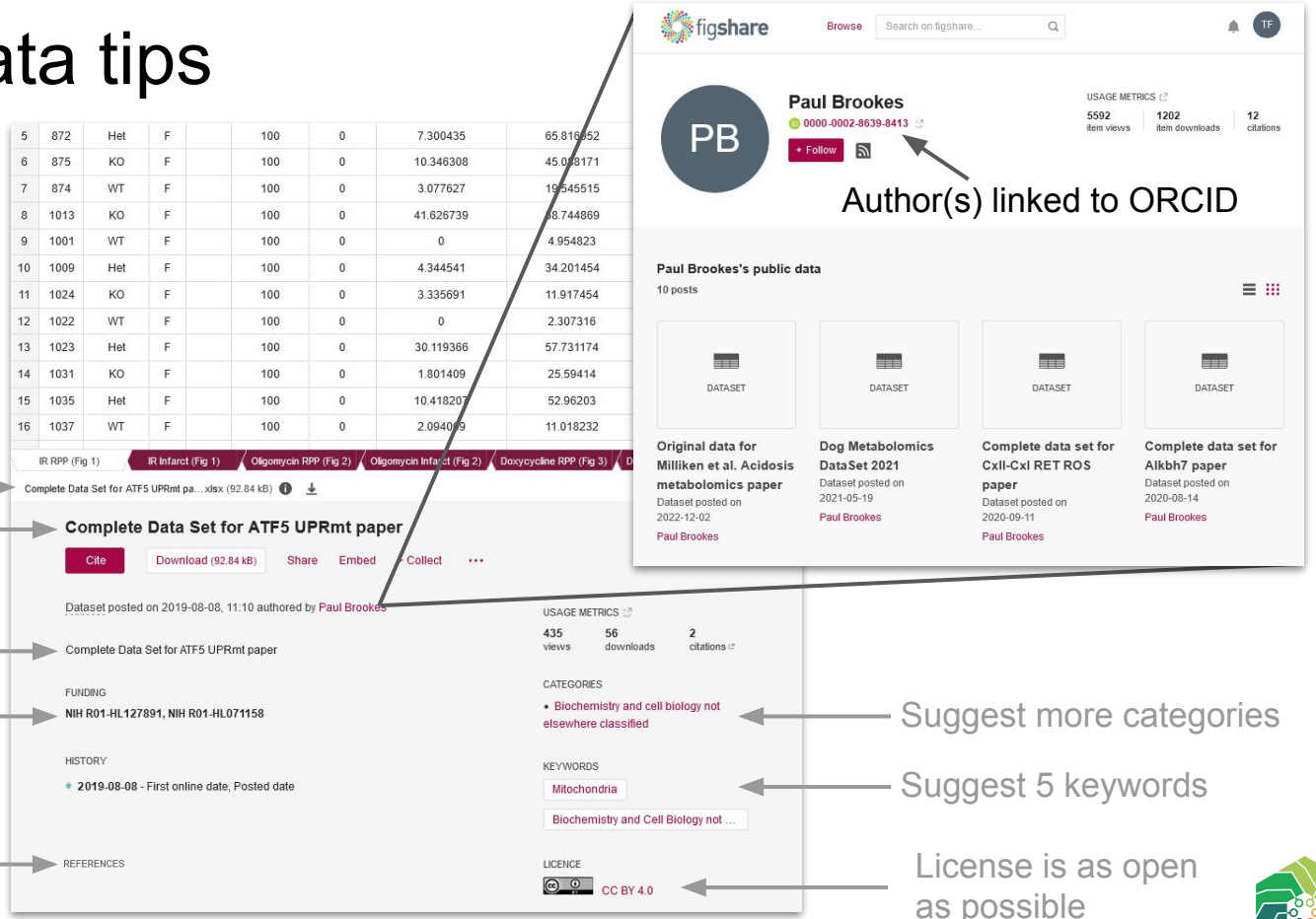
CC BY 4.0

Suggest more categories

Suggest 5 keywords

License is as open as possible

Useful metadata tips

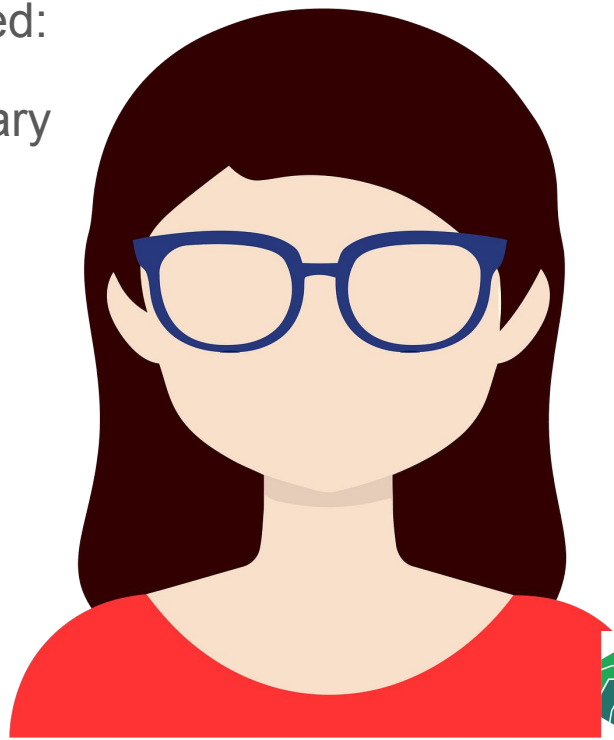


Scenario 2: Finding and Reusing Data

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

Institutional Repositories

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

Generalist repositories

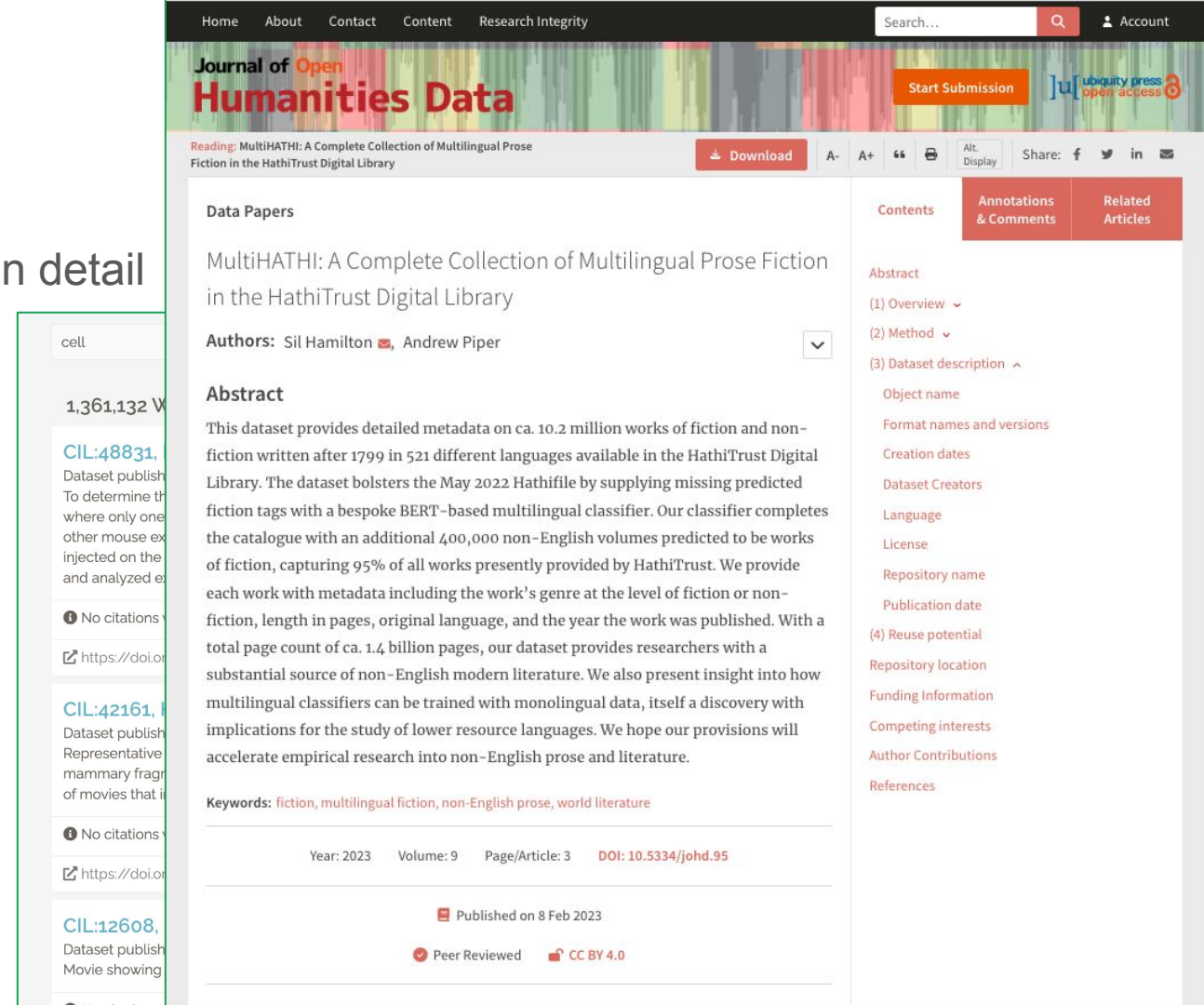
- Broad and interdisciplinary
- Commonly used metadata

Meta-search

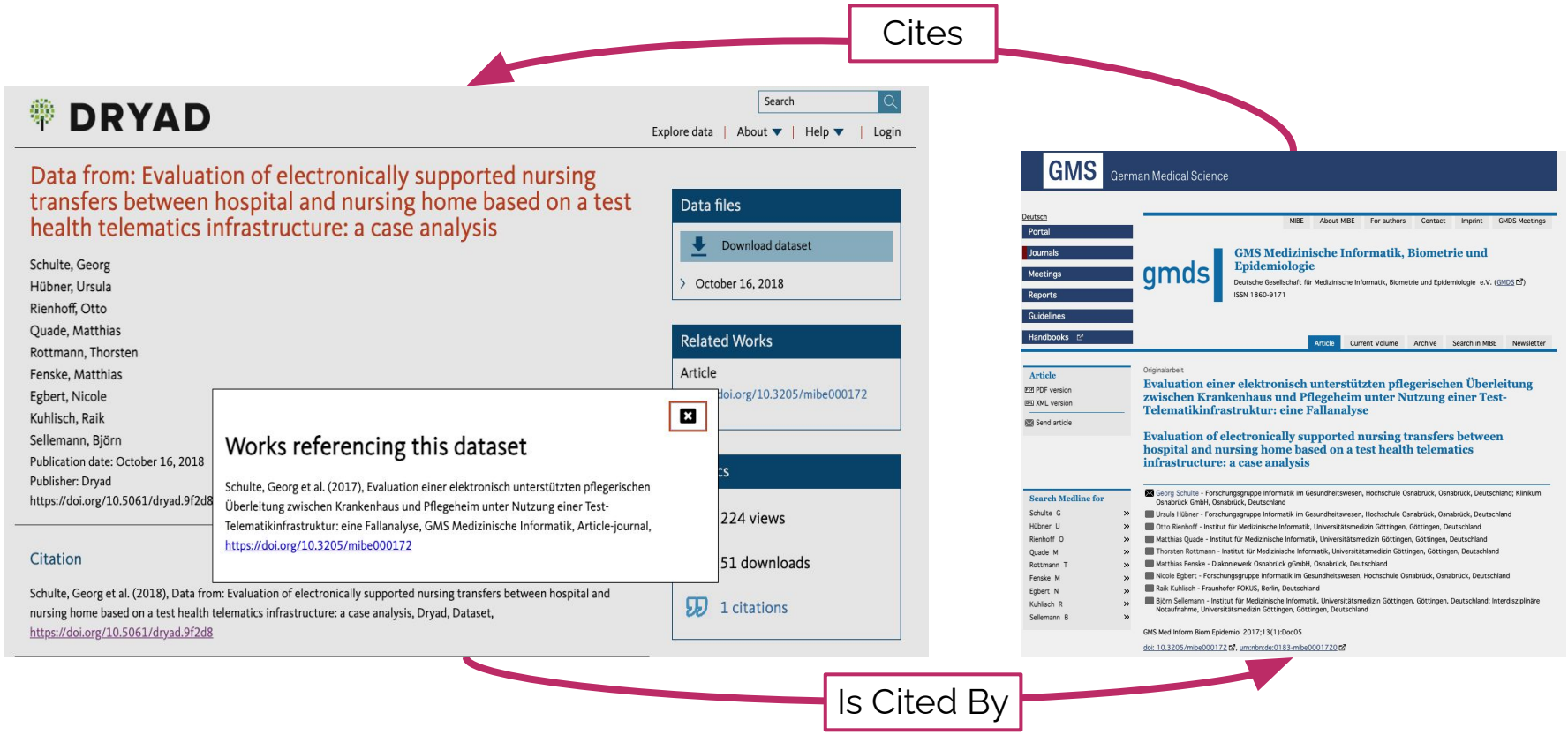
- Broad coverage
- Minimal metadata

Data Papers

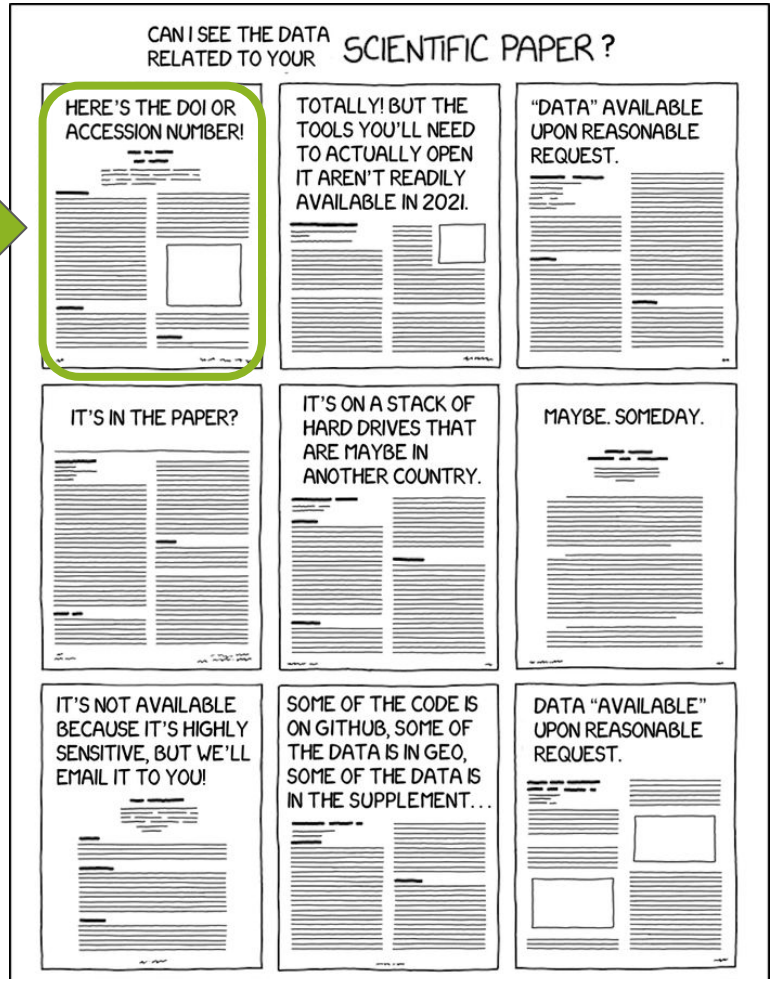
- Describes a dataset in detail
- Peer reviewed
- Includes reuse information



Citation Chaining



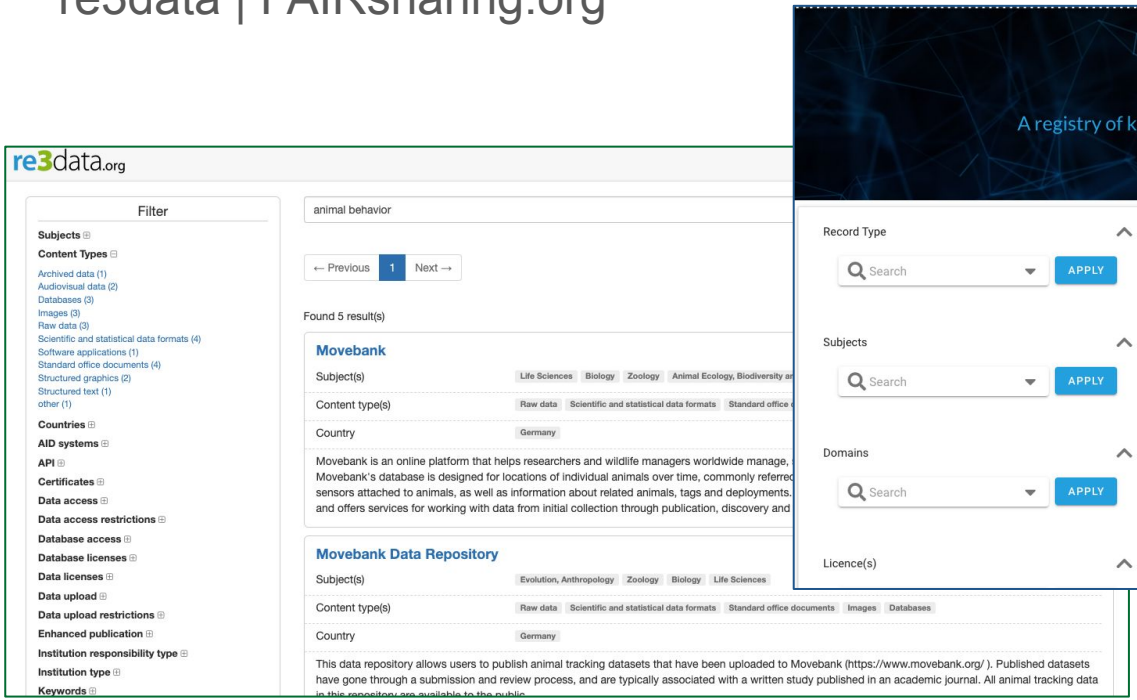
Data Citation

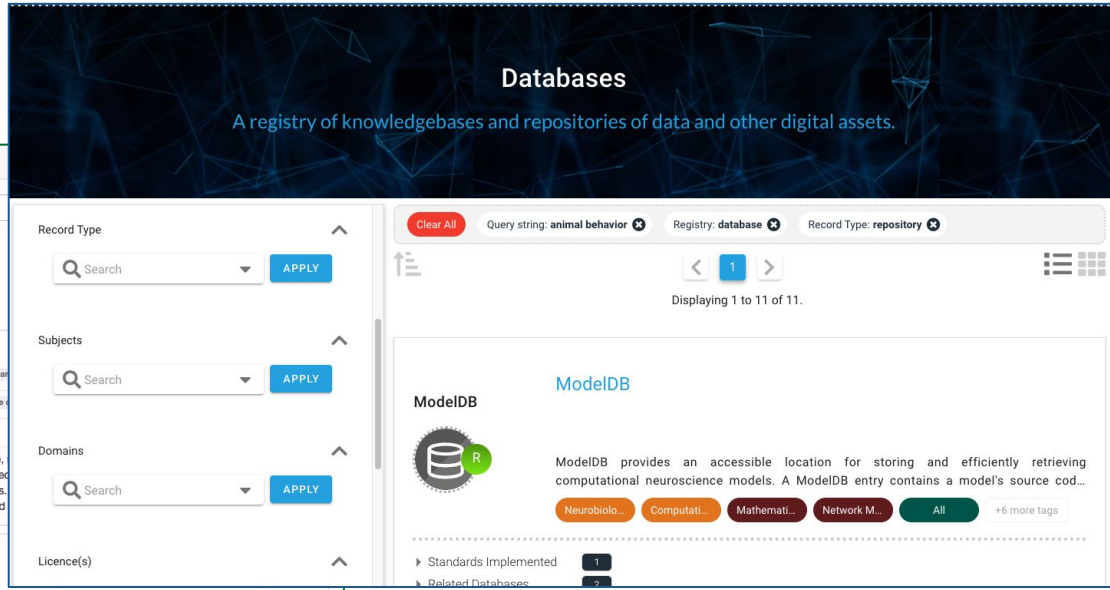


John Borghi [@JohnBorghi]. (2021, May 5). As seen today at @CSVConference: Can I see the data related to your scientific paper? Another riff on that @XKCD comic. Original here: Https://xkcd.com/2456/ #csvconf #commallama [Tweet]. Twitter. https://twitter.com/JohnBorghi/status/1390033561579835393

Generalist Repository Identification

re3data | FAIRsharing.org





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
Creators
Title
PublicationYear
Publisher
ResourceType

Best Practice: Document Search Strategy

Search Strategy2.docx

Page: 1 of 1 Automatic Zoom

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

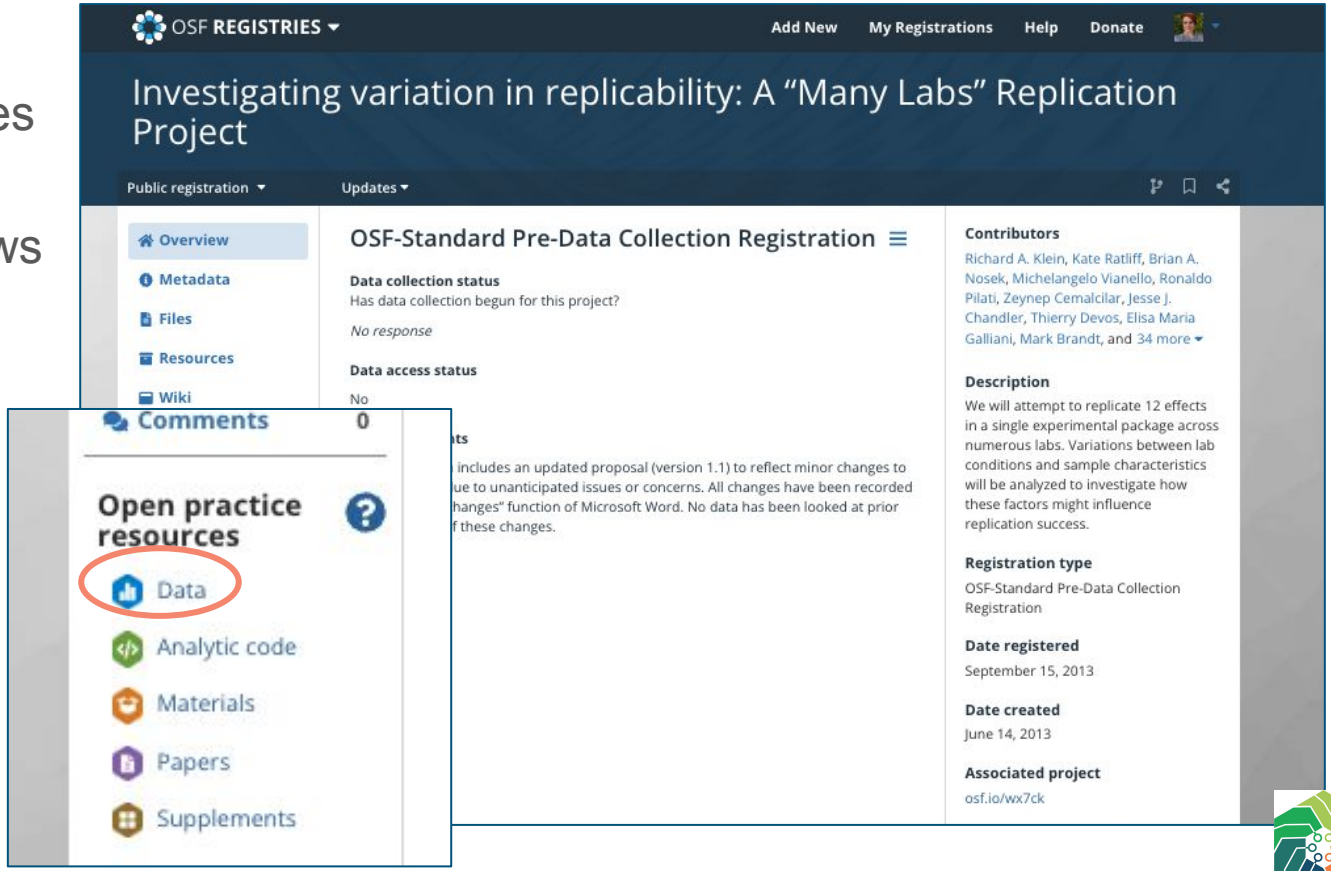
Databases:

- CINHAL

Datacite Metadata

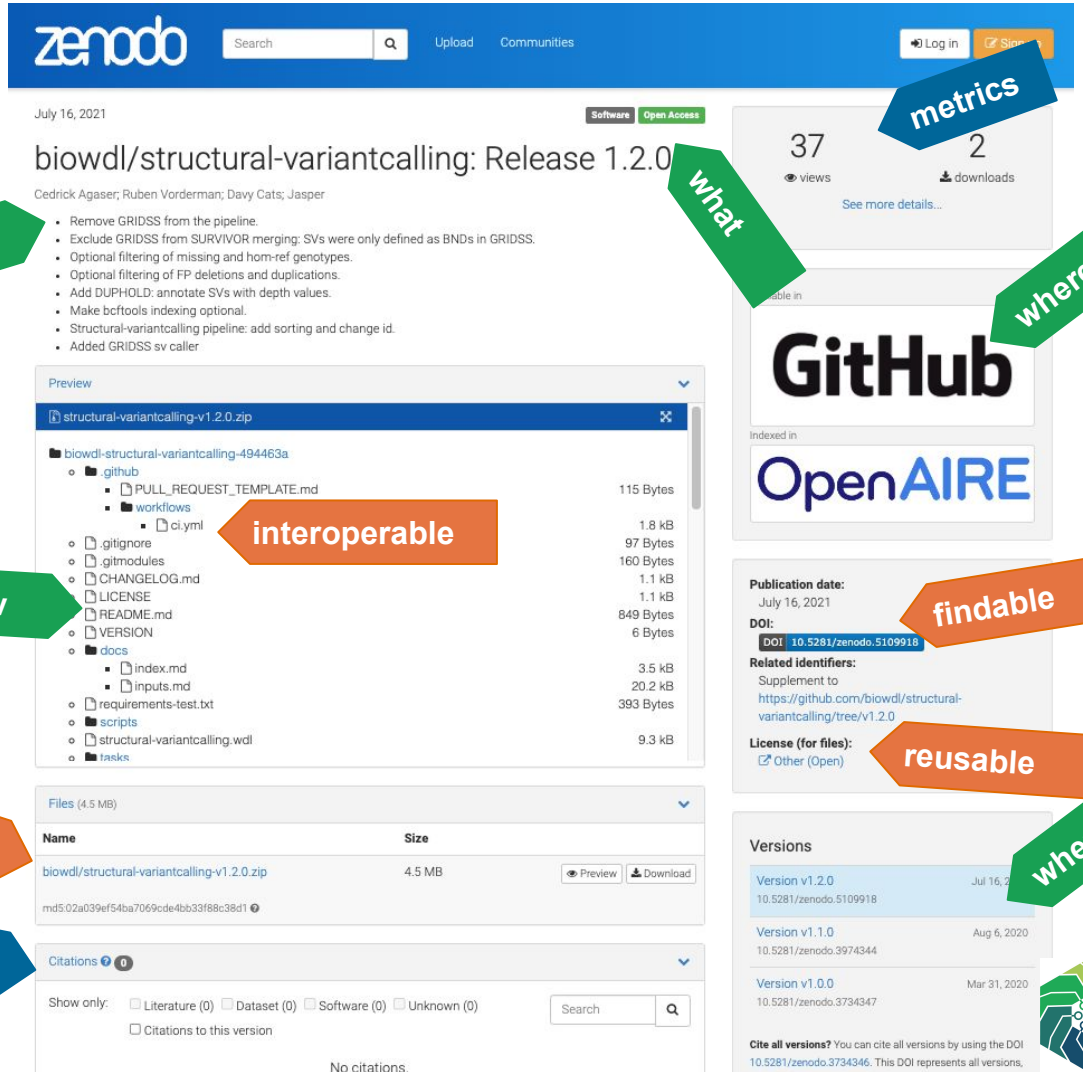
Generalist Repository Discovery

- Replication studies
- Meta-analysis
- Systematic reviews



Evaluate Datasets

- Metadata Quality
  - FAIR
    - Findable
    - Accessible
    - Interoperable
    - Reusable
- Documentation Quality
  - Who, what, when, where, why, how
- Reuse indicators
  - Metrics
  - Citations



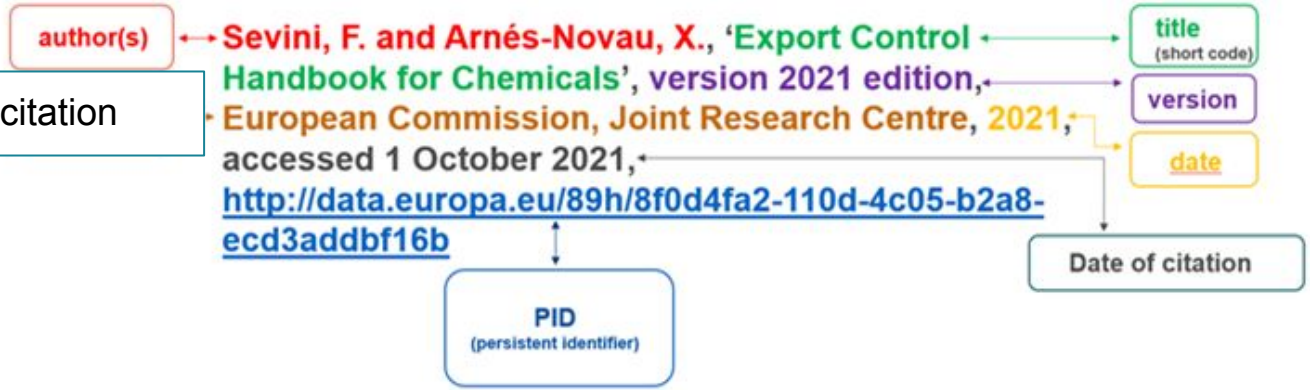
Select Data

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

Best Practice: Data citation

Completeness: data|docur

Data citation

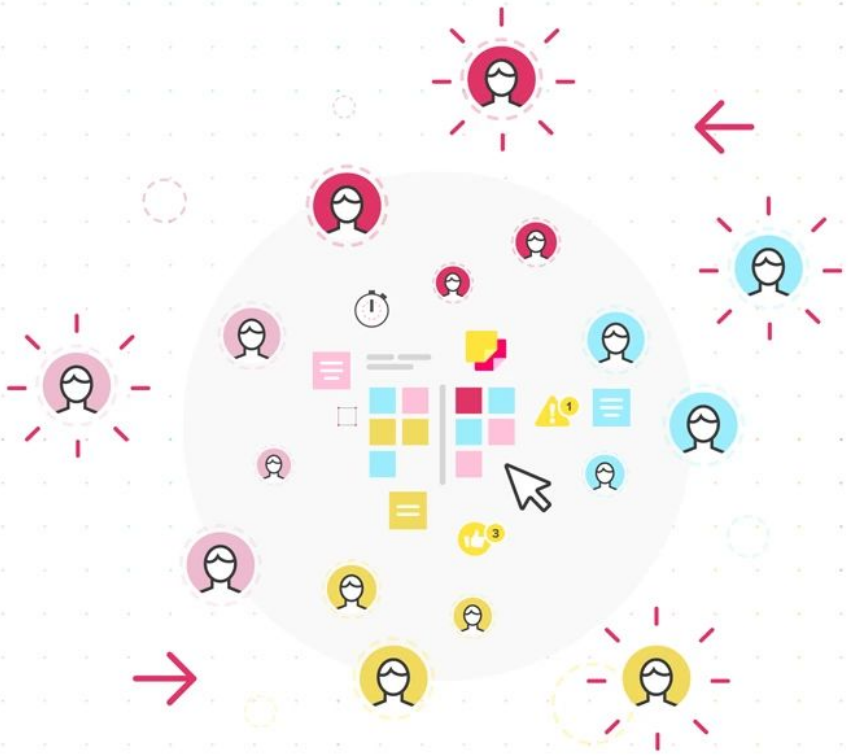


Poll Question

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

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

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

MOLECULAR ECOLOGY

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?

Open source research data repository software

95 Installations around the world

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

Budgeting for Data Management & Sharing

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

Budgeting for Data Management & Sharing

Questions to Ask & What to Include

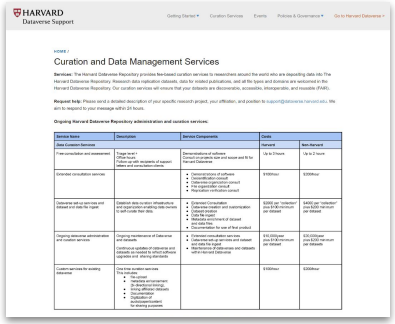
- Is there a deposit fee for any of the repositories you plan to use?
- Will you need dedicated personnel time or will you engage the services of a core or vendor?
  - Data curation?
  - Developing supporting documentation?
  - Formatting data according to accepted community standards or for transmission 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?

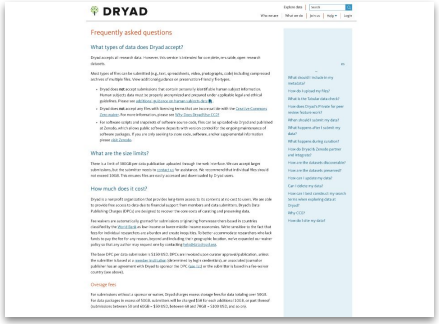
COGR Analysis - Project-Based and Institutional Cost Considerations: Budgeting & Costing

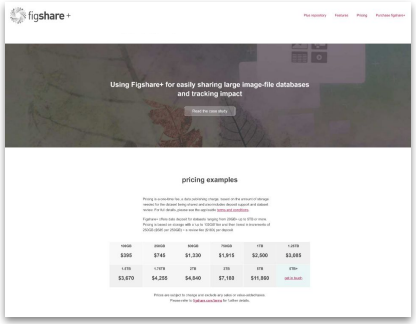
Poll Question

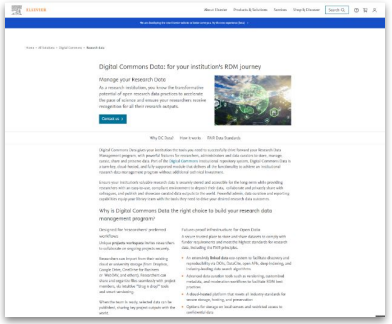
Do you plan to leverage generalist repositories as part of supporting data management and sharing plans?

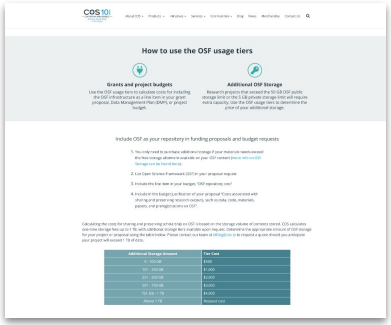
Repository Costs

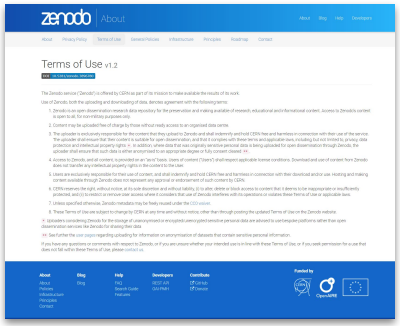












Harvard Dataverse Data Curation Services

Service Name	Description	Service Components	Costs	
Data Curation Services			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		• 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.	• 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	• 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: • 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

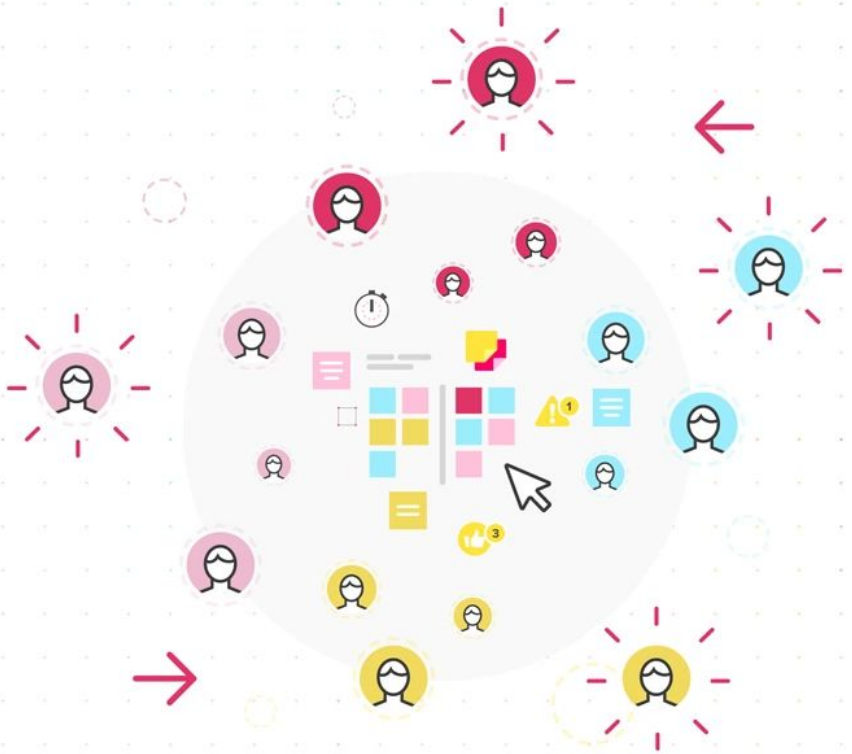
Hypothetical 5 Year Project Budget: $54,000

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!