Curating and Publishing Big Datasets Using CU Boulder High Performance Computing Infrastructure

Matthew Murray and Andrew Monaghan

View the slides

https://github.com/ResearchComputing/rmacc_2025

Introduction

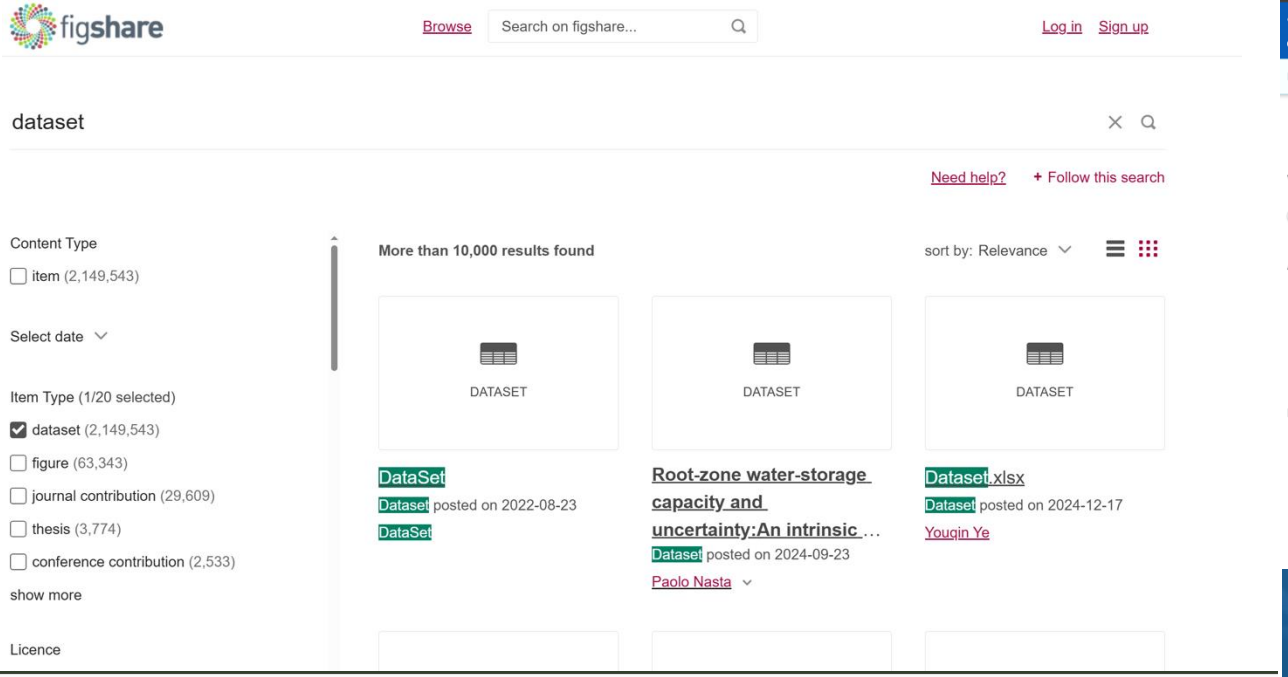
- No two people on any campus will agree on what “Data” actually is
- Publication of research datasets is now a requirement of most funding agencies and journals
- Just publishing datasets in repositories isn’t enough to make research reproducible or replicable
- Data curation is the process of ensuring that datasets are findable, accessible, and usable
- Big data creates added challenges for curating and publishing data

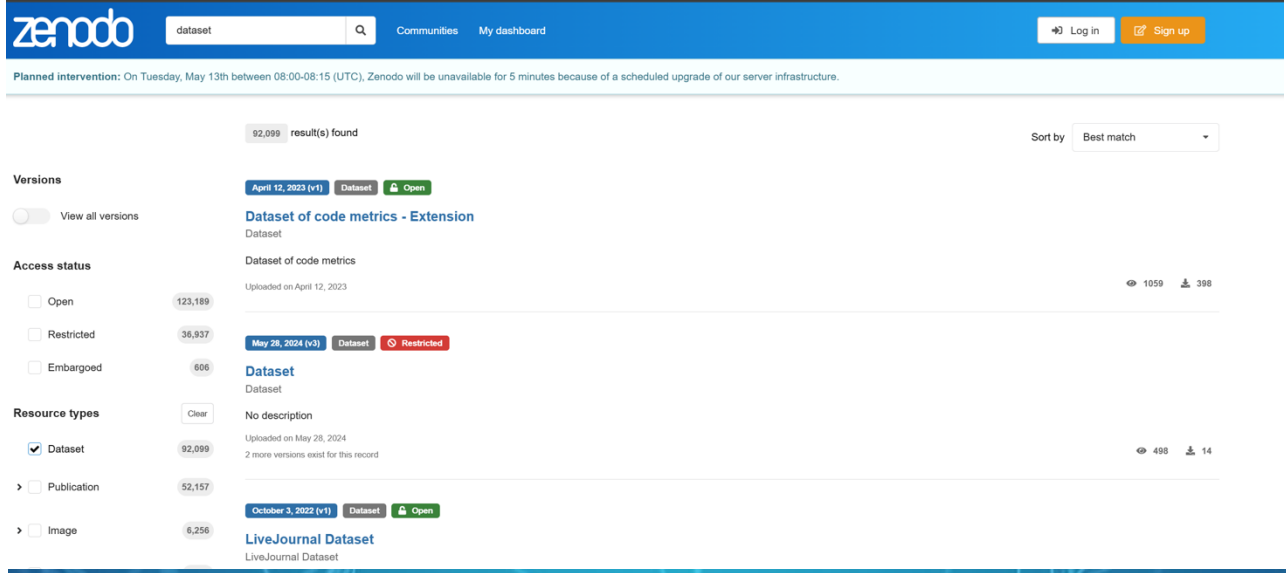
What is Research Data?

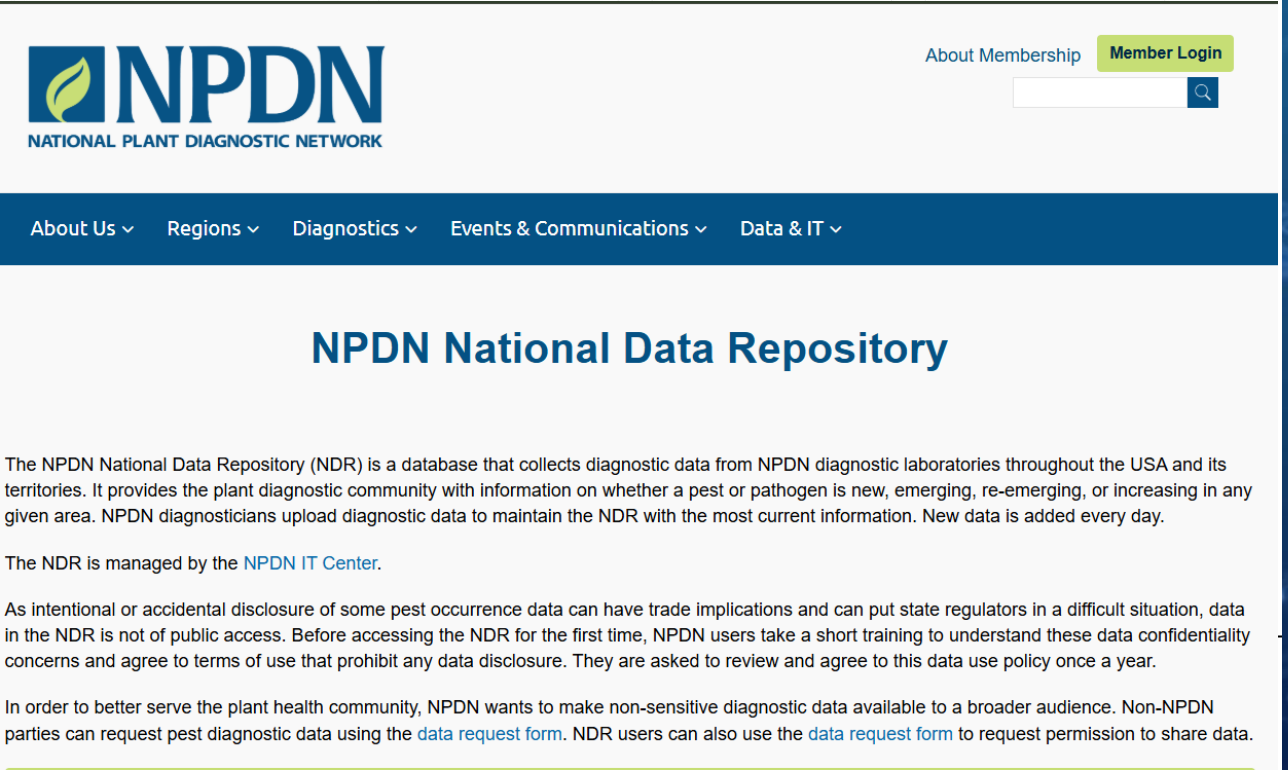
- 3D models and printable files
- Accreditation reports
- Archival university papers
- Artistry and performance materials
- Audio
- Books
- Computer code & scripts
- Conference proceedings
- Course catalogs
- Datasets
- Designs & blueprints
- Digital journals
- Dissertations
- Documentation
- GIS files
- Grant proposals
- “Grey” literature
- Historical documents
- Images
- Interviews
- Journals
- Lab notebooks
- Learning materials
- Lecture transcriptions
- Maps
- Methodologies & Workflows
- MOOCs
- Newsletters
- Oral History
- Physical artifacts and specimens
- Point clouds
- Posters
- Presentations
- Seismic recordings
- Software
- Spreadsheets / CSV files
- Surveys
- Technical reports
- Teaching tools designed by faculty
- Theses
- Transcripts
- Video
- Visualizations
- Websites
- White papers

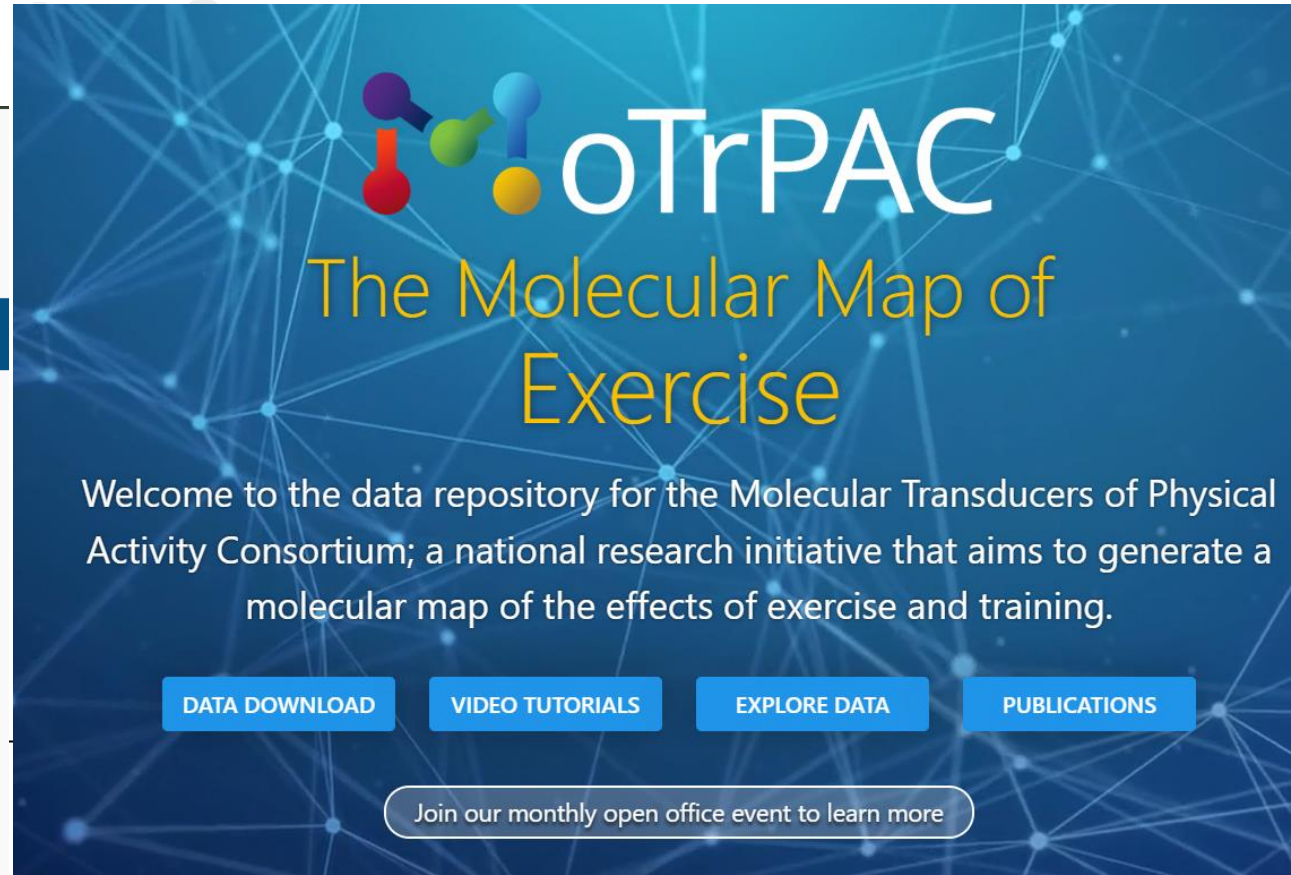
What is a Repository?

- A digital space used for publishing, sharing, and preserving works
- Repositories include articles, reports, slide decks, theses, data, and more
- There are general, domain-specific, national, and institutional
- Content may be openly accessible immediately, embargoed to a future date, or only accessible to certain users
- Gained popularity as many journals and funders now require data to be shared when articles are published


figshare
Browse
Search on figshare...
Log in
Sign up
dataset
Need help?
+ Follow this search
Content Type
item (2,149,543)
Select date
Item Type (1/20 selected)
dataset (2,149,543)
figure (63,343)
journal contribution (29,609)
thesis (3,774)
conference contribution (2,533)
show more
Licence
More than 10,000 results found
sort by: Relevance
DATASET
DataSet
Dataset posted on 2022-08-23
DataSet
DATASET
Root-zone water-storage capacity and uncertainty:An intrinsic ...
Dataset posted on 2024-09-23
Paolo Nasta
DATASET
Dataset.xlsx
Dataset posted on 2024-12-17
Yougin Ye


zenodo
dataset
Communities
My dashboard
Log in
Sign up
Planned intervention: On Tuesday, May 13th between 08:00-08:15 (UTC), Zenodo will be unavailable for 5 minutes because of a scheduled upgrade of our server infrastructure.
92,099 result(s) found
Sort by Best match
Versions
View all versions
Access status
Open 123,189
Restricted 36,937
Embargoed 606
Resource types
Clear
Dataset 92,099
Publication 52,157
Image 6,256
April 12, 2023 (v1) Dataset Open
Dataset of code metrics - Extension
Dataset
Dataset of code metrics
Uploaded on April 12, 2023
1059 398
May 28, 2024 (v3) Dataset Restricted
Dataset
Dataset
No description
Uploaded on May 28, 2024
2 more versions exist for this record
498 14
October 3, 2022 (v1) Dataset Open
LiveJournal Dataset
LiveJournal Dataset


NPDN
NATIONAL PLANT DIAGNOSTIC NETWORK
About Membership
Member Login
About Us
Regions
Diagnostics
Events & Communications
Data & IT
NPDN National Data Repository
The NPDN National Data Repository (NDR) is a database that collects diagnostic data from NPDN diagnostic laboratories throughout the USA and its territories. It provides the plant diagnostic community with information on whether a pest or pathogen is new, emerging, re-emerging, or increasing in any given area. NPDN diagnosticians upload diagnostic data to maintain the NDR with the most current information. New data is added every day.
The NDR is managed by the NPDN IT Center.
As intentional or accidental disclosure of some pest occurrence data can have trade implications and can put state regulators in a difficult situation, data in the NDR is not of public access. Before accessing the NDR for the first time, NPDN users take a short training to understand these data confidentiality concerns and agree to terms of use that prohibit any data disclosure. They are asked to review and agree to this data use policy once a year.
In order to better serve the plant health community, NPDN wants to make non-sensitive diagnostic data available to a broader audience. Non-NPDN parties can request pest diagnostic data using the data request form. NDR users can also use the data request form to request permission to share data.


oTrPAC
The Molecular Map of Exercise
Welcome to the data repository for the Molecular Transducers of Physical Activity Consortium; a national research initiative that aims to generate a molecular map of the effects of exercise and training.
DATA DOWNLOAD
VIDEO TUTORIALS
EXPLORE DATA
PUBLICATIONS
Join our monthly open office event to learn more

Institutional Repositories (IRs)

- For work created by those affiliated with the institution
- Usually run by libraries
- Can be general (include articles, etc.) or data-specific
- Data is usually associated with a published academic article
- Repositories may allow for self-deposit and instant publication or require data to be curated

Share Your Work

Terms of Use

Featured Works | Recently Uploaded

Domestic violence and sports news: How gender affects people's understanding

Creator: Painter, Chad, Ferrucci, Patrick, Tandoc, Edson, Willis, Erin
Subject: Domestic violence, Mass media and sports, Sports journalism, Sex role, Gender role, Sports news, Family violence
Resource Type: Article

Public Narratives of Domestic Violence: Giving a Public Face to Personal Transformations

Creator: Hirsch, Christine Courtade
Subject: no subject specified
Resource Type: Dissertation

Domestic Violence in Boulder County: Gendered Trends in Case Outcomes

Creator: McCullar, Sarah
Subject: Women and Gender Studies
Resource Type: Undergraduate Honors Thesis

Explore Collections | Featured Researcher

Series in Biology

University Libraries Fellows

Western States Government Information Virtual Conference

CIRES Center for Education, Engagement and Evaluation

Entrepreneurial Startup Skills

View all collections

© Regents of the University of Colorado

Search CU Scholar | data

Limit your search

- Creator
- Contributor
- Academic Affiliation
- Subject
- Language
- Location
- Publisher
- Resource type
  - Data Set ✖ 2,302
- Collections

Filtering by: data ✖ | Resource type › Data Set ✖ | Start Over

« Previous | **1 - 10** of **2,302** | Next »

Sort by relevance | 10 per page

Microlending 2017 Data Set

Creator: Sonboli, Nasim; Aird, Amanda; Burke, Robin D.

Academic Affiliation: Information Science

Subject: Recommender Systems and Microlending

Abstract: This dataset contains anonymized lending transactions from the crowdsources microlending site Kiva Microloans. The data has been transformed to make it suitable for recommender systems experimentation and research. See the attached README file and datasheet document for additional information. Copyright 2022 Robin Burke

Resource Type: Data Set

Data used for 2001-2020 spotted knapweed research

Creator: Seastedt, Timothy R

Academic Affiliation: Institute of Arctic & Alpine Research

Abstract: Data documentation required for submission of manuscript, "Biological control of spotted knapweed (Centaurea stoebe) in Colorado: A 20-year perspective"

Resource Type: Data Set

Data sets associated with "Sediment Production in French Alpine Rivers"

Creator: Pitlick, John

Academic Affiliation: Geography

Subject: Annual Sediment Yield and Sediment Transport

Abstract: This repository stores two data sets associated with the manuscript titled "Sediment Production in French Alpine Rivers". The first file, Ecrins_Site_Data.xlsx,is an Excel workbook listing data for the 16

Data Curation

- The process of ensuring that data is FAIR:
  - **F**indable
  - **A**ccessible
  - **I**nteroperable
  - **R**eusable
- Why do we need to do this?

Published March 25, 2025 | Version v1

Dataset | Open

data

Dataset, Subseasonal

Files

U_component_of_850hPa_wind.zip

U_component_of_850hPa_wind.zip

U_component_of_850hPa_wind

u_850_s2s_anso_week1.nc	1.7 GB
u_850_s2s_anso_week2.nc	1.7 GB
u_850_s2s_anso_week3.nc	1.7 GB
u_850_s2s_anso_week4.nc	1.7 GB
u_850_s2s_anso_week5.nc	1.7 GB
u_850_s2s_anso_week6.nc	1.7 GB

Files (45.9 GB)

Download all

Name	Size

45 VIEWS

37 DOWNLOADS

Show more details

Versions

Version v1 — Mar 25, 2025

10.5281/zenodo.15080493

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.15080492. This DOI represents all versions, and will always resolve to the latest one. Read more.

External resources

Indexed in

OpenAIRE

Citation

Dataset, S. (2025). data [Data set]. Zenodo.
https://doi.org/10.5281/zenodo.15080493

Style APA

Dataset, S. (2025). data [Data set]. Zenodo.

Data Curation Network's CURATE(D) Steps

- **C**heck files/code and read documentation
- **U**nderstand the data (or try to)
- **R**equest missing information or changes
- **A**ugment metadata for findability
- **T**ransform file formats for reuse
- **E**valuate for FAIRness
- **D**ocument all curation activities throughout the process

When we use HPC resources for data curation

- The data requires specific software or code that can only be run using HPC resources
- The data *is* specific software/code that can only be run using HPC resources

Example

- Curator cannot open/run files on their computer
- Files are uploaded to a HPC server and tested
- Oh no! They don’t work:
  - Scripts have hardcoded paths (researcher updates scripts)
  - Specific libraries are required (researcher adds documentation)
  - A different version of the software is required (researcher adds documentation)

Publishing Data

- The data has been curated and is ready to be published
  - The data is ready to be reused
  - The documentation is complete
  - There's accurate metadata (authors, etc.)
  - A DOI has been issued
- Problems
  - The dataset is too big for the institutional repository
  - The dataset requires HPC resources to be used

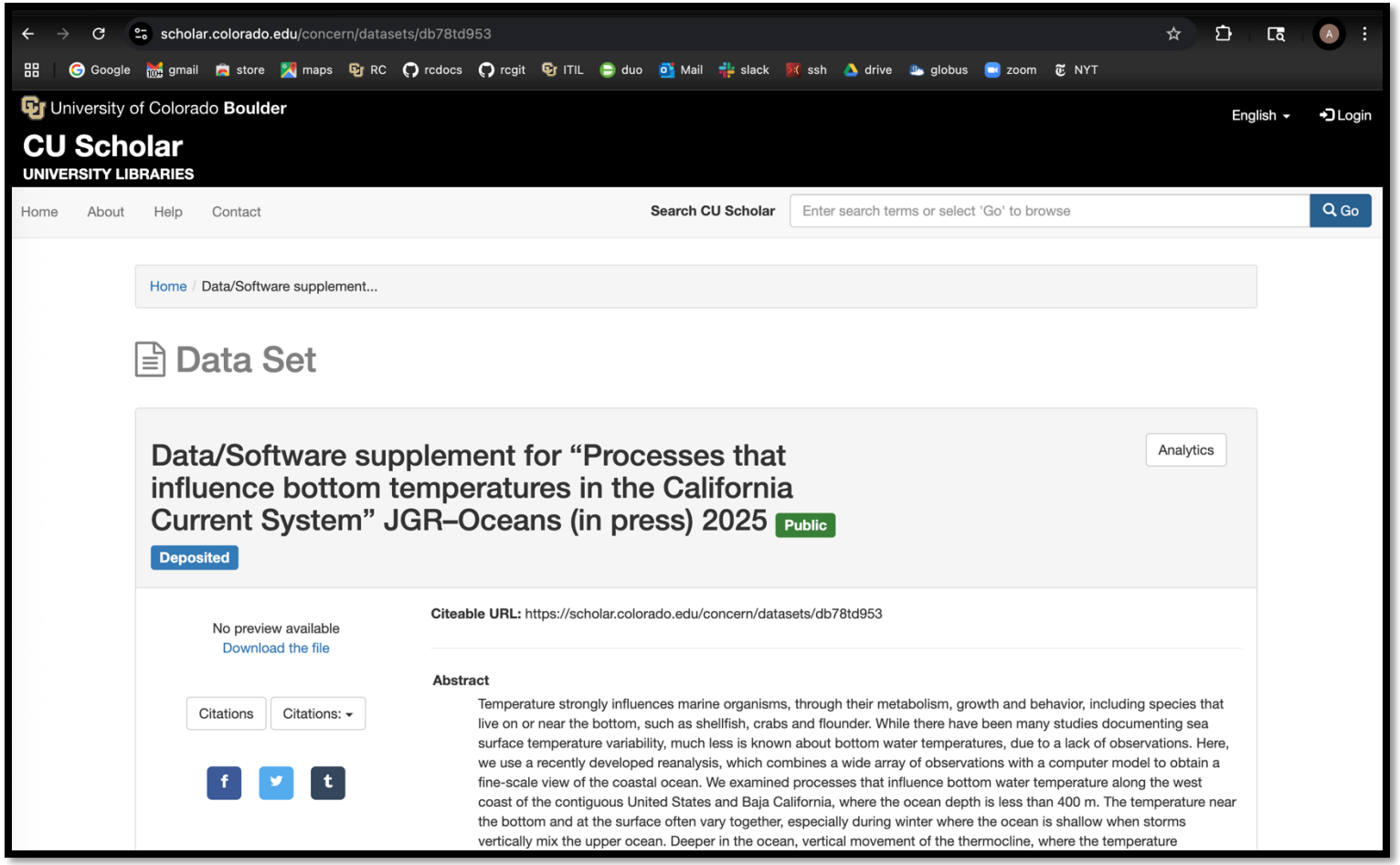
Framework for big data publishing at CU Boulder

1. Landing page
2. Storage infrastructure
3. Data transfer

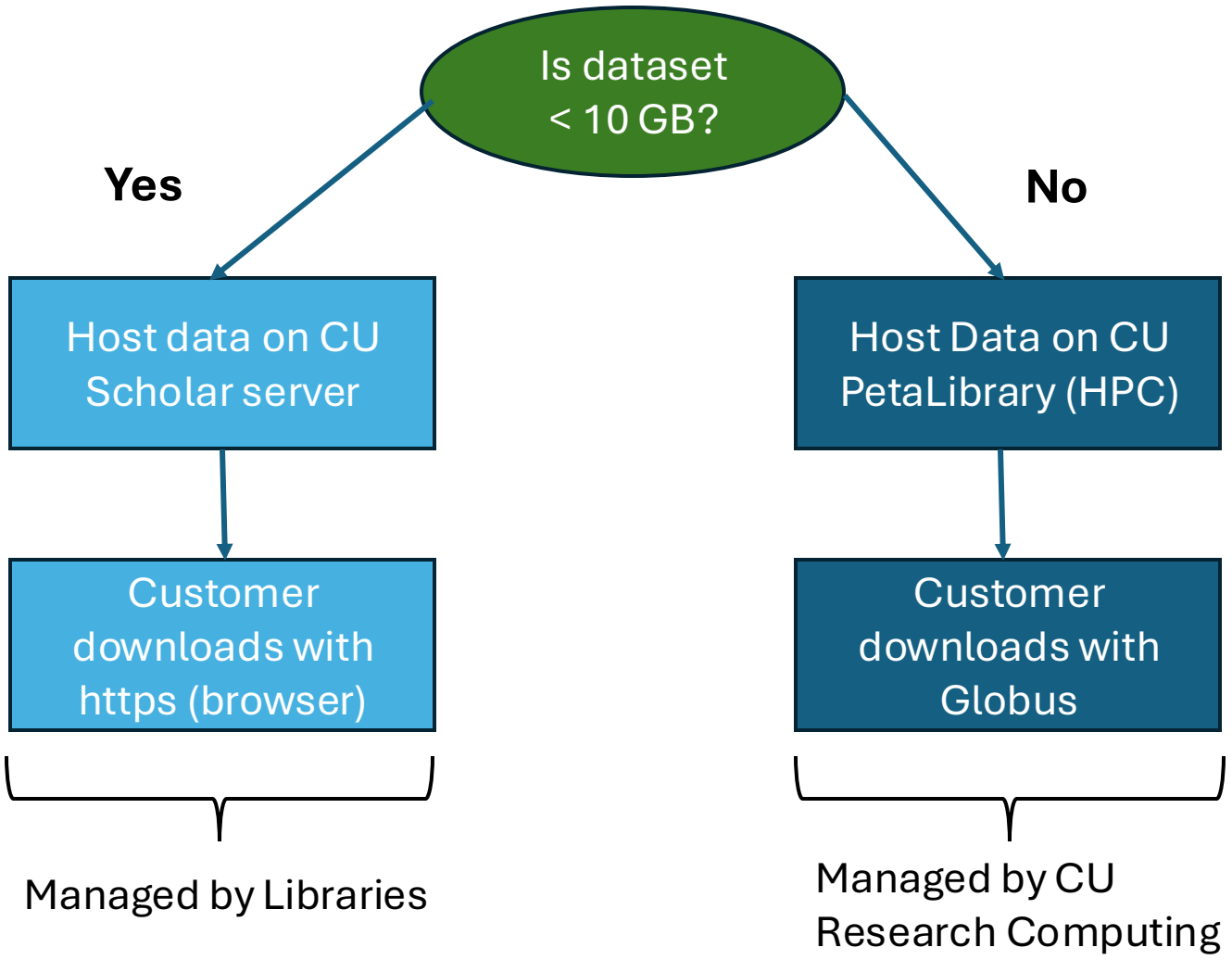
Landing Page

- CU Libraries host on CU Institutional repository:

  https://scholar.colorado.edu

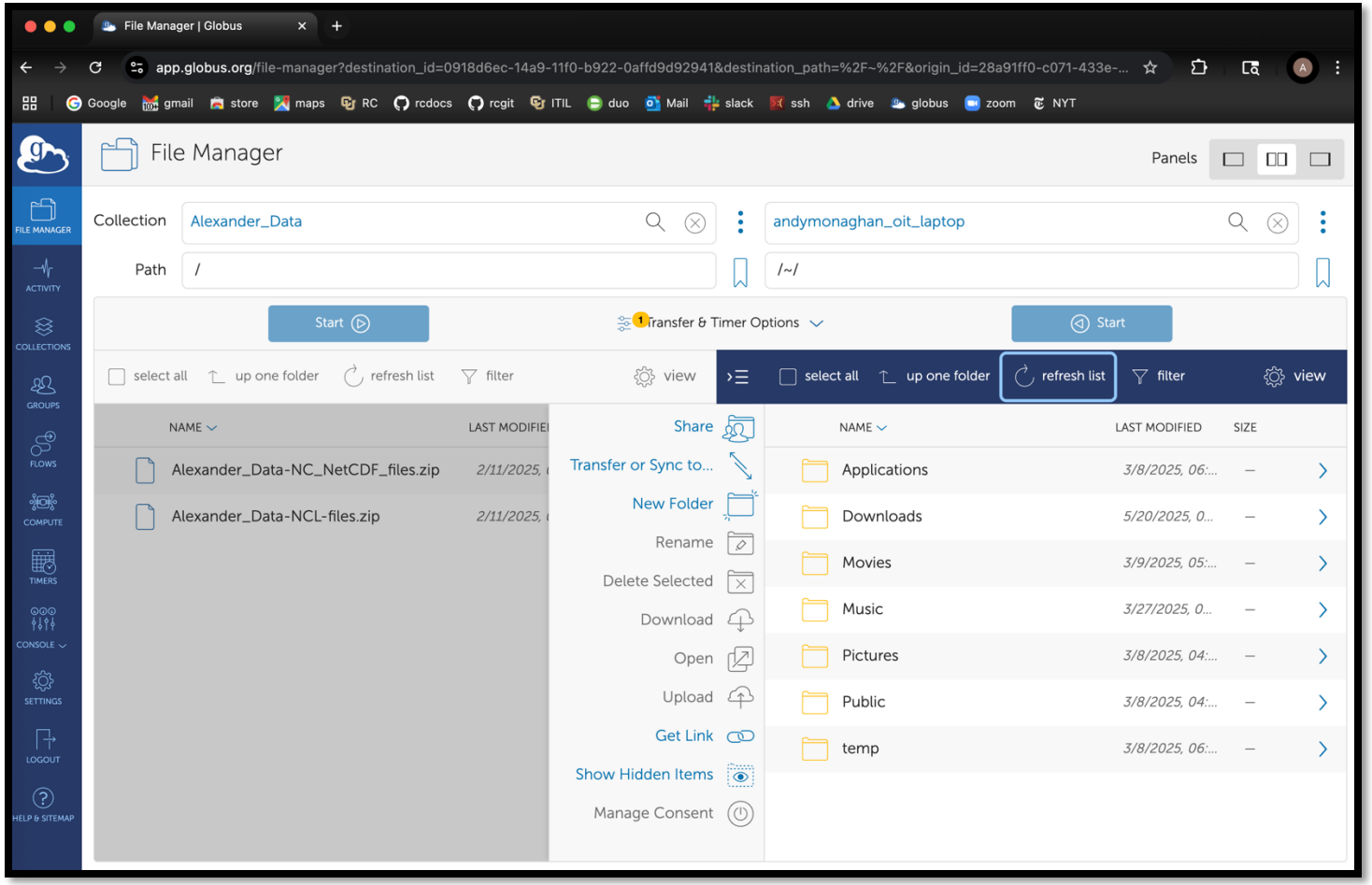


Storage Infrastructure



Data Transfer for "big data" publications

- Uses Globus
  - GridFTP transfer protocol
  - Fast, stable, fault tolerant
  - Anyone can use Globus
  - Guidance provided on CU Scholar landing page



Costs

Is there a cost for publishing large datasets at CU Boulder?

- If size < 500 GB: No
- If size > 500 GB: Yes ($450/TB)

If the data is related to a student's thesis or dissertation we'll talk with them to see what we can do (this hasn't happened yet)

How to get started:

- Email cuscholaradmin@colorado.edu
- We can advise on data curation, assist with data upload, etc.
- Reach out to the libraries/IR administrators at your institution

Future

- How can we expand use of HPC resources to provide increased access to datasets?
- Emulation
- Virtual access
- Containers