Finding Data: Opportunities for Data Services in Medical Libraries

National Center for Data Services
October 15, 2025

Michelle Yee, MPH
Research Data and Metadata Management Librarian
michelle.yee@nyulangone.org

NYU Langone Health

This session’s learning objectives

Today, learners will

- Become familiar with strategies for locating and evaluating data sources.
- Practice data reference skills in a guided activity.
- Explore opportunities for growing data services through collaborations with other academic departments.

National Center for Data Services Short Course Series

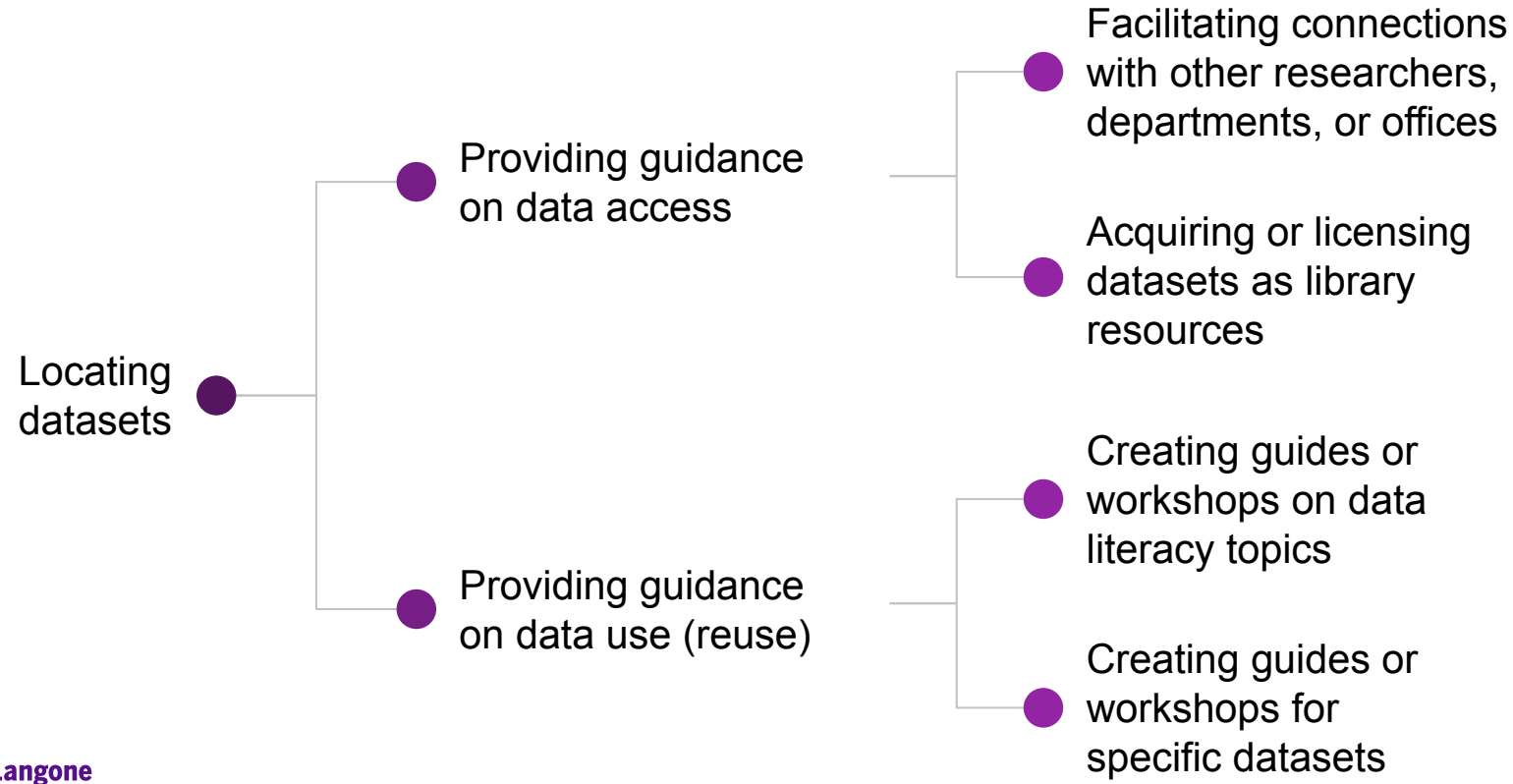
Over the course of the series, learners will

- Identify and explain different pathways to providing data services.
- Explore how low cost, high impact services like finding data can provide a foundation for building data services.

Data finding

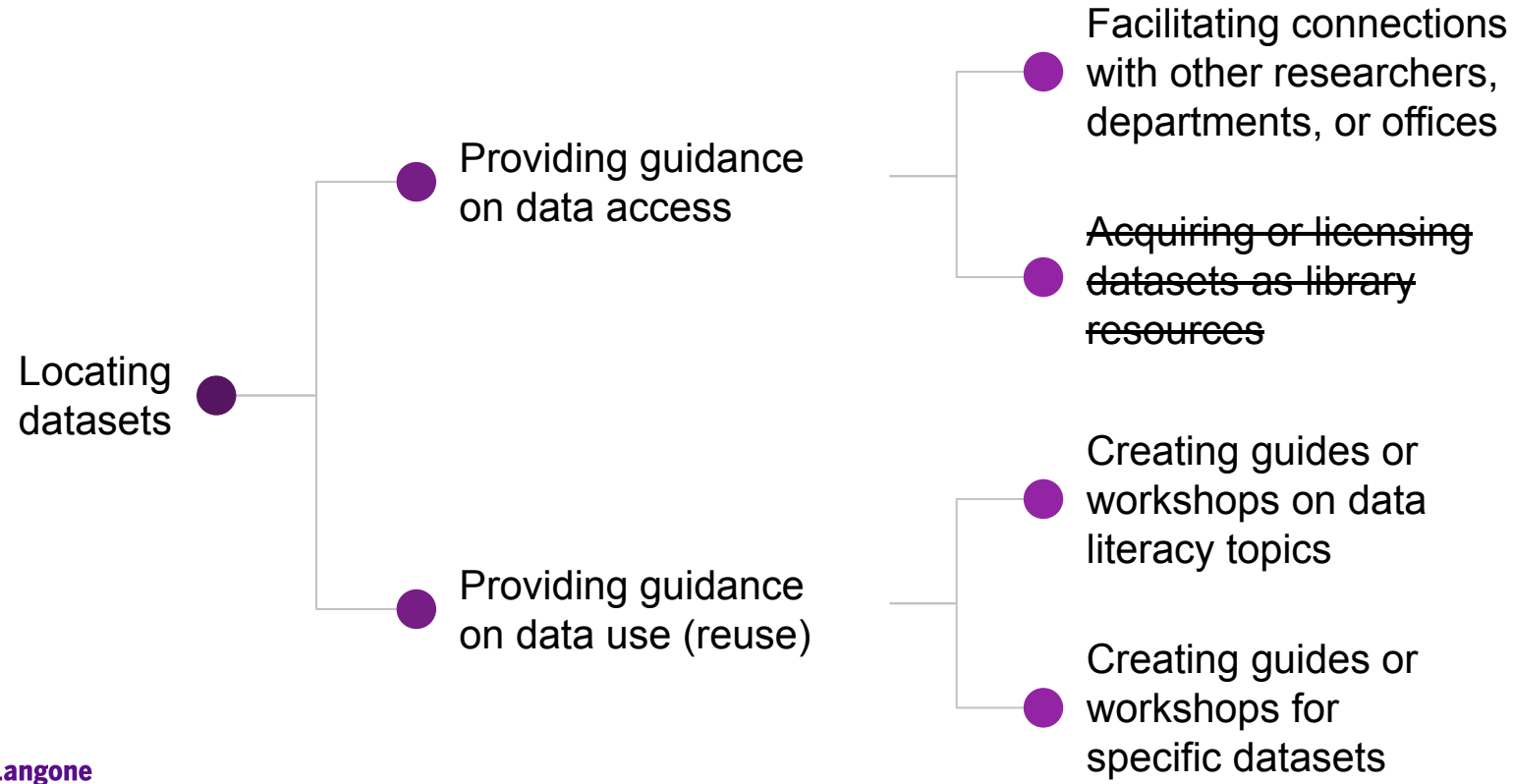
- A reference service: locating or identifying data sources to support a user need
- Not to be confused with processes that enable other people to find your/user/institution's data
  - October 23, 2025: The NIH Data Management and Sharing Policy
  - Archived webinar: "Understanding Data Discovery and Sharing Infrastructure and Leveraging It for Your Benefit"

Data finding may encompass…



NYU Langone Health

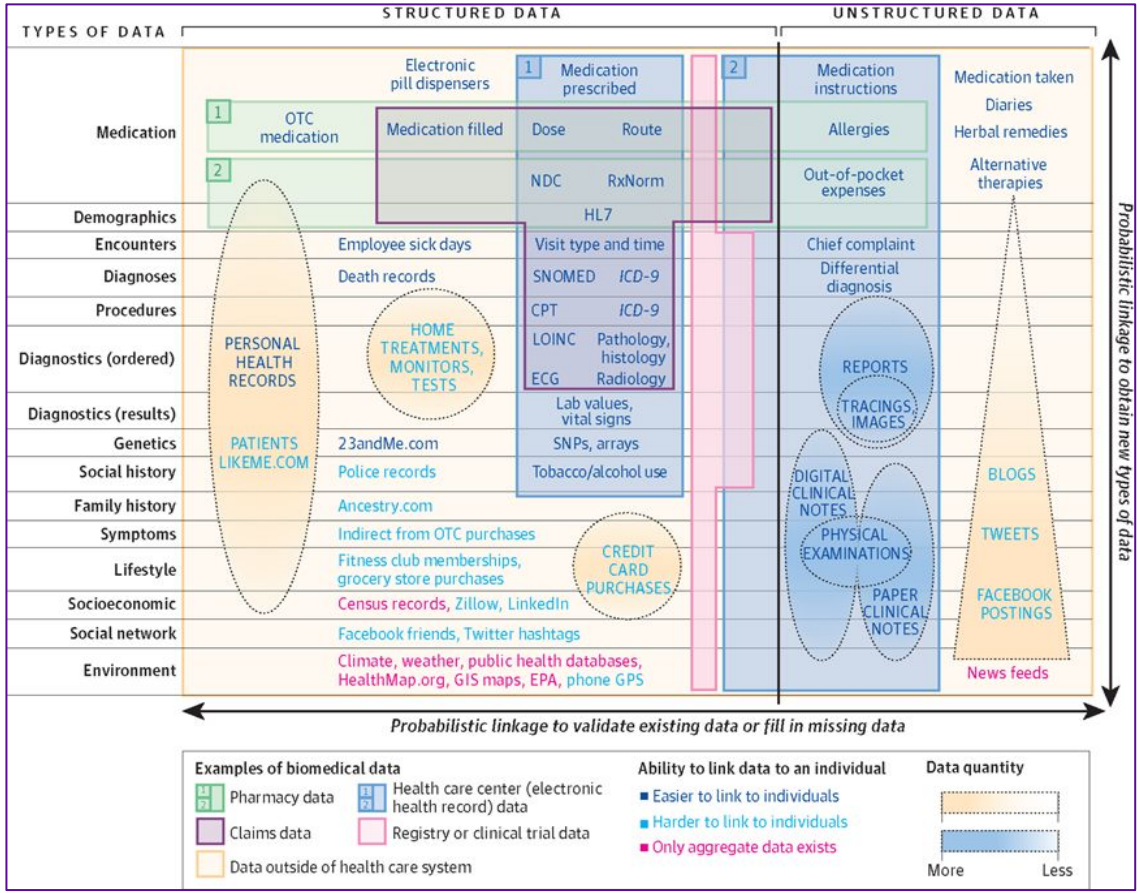
Low cost services



Strategies for finding data

Sources of data

- Surveys/polls
- Medical records
- Administrative data (e.g., claims, vital records)
- Public health surveillance
- Registries and clinical trials
- Peer-reviewed studies
- “Grey literature” (e.g., white papers, reports)



Weber GM, Mandl K, Kohane IS. Finding the Missing Link for Big Biomedical Data. *JAMA* 2014;311(24):2479-2480. doi: 10.1001/jama.2014.4228

Relating research questions and data

Are there sufficient senior care programs to keep pace with the projected growth of the aging population in the United States?

Possible data points (variables): population above 65 years old, number of senior living facilities, number of geriatric specialists

Possible data sources:

- Administrative data → population census, census or registries of healthcare providers, directory of senior living facilities, directory of home care services, number of jobs in care-related services
- Surveys or polls → senior preferences for aging care, household or family plans for senior care

Evaluating pre-existing data for new analyses

Accessibility

- Ease of finding, obtaining, and interpreting data

Documentation

- Information needed to understand the data

Limitations and biases

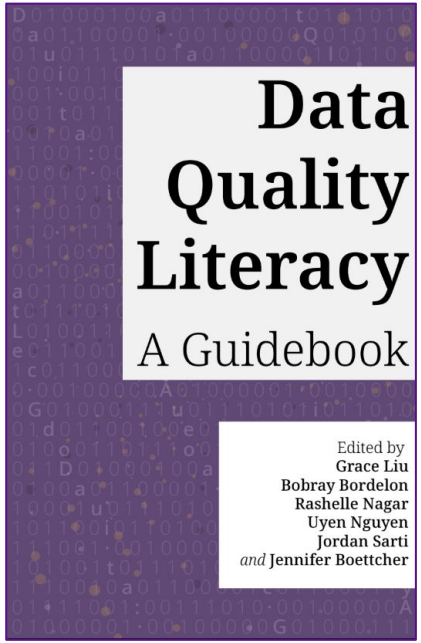
- Misinterpretation or perpetuation of errors and biases

Suppressed values

- Limited or no demographic information for data linkage or analysis of confounding variables

Resources for further reading

- Cheng HG, Phillips MR. Secondary analysis of existing data: opportunities and implementation. *Shanghai Arch Psychiatry.* 2014 Dec;26(6):371-5. doi: 10.11919/j.issn.1002-0829.214171
- Weston SJ, Ritchie SJ, Rohrer JM, Przybylski AK. Recommendations for Increasing the Transparency of Analysis of Preexisting Data Sets. *Adv Methods Pract Psychol Sci.* 2019 Sep;2(3):214-227. doi: 10.1177/2515245919848684
- Liu G, Bordelon B, Nagar R, Sarti J, Nguyen U, Boettcher J. *Data Quality Literacy: A Guidebook.* 2024. doi: 10.31219/osf.io/ruawm
- Rice R, Southall J. *The Data Librarian's Handbook*. London, UK: Facet Publishing; 2016. 192p.



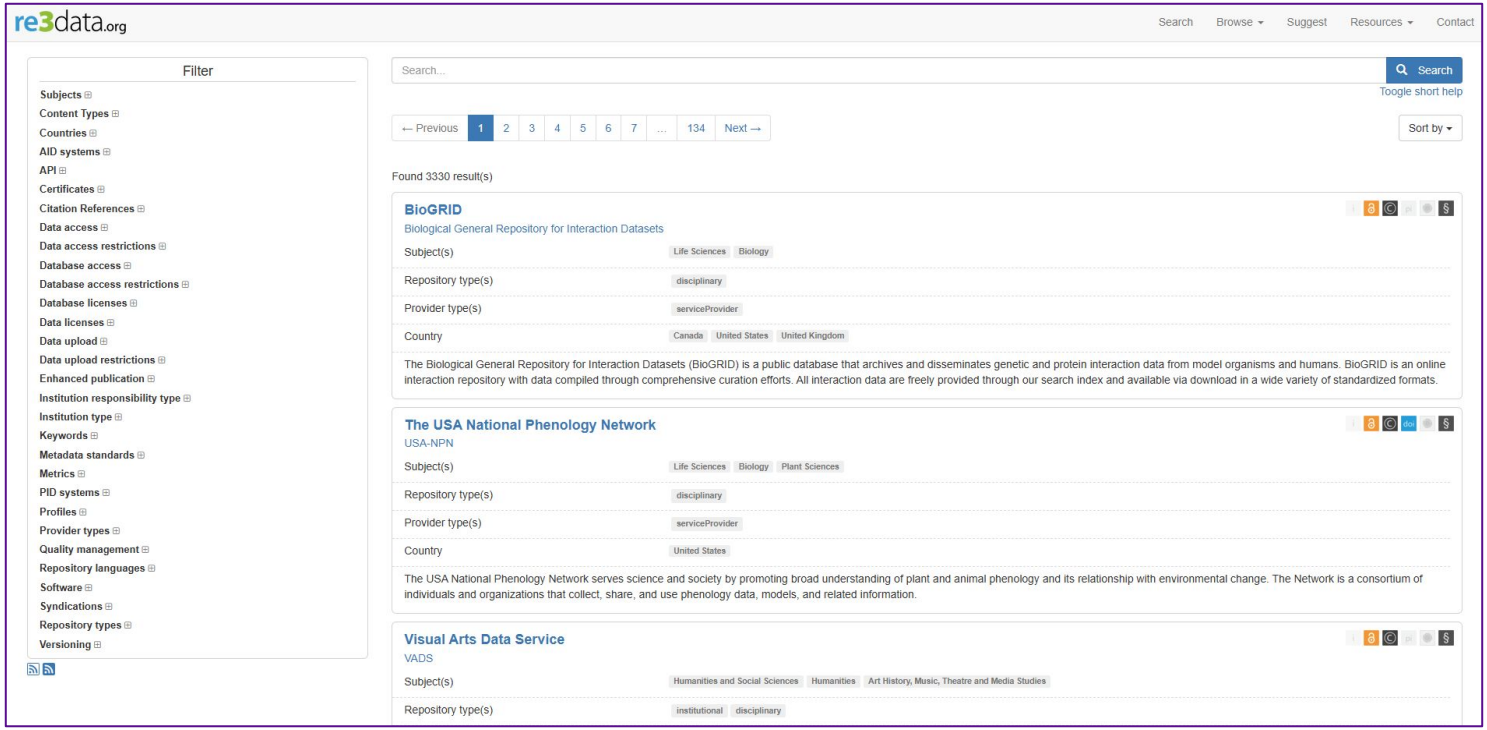
Launching points for a dataset search

- Data repositories
- Data catalogs
- Government data portals
- Professional associations
- Publications
- Search engines

Launching points for a dataset search

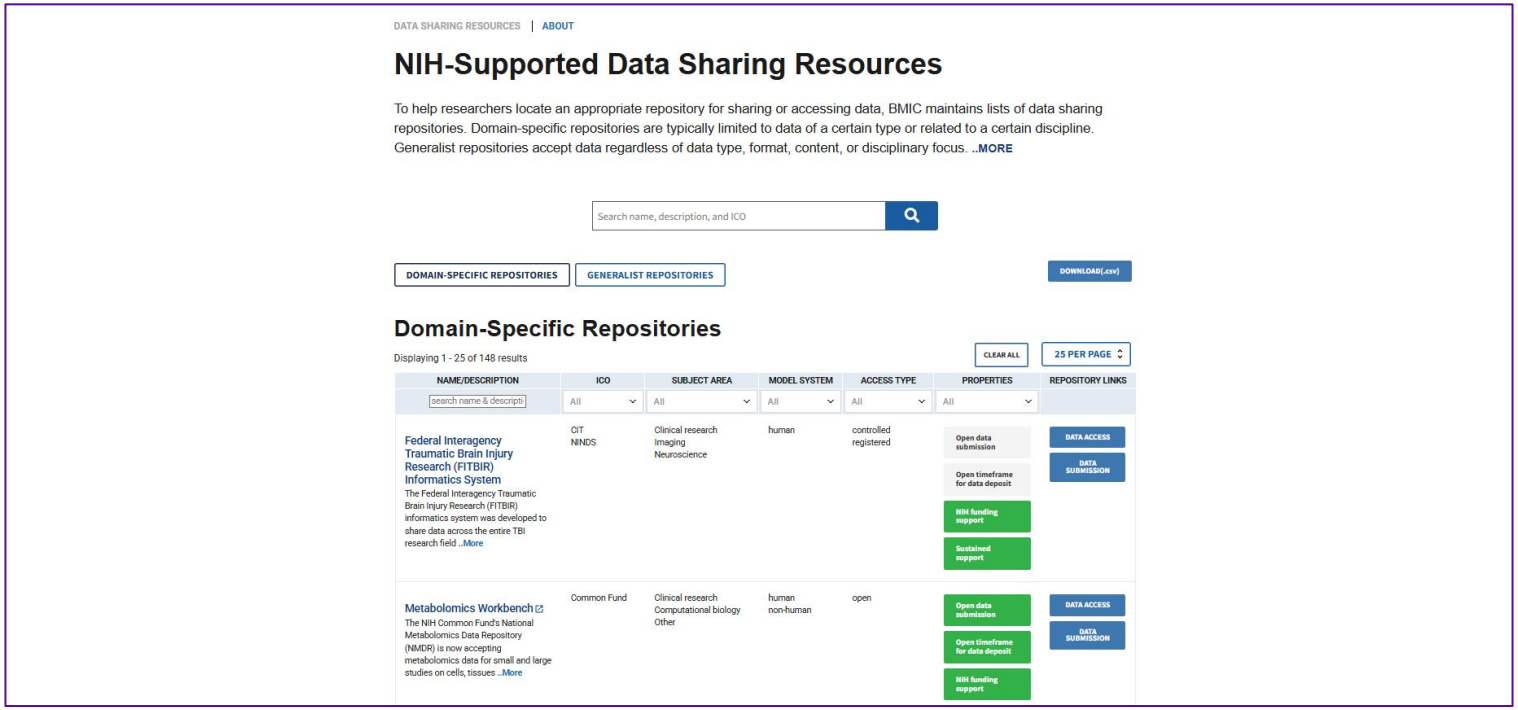
- **Data repositories**
- Data catalogs
- Government data portals
- Professional associations
- Publications
- Search engines

re3data.org



https://www.re3data.org/

NIH-supported repositories



DATA SHARING RESOURCES | ABOUT

NIH-Supported Data Sharing Resources

To help researchers locate an appropriate repository for sharing or accessing data, BMIC maintains lists of data sharing repositories. Domain-specific repositories are typically limited to data of a certain type or related to a certain discipline. Generalist repositories accept data regardless of data type, format, content, or disciplinary focus. ..MORE

Search name, description, and ICO

DOMAIN-SPECIFIC REPOSITORIES | GENERALIST REPOSITORIES | DOWNLOAD(.csv)

Domain-Specific Repositories

Displaying 1 - 25 of 148 results | CLEAR ALL | 25 PER PAGE

NAME/DESCRIPTION	ICO	SUBJECT AREA	MODEL SYSTEM	ACCESS TYPE	PROPERTIES	REPOSITORY LINKS
search name & descripti	All	All	All	All	All	
Federal Interagency Traumatic Brain Injury Research (FITBIR) Informatics System The Federal Interagency Traumatic Brain Injury Research (FITBIR) informatics system was developed to share data across the entire TBI research field ..More	CIT NINDS	Clinical research Imaging Neuroscience	human	controlled registered	Open data submission Open timeframe for data deposit NIH funding support Sustained support	DATA ACCESS DATA SUBMISSION
Metabolomics Workbench The NIH Common Fund's National Metabolomics Data Repository (NMDR) is now accepting metabolomics data for small and large studies on cells, tissues ..More	Common Fund	Clinical research Computational biology Other	human non-human	open	Open data submission Open timeframe for data deposit NIH funding support	DATA ACCESS DATA SUBMISSION

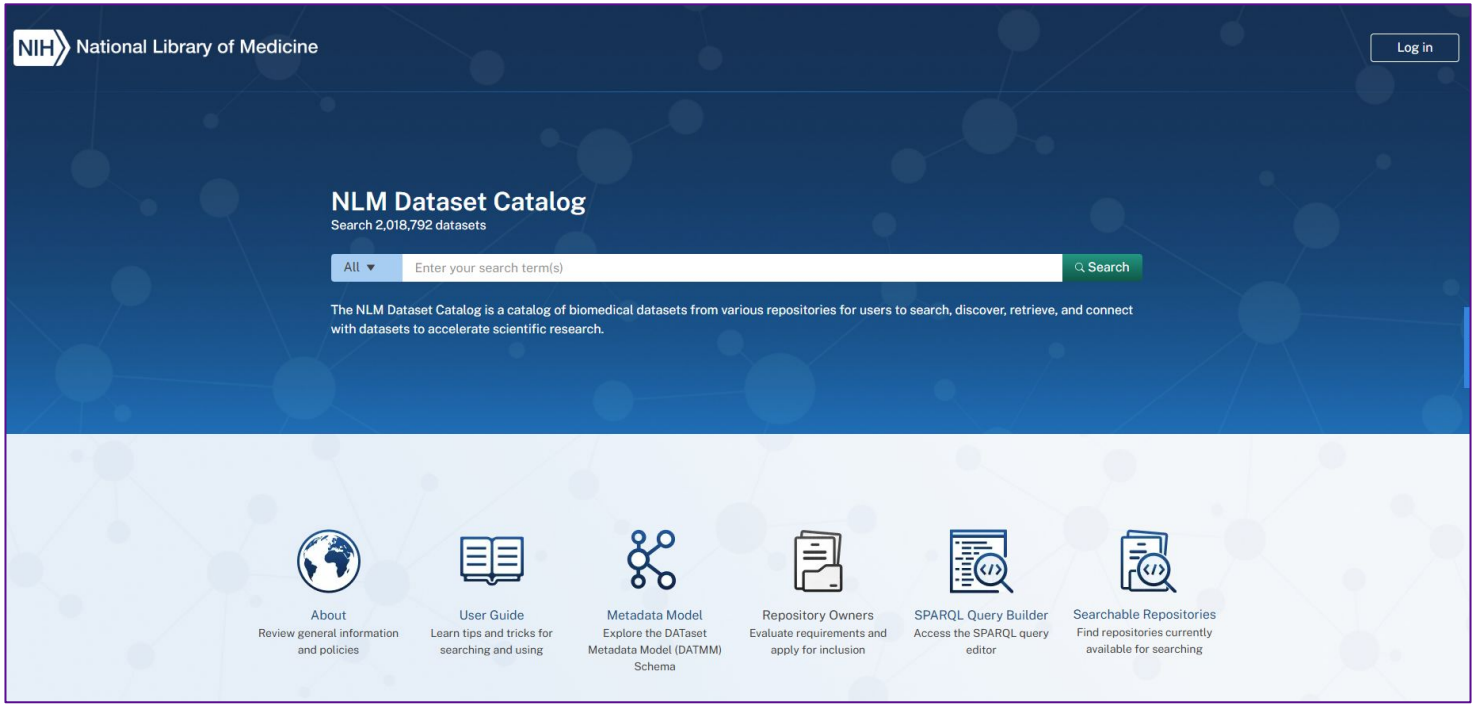
https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html

NYU Langone Health

Launching points for a dataset search

- Data repositories
- **Data catalogs**
- Government data portals
- Professional associations
- Publications
- Search engines

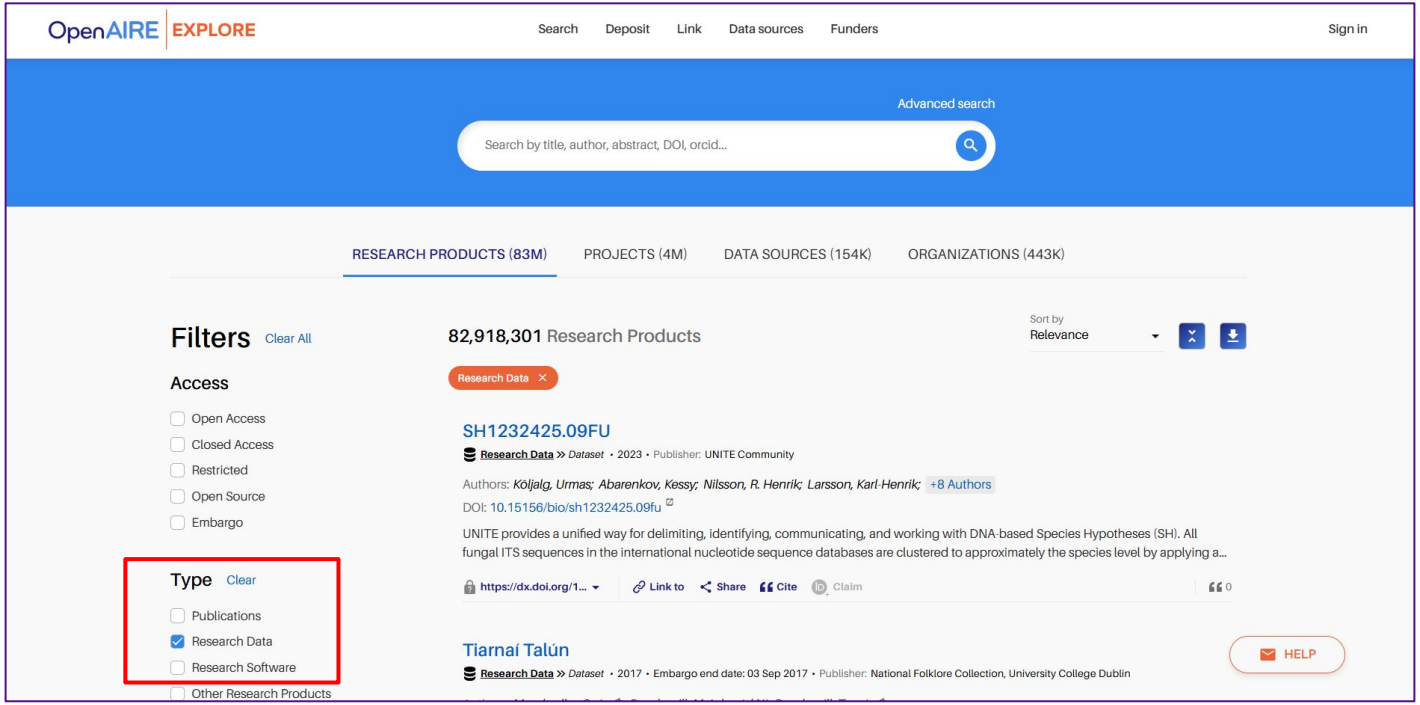
National Library of Medicine (NLM) Dataset Catalog



https://datasetcatalog.nlm.nih.gov/

NYU Langone Health

OpenAIRE Explore



https://explore.openaire.eu/search/find/research-outcomes

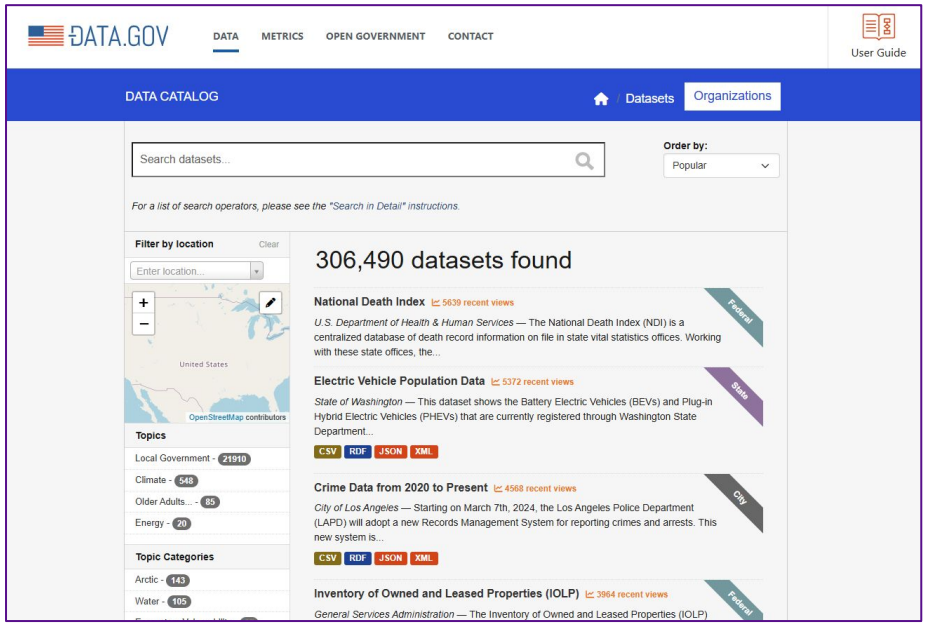
NYU Langone Health

Launching points for a dataset search

- Data repositories
- Data catalogs
- **Government data portals**
- Professional associations
- Publications
- Search engines

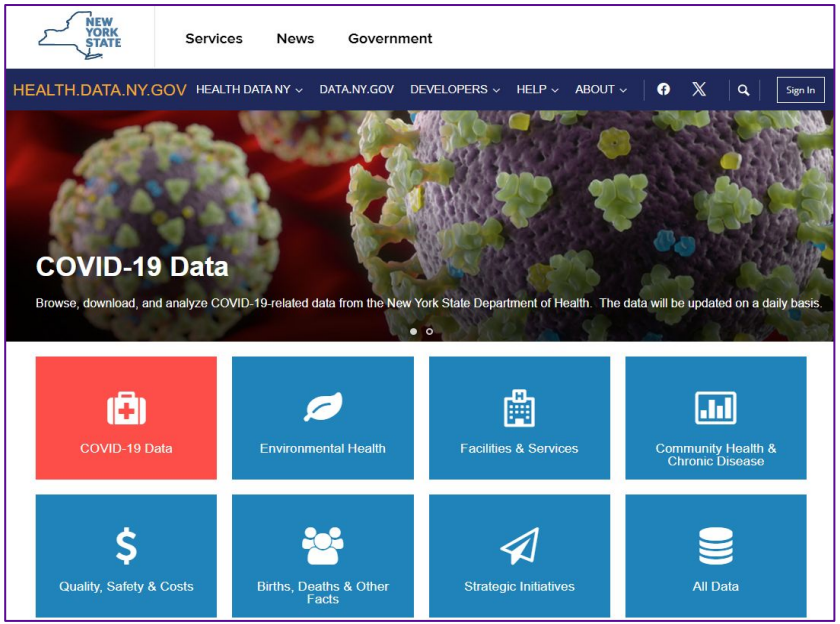
Government data portals

National



https://catalog.data.gov/dataset/

State

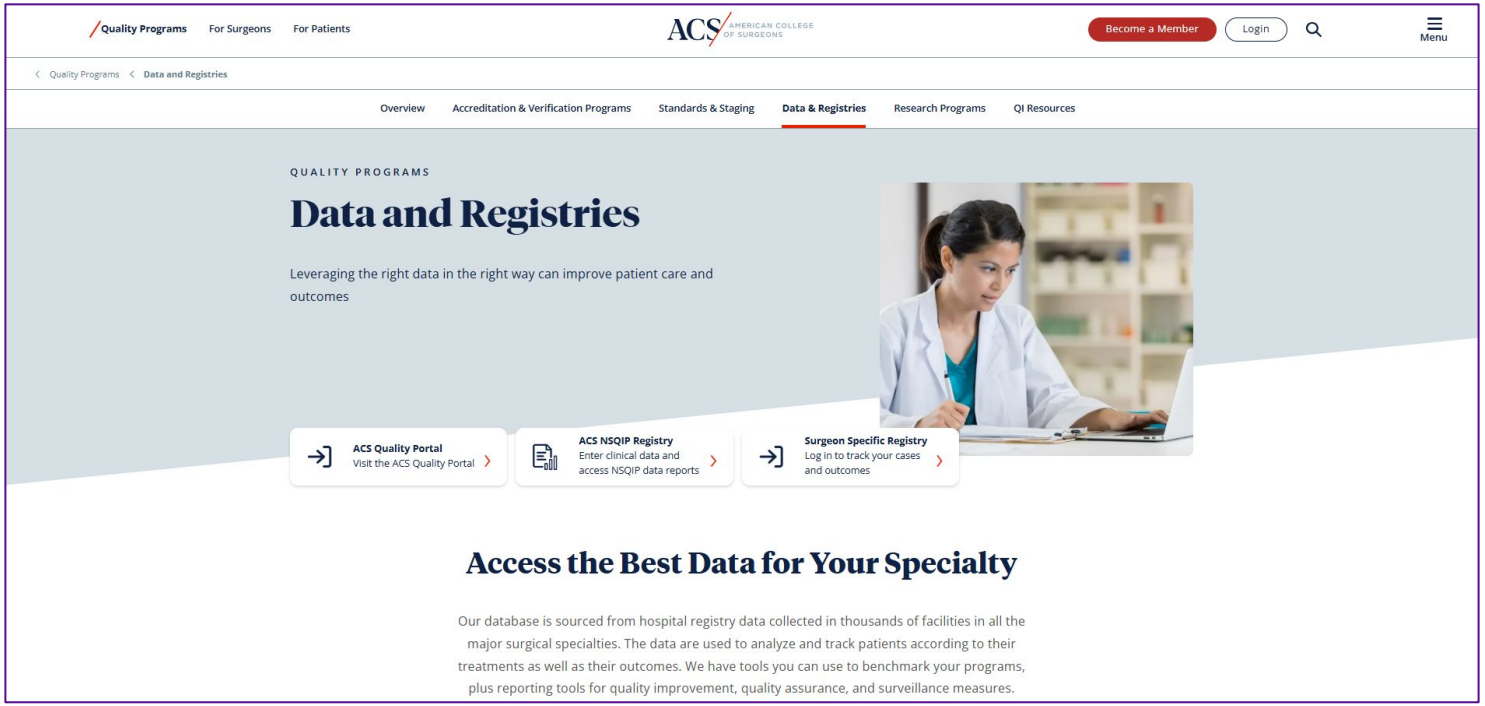


https://www.health.data.ny.gov/

Launching points for a dataset search

- Data repositories
- Data catalogs
- Government data portals
- **Professional associations**
- Publications
- Search engines

American College of Surgeons



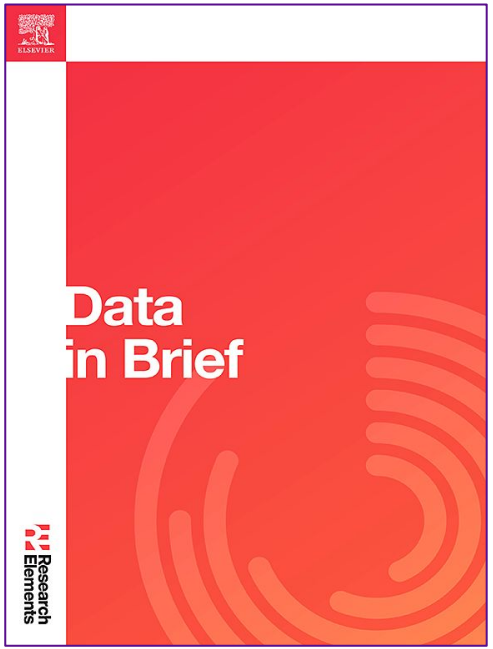
https://www.facs.org/quality-programs/data-and-registries/

NYU Langone Health

Launching points for a dataset search

- Data repositories
- Data catalogs
- Government data portals
- Professional associations
- **Publications**
- Search engines

Data journals



https://www.sciencedirect.com/journal/data-in-brief

(GIGA)n SCIENCE

https://academic.oup.com/gigascience

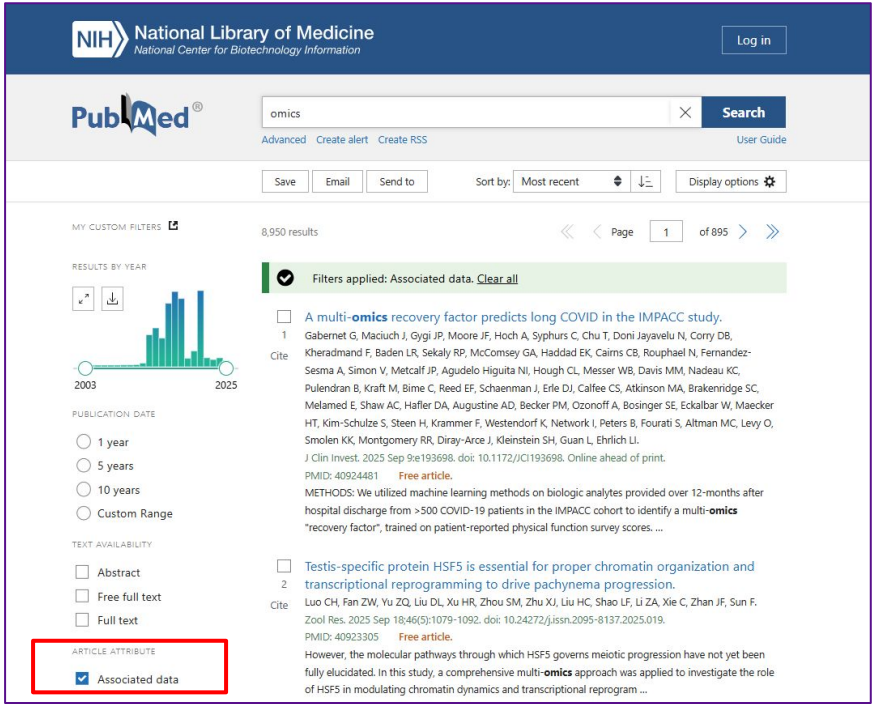
SCIENTIFIC DATA

https://www.nature.com/sdata/

BMC Research Notes

https://bmcresnotes.biomedcentral.com/

PubMed and PubMed Central



https://pubmed.ncbi.nlm.nih.gov/

By Article Property

Search or append a search with any of the following queries to find articles with these specific properties in PMC:

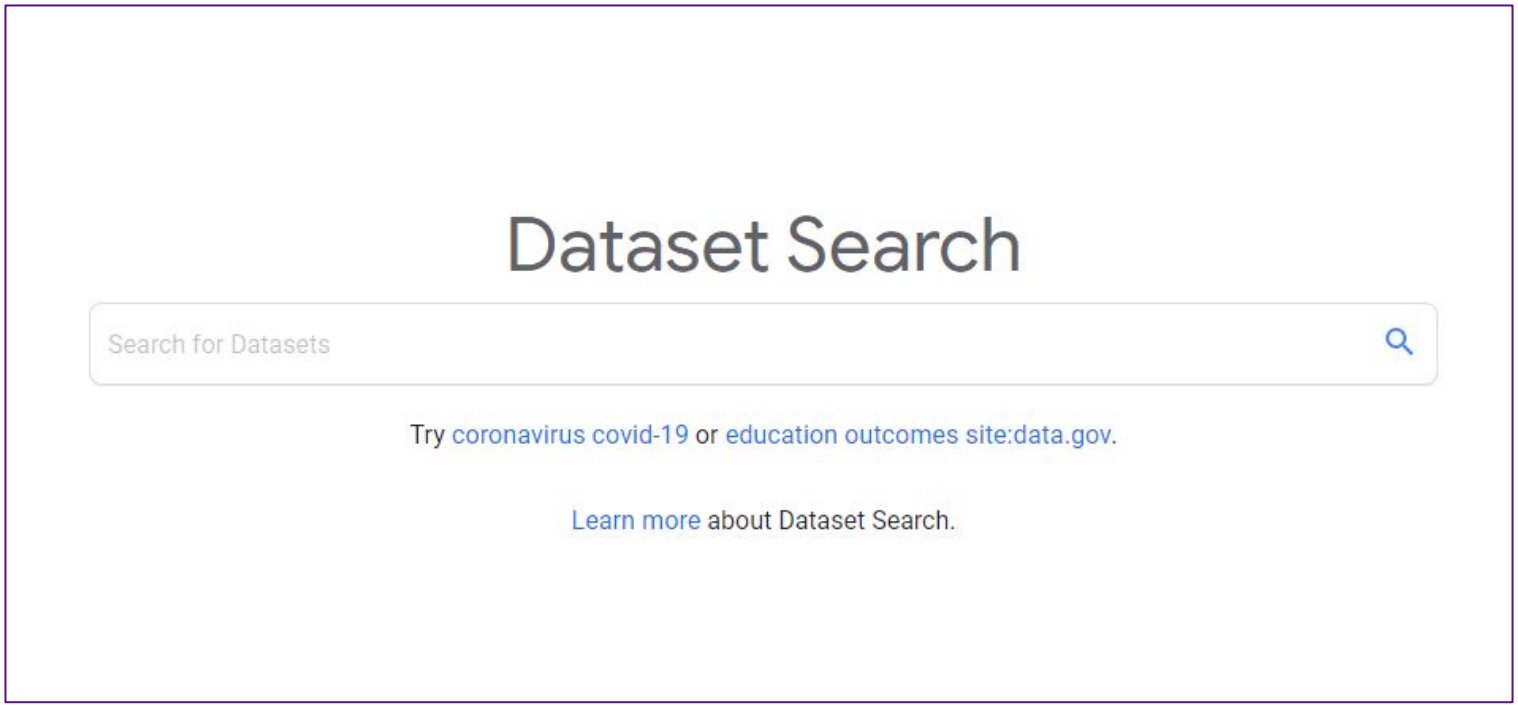
Article property	Search Query
Has supplementary material	hassuppdata
Has a "Data Availability" and/or a "Data Accessibility" section	hasdataavail
Has references tagged with a proper type to be rendered as data citations	hasdatacitations
Has supplementary material and/or data availability/accessibility section (best option for finding papers that may have openly accessible datasets associated)	hasassociateddata
Has abstract	hasabstract
Has plain language summary	hasplainlanguagesummary
Has a conflict of interest statement	hascoi
Has PDF	"has pdf"[filter]

https://pmc.ncbi.nlm.nih.gov/

Launching points for a dataset search

- Data repositories
- Data catalogs
- Government data portals
- Professional associations
- Publications
- **Search engines**

Google Dataset Search



https://datasetsearch.research.google.com

Resource roundup:
Launching points for a dataset search

- Data repositories
  - Re3data: https://re3data.org
  - NIH-supported: https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html
- Data catalogs
  - NLM Dataset Catalog: https://datasetcatalog.nlm.nih.gov/
  - OpenAIRE: https://explore.openaire.eu/search/find/research-outcomes
- Publications
  - PubMed: https://pubmed.ncbi.nlm.nih.gov/
  - PubMed Central: https://www.ncbi.nlm.nih.gov/pmc/
- Search engines
  - Google Dataset Search: https://datasetsearch.research.google.com/

\+ Data journals, government websites, and websites for professional associations

Obstacles to finding and accessing data

Obstacles to finding and accessing data

- Research data is infrequently shared

Research data is infrequently shared*

Jiao C, Li K, Fang Z. Data sharing practices across knowledge domains: A dynamic examination of data availability statements in PLOS ONE publications. *J Inf Sci*. 2022 Jun 29;*50*(3):673-689. https://doi.org/10.1177/01655515221101830

Table 5. Percentage distribution of data sharing methods in our study and Federer's work.

Data sharing mechanism	Our study (*n* = 127,935)	Federer's study (*n* = 47,593)
In paper and SI	48.20	45.34
In paper	17.15	24.3
Repository	13.62	15.4
Combination	4.68	4.5
Upon request	5.05	1.4
In SI	2.14	1.4

SI: supplemental information.

Characteristic	Sub-group sample size	Sharing IPD: Yes	No	Undecided	No response
Among all studies					
Overall, N = 313,990		32,210 (10.3%)	133,761 (42.6%)	35,862 (11.4%)	112,157 (35.7%)
Study type					
Clinical trials	237,147	27,232 (11.5%)	104,522 (44.1%)	22,172 (9.3%)	83,221 (35.1%)
Observational	76,835	4,978 (6.5%)	29,239 (38.1%)	13,690 (17.8%)	28,928 (37.6%)
Others	8	0 (0%)	0 (0%)	0 (0%)	8 (100%)
Funder					
NIH/Other USA Federal Agency	25,002	4,300 (17.2%)	9,198 (36.8%)	1,545 (6.2%)	9,959 (39.8%)
Industry	79,263	12,043 (15.2%)	31,770 (40.1%)	6,562 (8.3%)	28,888 (36.4%)
Others	209,725	15,867 (7.6%)	92,793 (44.2%)	27,755 (13.2%)	73,310 (35.0%)

Chen H, Zhao Y, Cao B, Petersen DJ, Valente MJ, Cen W. Breaking the silence of sharing data in medical research. *PLoS One*. 2024 May 29;19(5):e0301917. https://doi.org/10.1371/journal.pone.0301917

Biomedical data sharing requirements

Sponsors of Research

- NIH Data Management and Sharing Policy
- Office of Science and Technology Policy (OSTP) Public Access Memo
  - See guidance from Scholarly Publishing and Academic Resources Coalition (SPARC)

Publishers

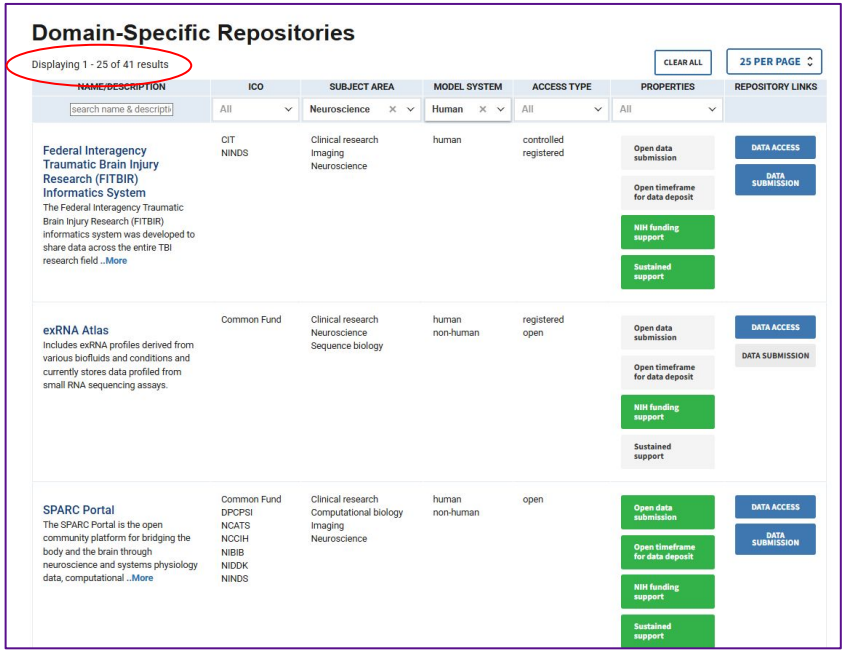
- International Committee of Medical Journal Editors (ICMJE)
- Nature Springer
- PLoS Journals

… and more

See MSKCC Library's Data Policy Finder

Obstacles to finding and accessing data

- Research data is infrequently shared
- Proliferation of data repositories



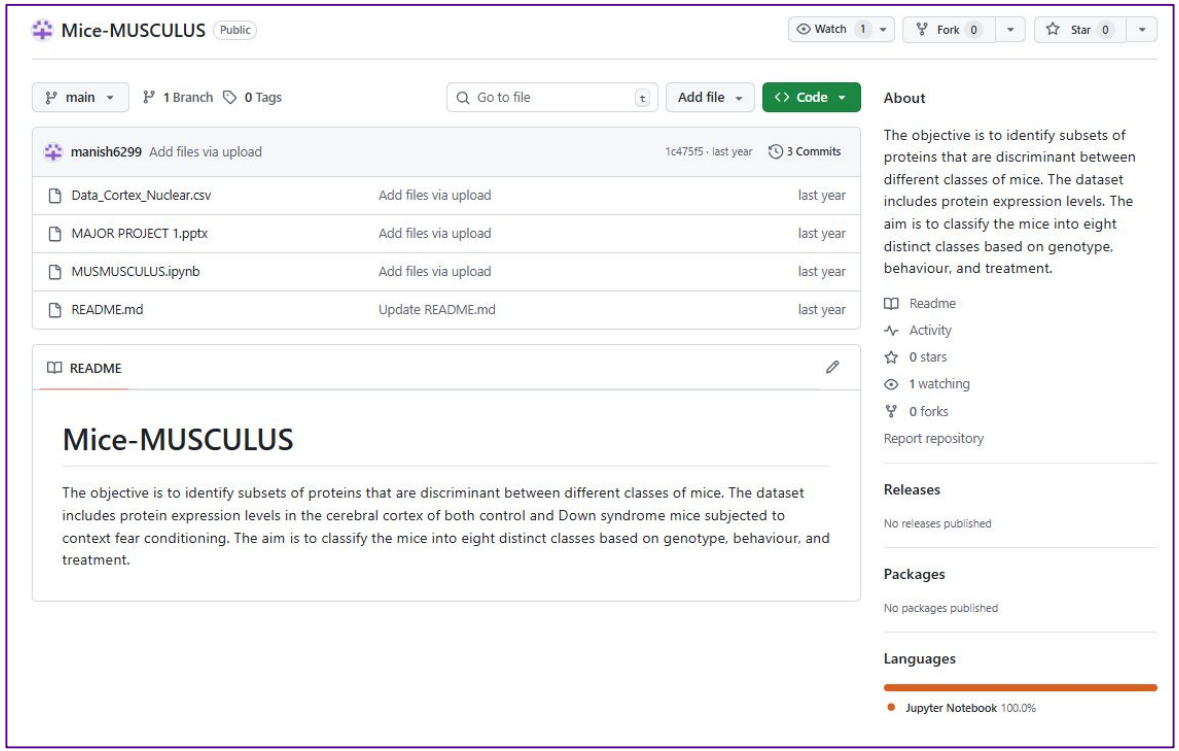
NIH Domain-Specific Repositories

NYU Langone Health

Obstacles to finding and accessing data

- Research data is infrequently shared
- Proliferation of data repositories
- Data may be poorly described

Data may be poorly described



https://github.com/manish6299/Mice-MUSCULUS

Obstacles to finding and accessing data

- Research data is infrequently shared
- Proliferation of data repositories
- Data may be poorly described
- Data may be in aggregate forms only

Obstacles to finding and accessing data

- Research data is infrequently shared
- Proliferation of data repositories
- Data may be poorly described
- Data may be in aggregate forms only
- Associated costs, timelines, and/or restrictions on access and use

Obstacles to finding and accessing data

- Research data is infrequently shared
- Proliferation of data repositories
- Data may be poorly described
- Data may be in aggregate forms only
- Associated costs, timelines, and/or restrictions on access and use
- Non-responsive creators/authors

> Availability of data and materials
>
> The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

The need for data finding experts (librarians!)

- Research data is infrequently shared
- Proliferation of data repositories
- Data may be poorly described
- Data may be in aggregate forms only
- Associated costs, timelines, and/or restrictions on access and use
- Non-responsive creators/authors

- Navigating data sources
- Exploring research feasibility
- Providing data literacy education

Enabling secondary data use for novel analyses (e.g., evidence synthesis, machine learning)

Understanding researchers' data requests

Conversation starters

- Who needs the data?
  - Level of expertise with the subject matter and/or data
  - Availability of funding
- What data is needed?
  - Data attributes (e.g., subject, level of aggregation, data collection period, locale)
  - Data type (e.g., images, surveys, electronic health records)
  - Research plan (e.g., research question(s), methodology)
- When is it needed?
  - Timelines and deadlines

Reference: Data Quality Literacy - A Guidebook

Activity:

Pick one query and consider:

- Is there enough information to understand the patron's need or request?
  - If not, what follow-up questions would you ask?
- What search strategies would you employ?
- What recommendations or next steps would you share with the patron?

We will reconvene in ~10 minutes!

Case #1 - From a postdoctoral fellow
I would like to get datasets related to air quality and cardiovascular disease risk.

Case #2 - From a medical resident
I am trying to research the trends within Crohn's disease management with [name of investigator]. He wants me to use data from the Centers of Medicare and Medicaid Services to look at the number of cases in the past 5-10 years.

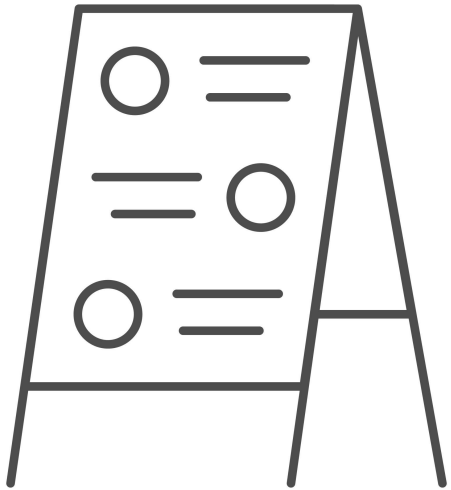
Case #3 - From a PhD student
I am conducting research on the relationship between prescription trends, health outcomes, and mental health. As part of my study, I am seeking access to large-scale healthcare data sources and sales data on [drug A] and [drug B]. Do we have access to this data?

Case #4 - From a medical resident
Hello, I am interested in doing a study using the 1. National (Nationwide) Inpatient Sample (NIS) and/or 2. Nationwide Readmissions Database (NRD). Is there any way I can get access for research purposes? Thank you for your help!

Beyond the library space: Outreach and collaboration

Collaboration strategies

- Library guides for specific research areas
- Guest lectures, workshops, or training sessions to highlight datasets or data literacy skills
- Co-hosting programming to highlight datasets or data literacy skills



Finding collaborators

Consultations

- Who (people, departments, labs, etc) reaches out with data finding/reference questions?
- What datasets or types of data are most commonly requested?

Gap or needs analysis

- What skills or resources would users/patrons like to acquire from the library?

Outreach

- In what way(s) would you like to work with others outside of the library?

Potential collaborators

- Academic departments
- Clinical and Translational Science Award (CTSA) programs
- Information technology (IT)
- Nursing/research nurses
- Research support offices

Collaboration case study #1

Annual invited talk for an intensive training program coordinated by the Clinical and Translational Science Institute.

Scope: Healthcare professionals acquiring skills in clinical research methods.

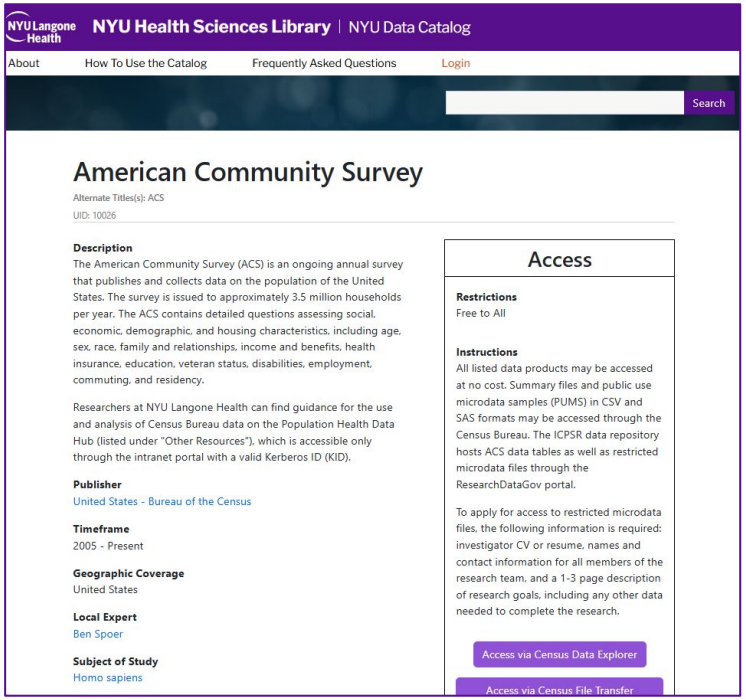
Topics:

1. The Data Sharing Landscape
2. Data Resources
3. Challenges to Finding Data
4. Accessing Data
5. Making Your Data Discoverable

Collaboration case study #2

NYU Data Catalog and the Department of Population Health (DPH):

- Early records described secondary data from US Census, National Health and Nutrition Examination Survey, Health and Retirement Study, among others.
- Records from these external datasets include a local expert.
- Currently working with DPH to catalog new datasets purchased or licensed by the department.



https://datacatalog.med.nyu.edu/dataset/10026

Collaboration case study #3

Data Day to Day workshops

- Periodic, library-organized workshops with a central theme.
- Utilized a couple of approaches to find guest speakers for a recurring series featuring secondary data:
  - Existing partnerships
  - Referrals
  - Cold emailing

Atherosclerosis Risk in Communities (ARIC) Study
Josef Coresh, MD, PhD
Optimal Aging Institute

Epic Cosmos
Andrew Fair, MS, ScM
Krutika Pandit, MS
Dept of Population Health

Perlmutter Cancer Center Data Hub
Rimma Belenkaya, MS, MA
Perlmutter Cancer Center

Q and A

Michelle Yee, MPH
michelle.yee@nyulangone.org