

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

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

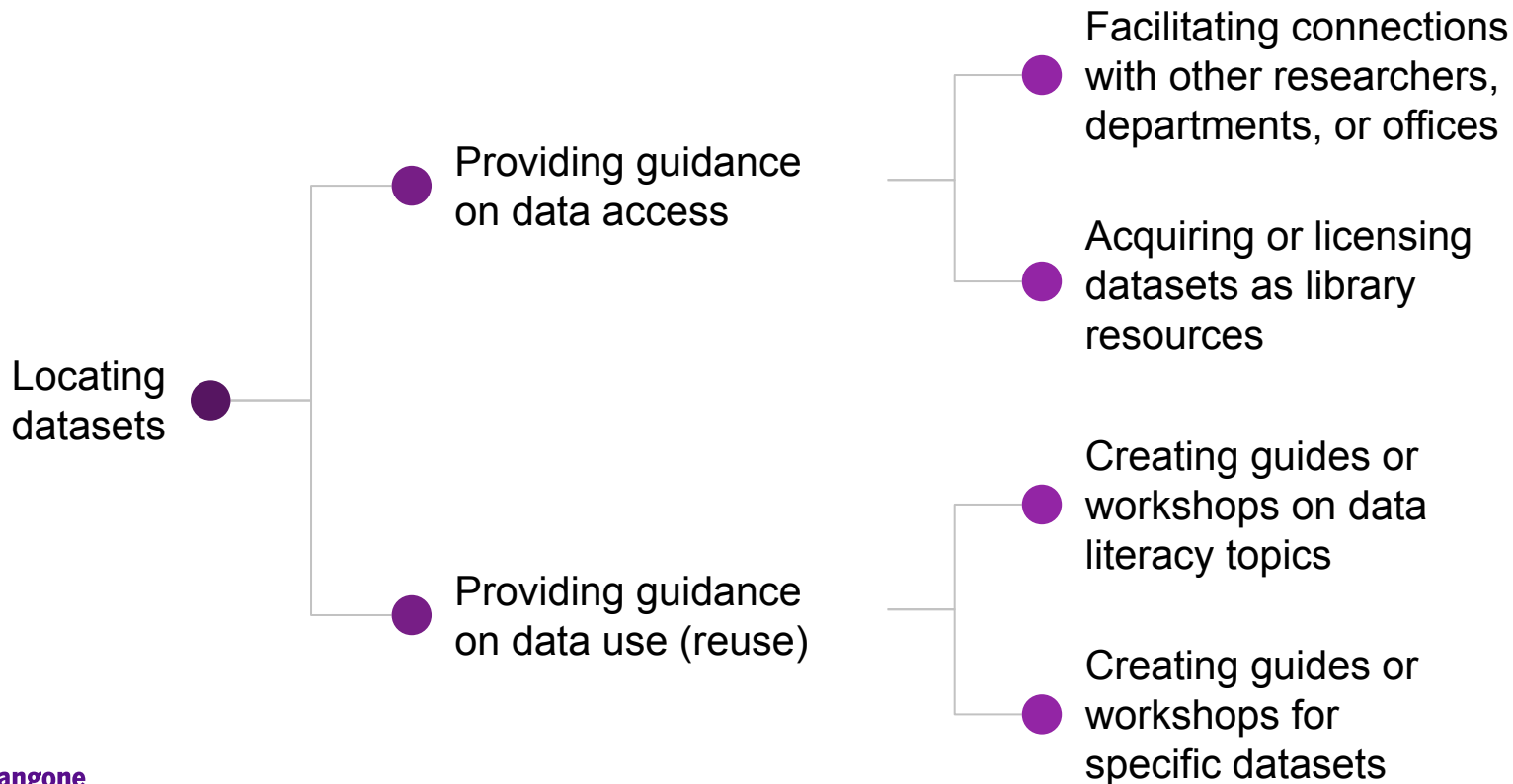
- Identify and explain different pathways to providing data services.
- Explore how low cost, high impact services like finding data can provide a foundation for building data services.

Data finding

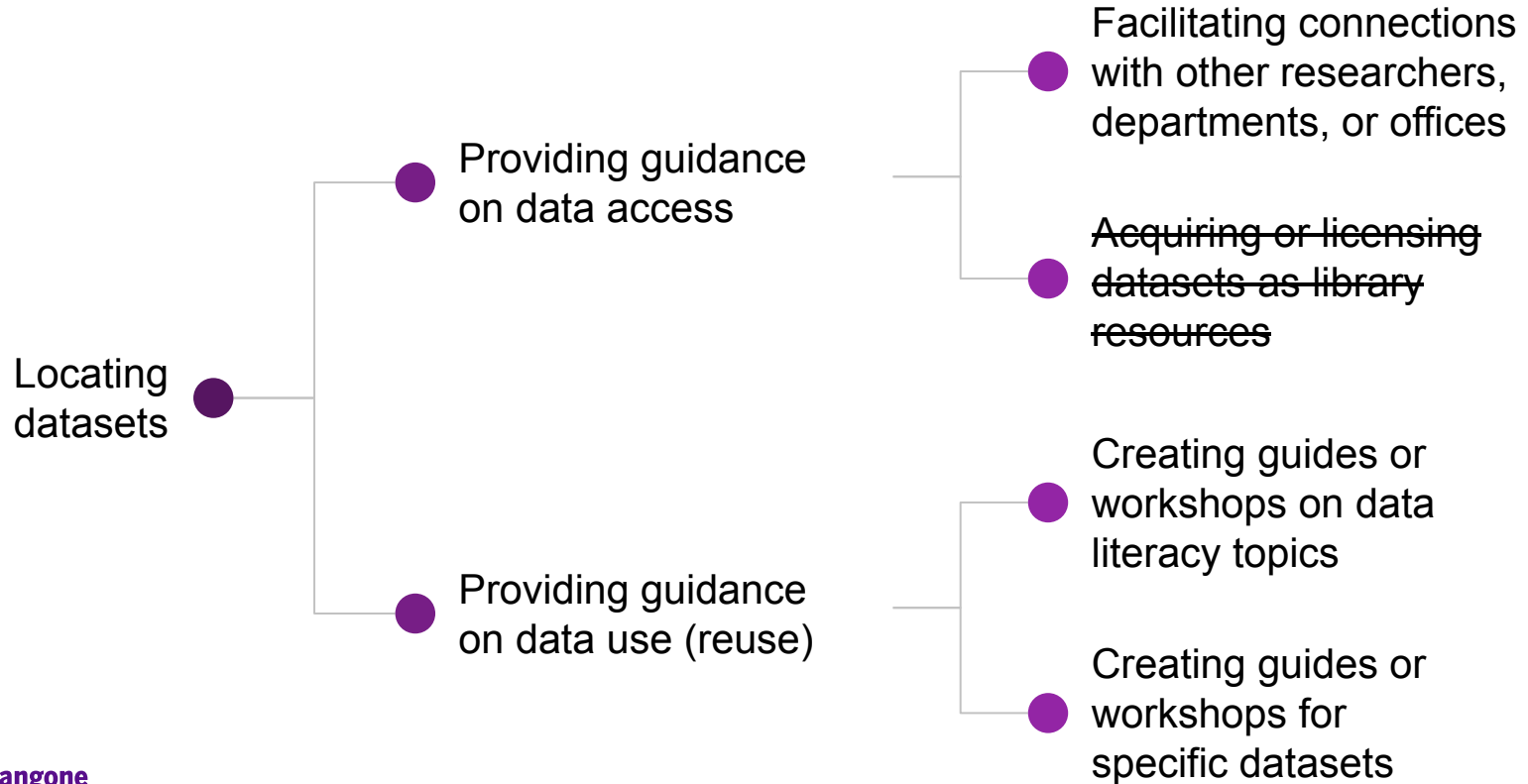
- A reference service: locating or identifying data sources to support a user need
- Not to be confused with processes that enable other people to find your/user/institution's data
 - October 23, 2025: The NIH Data Management and Sharing Policy
 - Archived webinar: [“Understanding Data Discovery and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



Data finding may encompass...



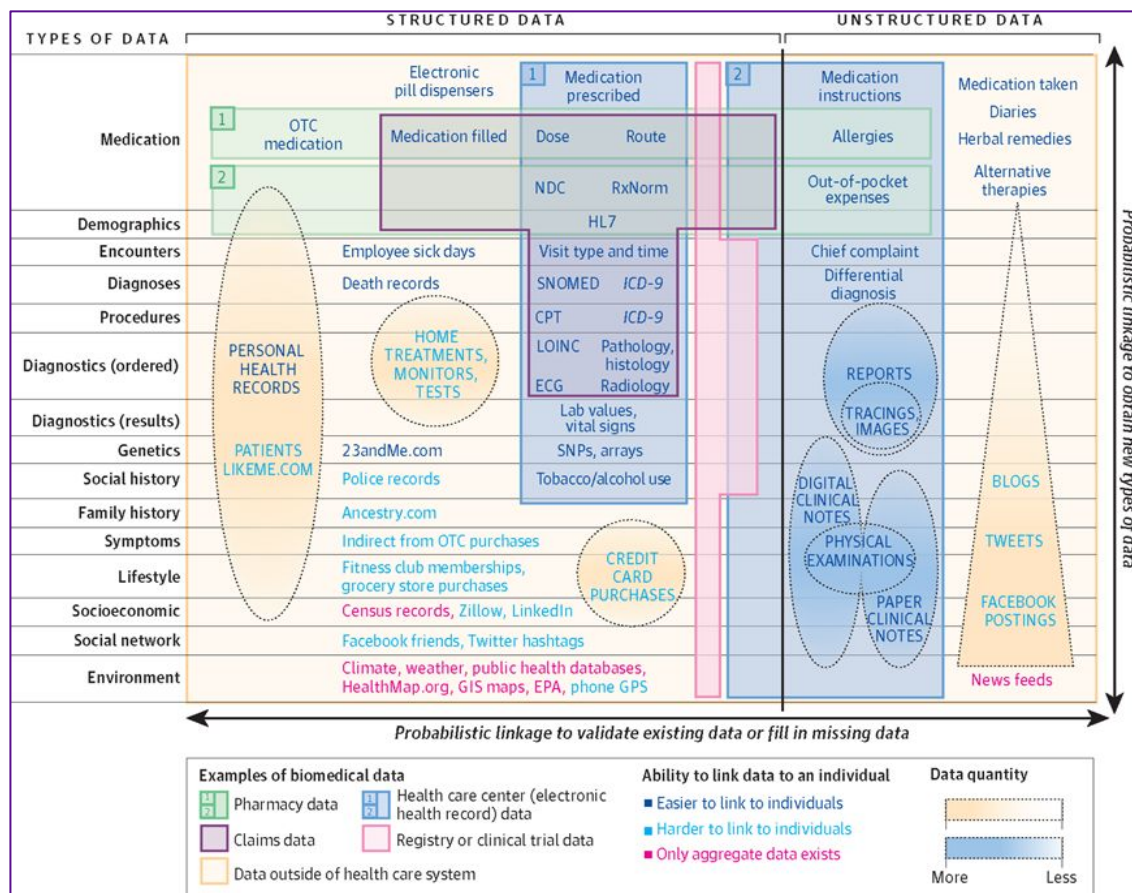
Low cost services



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](https://doi.org/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

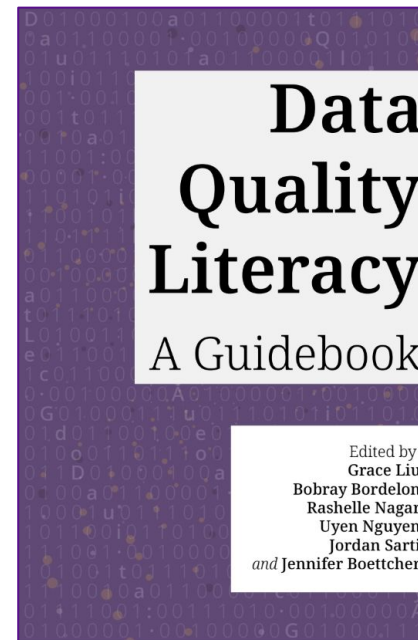
- Misinterpretation or perpetuation of errors and biases

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](https://doi.org/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](https://doi.org/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](https://doi.org/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

The screenshot displays the re3data.org homepage. On the left is a 'Filter' sidebar with a list of categories such as Subjects, Content Types, Countries, and AID systems, each with a dropdown arrow. The main content area features a search bar at the top right with a 'Search' button. Below the search bar is a pagination control showing '1' as the current page out of 134 results. The results list includes three entries:

- BioGRID**: Biological General Repository for Interaction Datasets. Subject(s): Life Sciences, Biology. Repository type(s): disciplinary. Provider type(s): serviceProvider. Country: Canada, United States, United Kingdom. Description: The Biological General Repository for Interaction Datasets (BioGRID) is a public database that archives and disseminates genetic and protein interaction data from model organisms and humans.
- The USA National Phenology Network**: USA-NPN. Subject(s): Life Sciences, Biology, Plant Sciences. Repository type(s): disciplinary. Provider type(s): serviceProvider. Country: United States. Description: The USA National Phenology Network serves science and society by promoting broad understanding of plant and animal phenology and its relationship with environmental change.
- Visual Arts Data Service**: VADS. Subject(s): Humanities and Social Sciences, Humanities, Art History, Music, Theatre and Media Studies. Repository type(s): institutional, disciplinary.

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](#)

[DOMAIN-SPECIFIC REPOSITORIES](#) [GENERALIST REPOSITORIES](#) [DOWNLOAD \(.csv\)](#)

Domain-Specific Repositories

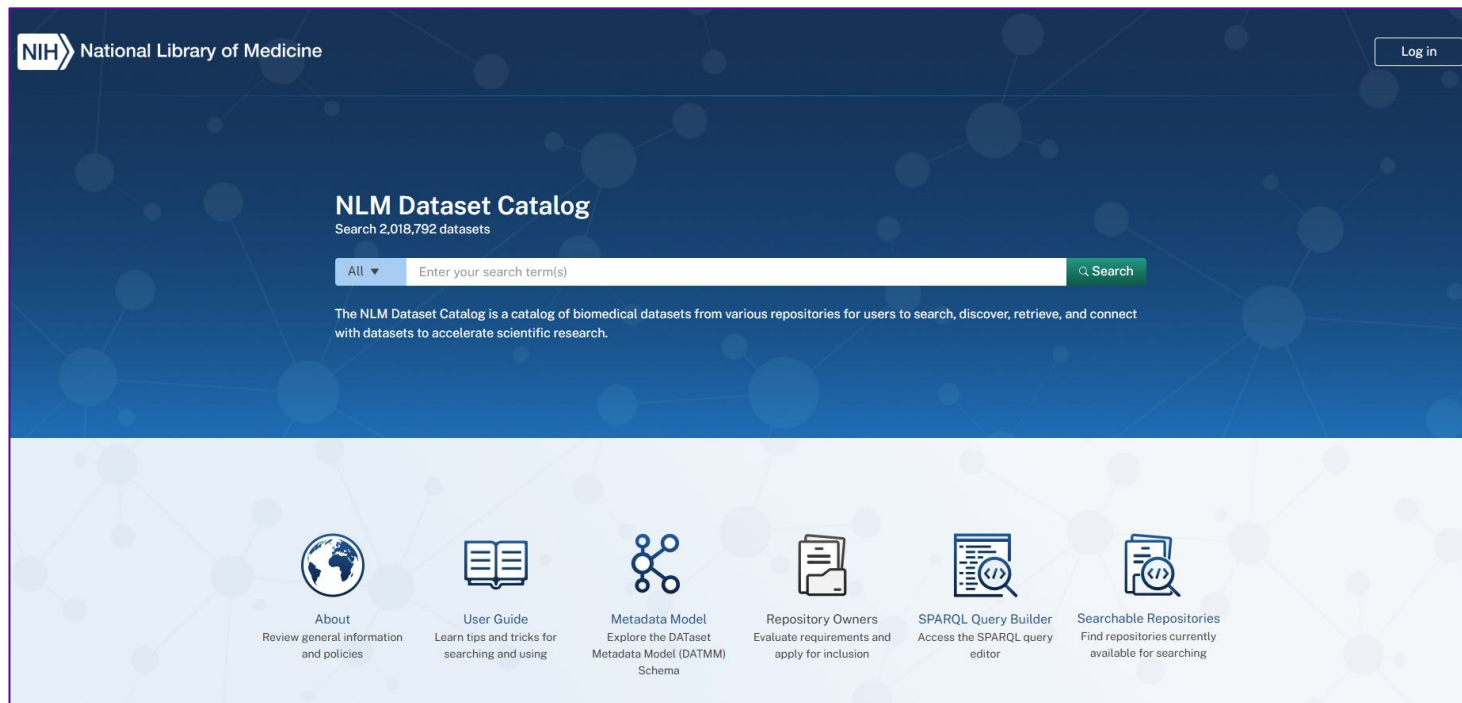
Displaying 1 - 25 of 148 results CLEAR ALL 25 PER PAGE

NAME/DESCRIPTION	ICO	SUBJECT AREA	MODEL SYSTEM	ACCESS TYPE	PROPERTIES	REPOSITORY LINKS
search name & description	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	QIT NNDS	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

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



The screenshot shows the NLM Dataset Catalog homepage. At the top left is the NIH logo and the text "National Library of Medicine". At the top right is a "Log in" button. The main heading is "NLM Dataset Catalog" with the subtitle "Search 2,018,792 datasets". Below this is a search bar with a dropdown menu set to "All", a text input field for "Enter your search term(s)", and a green "Search" button. A paragraph below the search bar states: "The NLM Dataset Catalog is a catalog of biomedical datasets from various repositories for users to search, discover, retrieve, and connect with datasets to accelerate scientific research." At the bottom, there are six icons with corresponding text links: a globe for "About" (Review general information and policies), an open book for "User Guide" (Learn tips and tricks for searching and using), a network diagram for "Metadata Model" (Explore the DATaset Metadata Model (DATMM) Schema), a folder for "Repository Owners" (Evaluate requirements and apply for inclusion), a document with code for "SPARQL Query Builder" (Access the SPARQL query editor), and a document with a magnifying glass for "Searchable Repositories" (Find repositories currently available for searching).

NIH National Library of Medicine

Log in

NLM Dataset Catalog

Search 2,018,792 datasets

All Enter your search term(s) Search

The NLM Dataset Catalog is a catalog of biomedical datasets from various repositories for users to search, discover, retrieve, and connect with datasets to accelerate scientific research.

-  **About**
Review general information and policies
-  **User Guide**
Learn tips and tricks for searching and using
-  **Metadata Model**
Explore the DATaset Metadata Model (DATMM) Schema
-  **Repository Owners**
Evaluate requirements and apply for inclusion
-  **SPARQL Query Builder**
Access the SPARQL query editor
-  **Searchable Repositories**
Find repositories currently available for searching

OpenAIRE Explore

The screenshot displays the OpenAIRE Explore homepage. At the top, the navigation bar includes the OpenAIRE logo, the word 'EXPLORE', and links for Search, Deposit, Link, Data sources, Funders, and Sign in. Below this is a large blue header with an 'Advanced search' button and a search bar containing the text 'Search by title, author, abstract, DOI, orcid...'. The main content area features a tabbed interface with 'RESEARCH PRODUCTS (83M)' selected, alongside 'PROJECTS (4M)', 'DATA SOURCES (154K)', and 'ORGANIZATIONS (443K)'. On the left, a 'Filters' section is visible, with a red box highlighting the 'Type' filter. Under 'Type', the 'Research Data' option is selected with a blue checkmark. Other filter categories include 'Access' (Open Access, Closed Access, Restricted, Open Source, Embargo) and 'Sort by' (Relevance). The search results for 'Research Products' show a total of 82,918,301 items. A specific result is highlighted: 'SH1232425.09FU', a Research Data Dataset from 2023 published by UNITE Community. The authors listed are Kõljalg, Urmas; Abarenkov, Kessy; Nilsson, R. Henrik; Larsson, Karl-Henrik; and +8 Authors. The DOI is 10.15156/bio/sh1232425.09fu. The description states: 'UNITE provides a unified way for delimiting, identifying, communicating, and working with DNA-based Species Hypotheses (SH). All fungal ITS sequences in the international nucleotide sequence databases are clustered to approximately the species level by applying a...'. Below the description are links for the DOI, a 'Link to' button, and social media icons for Share, Cite, and Claim. At the bottom right, there is a 'HELP' button.

OpenAIRE EXPLORE

Search Deposit Link Data sources Funders Sign in

Advanced search

Search by title, author, abstract, DOI, orcid...

RESEARCH PRODUCTS (83M) PROJECTS (4M) DATA SOURCES (154K) ORGANIZATIONS (443K)

Filters Clear All

Access

- ☐ Open Access
- ☐ Closed Access
- ☐ Restricted
- ☐ Open Source
- ☐ Embargo

Type Clear

- ☐ Publications
- ☒ Research Data
- ☐ Research Software
- ☐ Other Research Products

82,918,301 Research Products

Research Data x

SH1232425.09FU

Research Data » Dataset • 2023 • Publisher: UNITE Community

Authors: Kõljalg, Urmas; Abarenkov, Kessy; Nilsson, R. Henrik; Larsson, Karl-Henrik; +8 Authors

DOI: 10.15156/bio/sh1232425.09fu

UNITE provides a unified way for delimiting, identifying, communicating, and working with DNA-based Species Hypotheses (SH). All fungal ITS sequences in the international nucleotide sequence databases are clustered to approximately the species level by applying a...

<https://dx.doi.org/1...> Link to Share Cite Claim

Tiarnai Talún

Research Data » Dataset • 2017 • Embargo end date: 03 Sep 2017 • Publisher: National Folklore Collection, University College Dublin

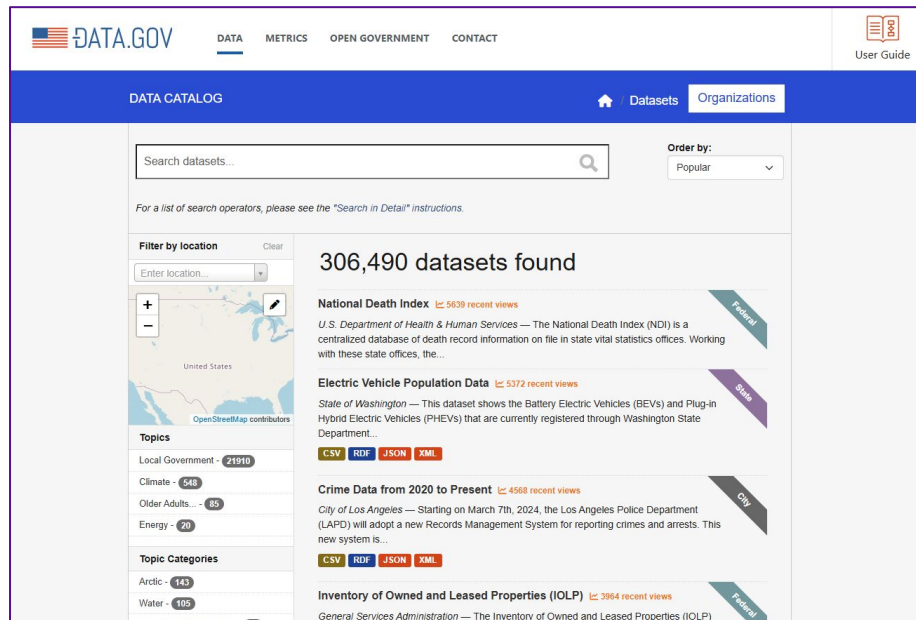
HELP

Launching points for a dataset search

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

Government data portals

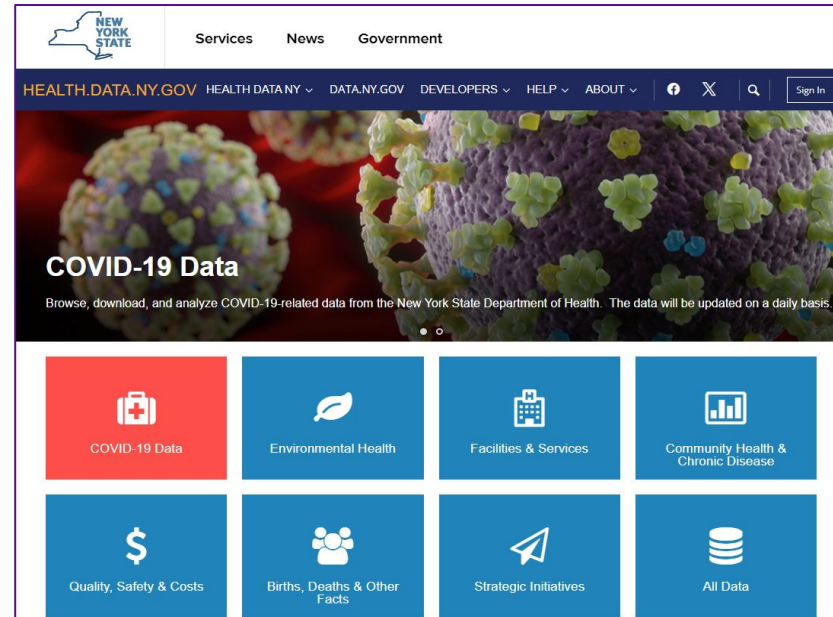
National



The screenshot shows the Data.GOV website interface. At the top, there's a navigation bar with links for DATA, METRICS, OPEN GOVERNMENT, and CONTACT. Below this is a blue header with 'DATA CATALOG' and tabs for 'Datasets' and 'Organizations'. A search bar is prominently displayed with the text 'Search datasets...'. To the right of the search bar, there's a dropdown menu for 'Order by:' set to 'Popular'. Below the search bar, a message states 'For a list of search operators, please see the "Search in Detail" instructions.' On the left side, there's a 'Filter by location' section with a map of the United States and a list of topics including Local Government (21910), Climate (546), Older Adults (85), Energy (20), and Topic Categories (Arctic 143, Water 105). The main content area displays '306,490 datasets found'. Below this, several dataset cards are visible, including 'National Death Index' (5639 recent views), 'Electric Vehicle Population Data' (5372 recent views), 'Crime Data from 2020 to Present' (4568 recent views), and 'Inventory of Owned and Leased Properties (IOLP)' (3964 recent views). Each card includes a brief description and download options (CSV, RDF, JSON, XML).

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

State



The screenshot shows the Health.Data.NY.GOV website interface. At the top, there's a navigation bar with links for Services, News, and Government. Below this is a blue header with 'HEALTH.DATA.NY.GOV' and links for HEALTH DATA NY, DATA.NY.GOV, DEVELOPERS, HELP, ABOUT, and a Sign In button. The main content area features a large banner for 'COVID-19 Data' with a background image of a virus particle. Below the banner, there's a grid of eight data categories, each with an icon and a title: COVID-19 Data, Environmental Health, Facilities & Services, Community Health & Chronic Disease, Quality, Safety & Costs, Births, Deaths & Other Facts, Strategic Initiatives, and All Data.

<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

Quality Programs For Surgeons For Patients

ACS AMERICAN COLLEGE OF SURGEONS

Become a Member Login

Menu

Quality Programs < Data and Registries

Overview Accreditation & Verification Programs Standards & Staging **Data & Registries** Research Programs QI Resources

QUALITY PROGRAMS

Data and Registries

Leveraging the right data in the right way can improve patient care and outcomes

→ ACS Quality Portal
Visit the ACS Quality Portal

→ ACS NSQIP Registry
Enter clinical data and access NSQIP data reports

→ Surgeon Specific Registry
Log in to track your cases and outcomes

Access the Best Data for Your Specialty

Our database is sourced from hospital registry data collected in thousands of facilities in all the major surgical specialties. The data are used to analyze and track patients according to their treatments as well as their outcomes. We have tools you can use to benchmark your programs, plus reporting tools for quality improvement, quality assurance, and surveillance measures.

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://bmcresearchnotes.biomedcentral.com/>

PubMed and PubMed Central

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

PubMed®

omics

Advanced Create alert Create RSS User Guide

Save Email Send to Sort by: Most recent Display options

MY CUSTOM FILTERS 12 8,950 results

RESULTS BY YEAR

2003 2025

PUBLICATION DATE

1 year
5 years
10 years
Custom Range

TEXT AVAILABILITY

Abstract
Free full text
Full text

ARTICLE ATTRIBUTE

☒ Associated data

Filters applied: Associated data. Clear all

1 A multi-omics recovery factor predicts long COVID in the IMPACC study.
Gabernet G, Maciuch J, Gygi JP, Moore JF, Hoch A, Syphurs C, Chu T, Doni Jayavelu N, Corry DB, Kheradmand F, Baden LR, Sekaly RP, McComsey GA, Haddad EK, Cairns CB, Roupael N, Fernandez-Sesma A, Simon V, Metcalf JP, Agudelo Higuita NJ, Hough CL, Messer WB, Davis MM, Nadeau KC, Pulendran B, Kraft M, Bime C, Reed EF, Schaenman J, Erle DJ, Calfee CS, Atkinson MA, Brakenridge SC, Melamed E, Shaw AC, Haffer DA, Augustine AD, Becker PM, Ozonoff A, Bostinger SE, Eckalbar W, Maelcker HT, Kim-Schulze S, Steen H, Krammer F, Westendorf K, Network I, Peters B, Fourati S, Altman MC, Levy O, Smolen KK, Montgomery RR, Diray-Arce J, Kleinstein SH, Guan L, Ehrlich LJ.
J Clin Invest. 2025 Sep 9:e193698. doi: 10.1172/JCI193698. Online ahead of print.
PMID: 40924481 Free article.

2 Testis-specific protein HSF5 is essential for proper chromatin organization and transcriptional reprogramming to drive pachynema progression.
Luo CH, Fan ZW, Yu ZQ, Liu DL, Xu HR, Zhou SM, Zhu XJ, Liu HC, Shao LF, Li ZA, Xie C, Zhan JF, Sun F.
Zool Res. 2025 Sep 18;46(5):1079-1092. doi: 10.24272/zjssn.2095-8137.2025.019.
PMID: 40923305 Free article.

However, the molecular pathways through which HSF5 governs meiotic progression have not yet been fully elucidated. In this study, a comprehensive multi-omics approach was applied to investigate the role of HSF5 in modulating chromatin dynamics and transcriptional reprogram ...

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

NIH National Library of Medicine
National Center for Biotechnology Information

Search Log in

PMC PubMed Central®

Search PMC Full-Text Archive Search in PMC

Journal List

PubMed Central® (PMC) is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM)

By Article Property

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

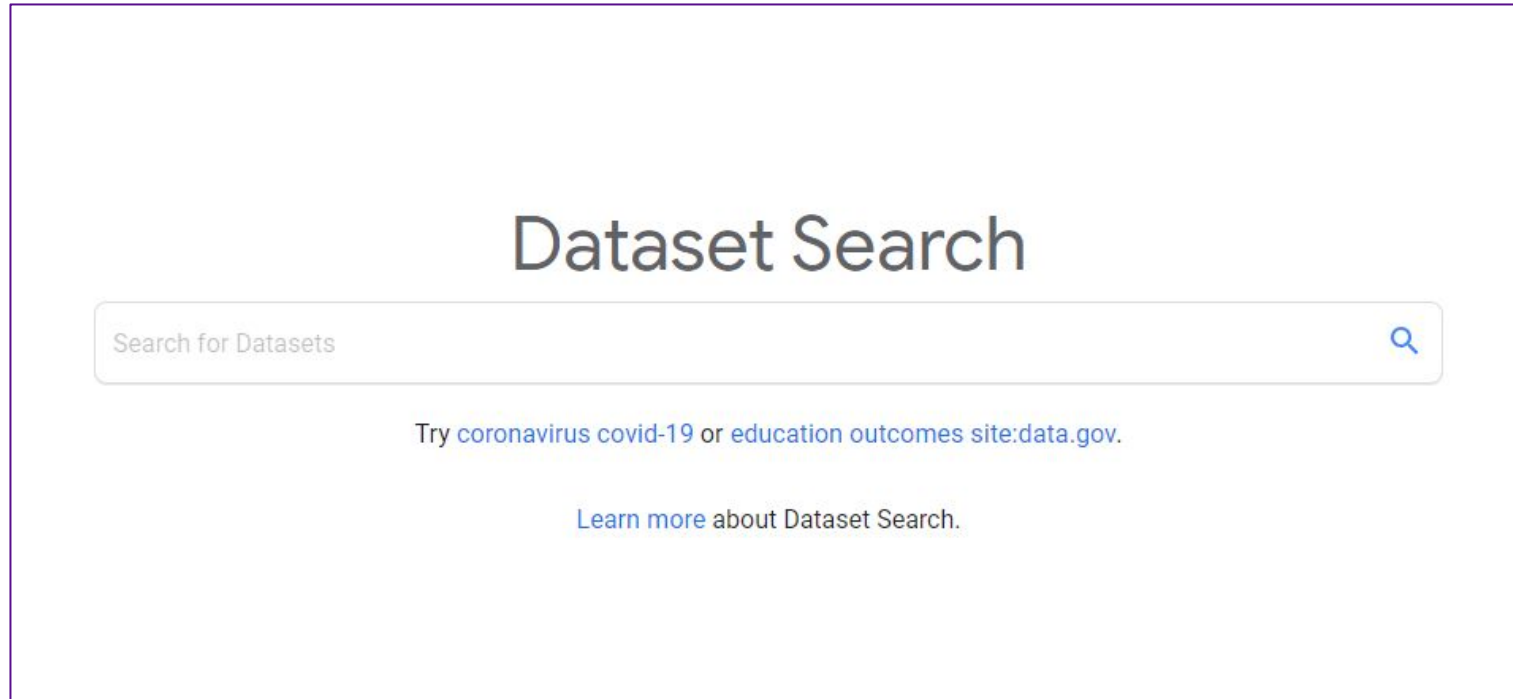
Article property	Search Query
Has supplementary material	hassuppdata
Has a "Data Availability" and/or a "Data Accessibility" section	hasdataavail
Has references tagged with a proper type to be rendered as data citations	hasdatacitations
Has supplementary material and/or data availability/accessibility 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

A screenshot of the Google Dataset Search homepage. The page has a light gray background. At the top center, the text "Dataset Search" is displayed in a large, dark gray font. Below this, there is a search bar with the placeholder text "Search for Datasets" and a magnifying glass icon on the right. Under the search bar, there is a suggestion: "Try coronavirus covid-19 or education outcomes site:data.gov." and a link "Learn more about Dataset Search." in blue text.

Dataset Search

Search for Datasets

Try coronavirus covid-19 or education outcomes site:data.gov.

[Learn more](#) about Dataset Search.

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			Sharing IPD		
	Sub-group sample size	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

Domain-Specific Repositories							
Displaying 1 - 25 of 41 results							
NAME/DESCRIPTION	ICO	SUBJECT AREA	MODEL SYSTEM	ACCESS TYPE	PROPERTIES	REPOSITORY LINKS	
search name & description	All	Neuroscience	Human	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	
exRNA Atlas Includes exRNA profiles derived from various brainfolds and conditions and currently stores data profiled from small RNA sequencing assays.	Common Fund	Clinical research Neuroscience Sequence biology	human non-human	registered open	Open data submission Open timeframe for data deposit NIH funding support Sustained support	DATA ACCESS DATA SUBMISSION	
SPARC Portal The SPARC Portal is the open community platform for bridging the body and the brain through neuroscience and systems physiology data, computational...More	Common Fund DPCPSI NCATS NCCIH NIBIB NIDDK NINDS	Clinical research Computational biology Imaging Neuroscience	human non-human	open	Open data submission Open timeframe for data deposit NIH funding support Sustained support	DATA ACCESS DATA SUBMISSION	

NIH Domain-Specific Repositories

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

The screenshot shows the GitHub repository page for 'Mice-MUSCULUS' by user 'manish6299'. The repository is public and has 1 watch, 0 forks, and 0 stars. It is on the 'main' branch with 1 branch and 0 tags. The repository contains four files: 'Data_Cortex_Nuclear.csv', 'MAJOR PROJECT 1.pptx', 'MUSMUSCULUS.ipynb', and 'README.md'. The 'README.md' file is selected and shows the following content:

Mice-MUSCULUS

The objective is to identify subsets of proteins that are discriminant between different classes of mice. The dataset includes protein expression levels in the cerebral cortex of both control and Down syndrome mice subjected to context fear conditioning. The aim is to classify the mice into eight distinct classes based on genotype, behaviour, and treatment.

The right sidebar provides additional information:

- About:** The objective is to identify subsets of proteins that are discriminant between different classes of mice. The dataset includes protein expression levels. The aim is to classify the mice into eight distinct classes based on genotype, behaviour, and treatment.
- Readme:** A link to the README file.
- Activity:** A link to the repository's activity.
- Stars:** 0 stars.
- Watching:** 1 watching.
- Forks:** 0 forks.
- Report repository:** A link to report the repository.
- Releases:** No releases published.
- Packages:** No packages published.
- Languages:** A bar chart showing the language distribution: Jupyter Notebook (100.0%).

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?

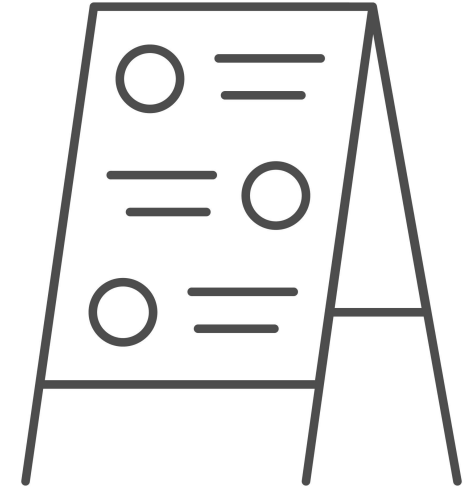
Case #4 - From a medical resident

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.

The screenshot shows the NYU Health Sciences Library NYU Data Catalog interface. The header includes the NYU Langone Health logo and navigation links: About, How To Use the Catalog, Frequently Asked Questions, and Login. A search bar is located on the right. The main content area displays the dataset 'American Community Survey' with the alternate title 'ACS' and UID '10026'. The 'Description' section explains that the ACS is an ongoing annual survey of the US population. The 'Access' section on the right includes 'Restrictions' (Free to All) and 'Instructions' on how to access the data. The left sidebar lists metadata: Publisher (United States - Bureau of the Census), Timeframe (2005 - Present), Geographic Coverage (United States), Local Expert (Ben Spoer), and Subject of Study (Homo sapiens). Two buttons at the bottom right provide access options: 'Access via Census Data Explorer' and 'Access via Census File Transfer'.

NYU Health Sciences Library | NYU Data Catalog

About How To Use the Catalog Frequently Asked Questions Login

Search

American Community Survey

Alternate Title(s): ACS
UID: 10026

Description
The American Community Survey (ACS) is an ongoing annual survey that publishes and collects data on the population of the United States. The survey is issued to approximately 3.5 million households per year. The ACS contains detailed questions assessing social, economic, demographic, and housing characteristics, including age, sex, race, family and relationships, income and benefits, health insurance, education, veteran status, disabilities, employment, commuting, and residency.

Researchers at NYU Langone Health can find guidance for the use and analysis of Census Bureau data on the Population Health Data Hub (listed under "Other Resources"), which is accessible only through the intranet portal with a valid Kerberos ID (KID).

Publisher
United States - Bureau of the Census

Timeframe
2005 - Present

Geographic Coverage
United States

Local Expert
Ben Spoer

Subject of Study
Homo sapiens

Access

Restrictions
Free to All

Instructions
All listed data products may be accessed at no cost. Summary files and public use microdata samples (PUMS) in CSV and SAS formats may be accessed through the Census Bureau. The ICPSR data repository hosts ACS data tables as well as restricted microdata files through the ResearchDataGov portal.

To apply for access to restricted microdata files, the following information is required: investigator CV or resume, names and contact information for all members of the research team, and a 1-3 page description of research goals, including any other data needed to complete the research.

Access via Census Data Explorer

Access via Census File Transfer

<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