Choosing a Repository for Scientific Data

Lisa Chinn, PhD, MLIS
Data Services Librarian
Leon S. McGoogan Health Sciences Library
March 28, 2024

University of Nebraska Medical Center℠

Objective

Help you evaluate data repositories, focusing on the NIH Data Management and Sharing Policy

What We Will Cover:

1) Underlying motivation
2) What is a Data Repository?
3) Two Types of Repositories
    1) Discipline-Specific
    2) Generalist
4) How to evaluate a Repository for your data

Underlying Motivation

Data Management and Sharing Plans for Federally Funded Research

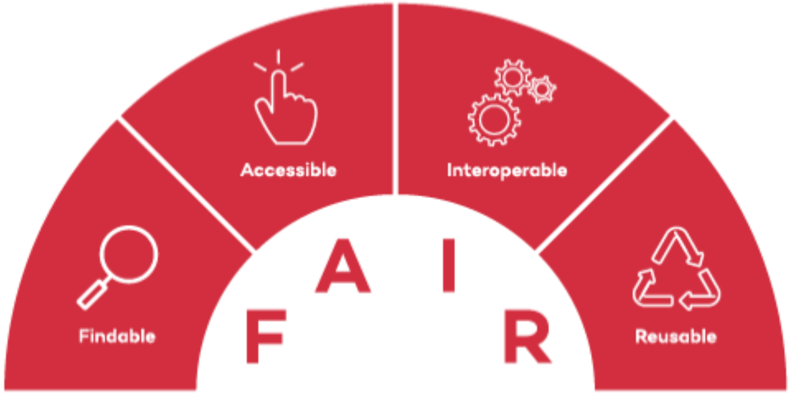
-Requires a description of how you plan to preserve and share your research data with others

-Preservation and sharing are key components of the new NIH DMSP

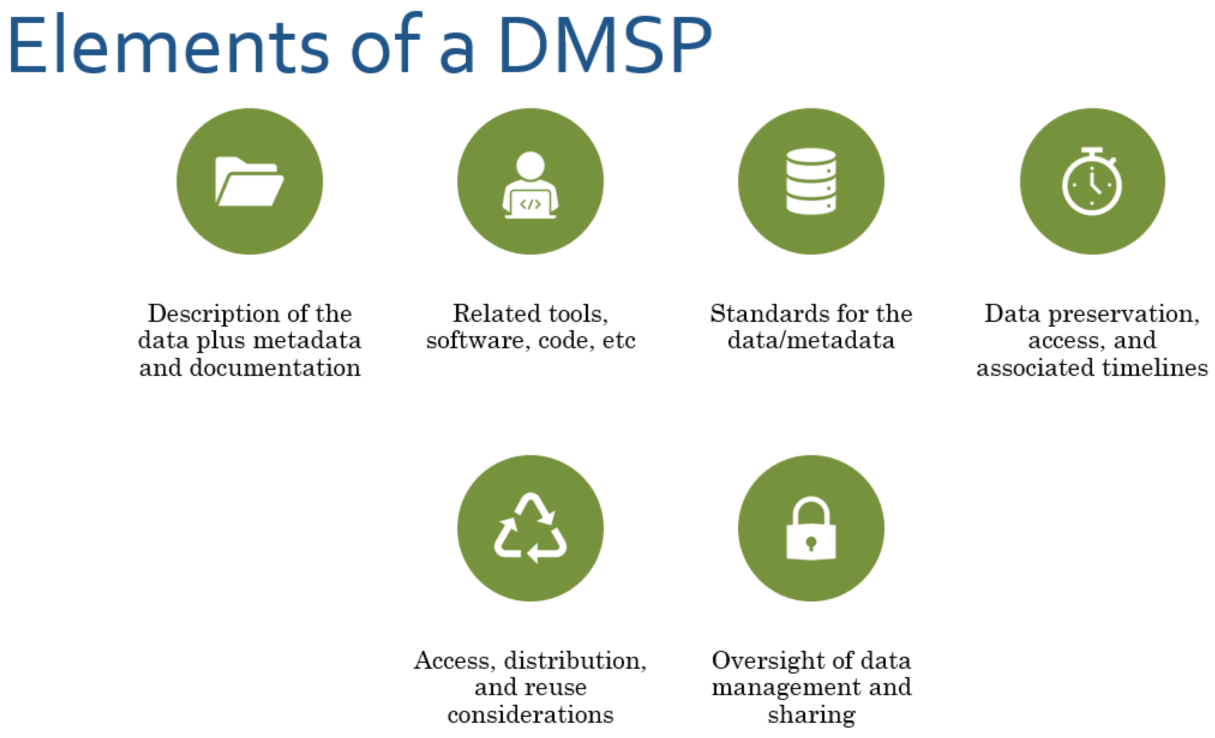
- Elements 4 and 5 of the NIH DMSP directly address preservation and sharing

Why Preserve & Share?

Preserving and sharing scientific data promotes FAIR data use:



6 Elements of the NIH DMSP



Elements of a DMSP

Description of the data plus metadata and documentation

Related tools, software, code, etc

Standards for the data/metadata

Data preservation, access, and associated timelines

Access, distribution, and reuse considerations

Oversight of data management and sharing

https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-014.html

NIH DMSP Element 4: Preservation

Data Preservation, Access, and Associated Timelines

4.1 Repository where scientific data and metadata will be archived

4.2 Describe how the scientific data will be findable and identifiable

4.3 When and how long the scientific data will be made available

NIH DMSP Element 5: Sharing

Access, Distribution, or Reuse Considerations

5.1 Factors affecting access, distribution, or reuse of scientific data

5.2 Controlled access to scientific data

5.3 Protection for privacy, rights, and confidentiality of human research participants

To Keep in Mind:

Some NIH Institute, Center, Office (ICO) policies and Funding Opportunity Announcements (FOAs) already have designated repositories for preserving and sharing data.

If an ICO/FOA has a designated respiratory, use the designated repository.

National Institutes of Health, *Supplementary Information to the NIH Policy for Data Management and Sharing: Selecting a Repository for Data Resulting from NIH-Supported Research*, 2020, https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-016.html.

To Keep in Mind:

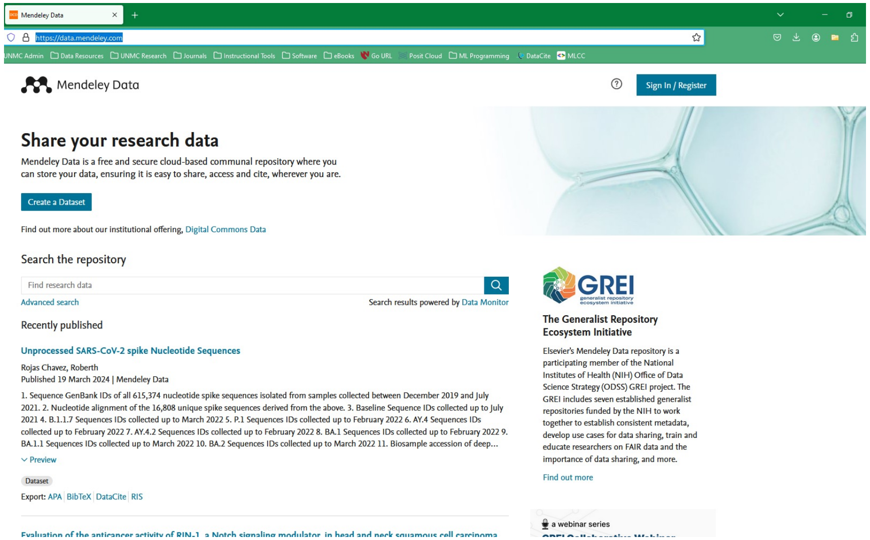
If dataset is small (up to 2 GB), then it may be included as supplementary material to articles submitted to PubMed Central.

National Institutes of Health, *Supplementary Information to the NIH Policy for Data Management and Sharing: Selecting a Repository for Data Resulting from NIH-Supported Research*, 2020, https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-016.html.

To Keep in Mind:

Publishers are requiring datasets to be uploaded in repositories.

For example, some Elsevier publications require supplementary data to be uploaded to Mendeley Data.



Share your research data

Mendeley Data is a free and secure cloud-based communal repository where you can store your data, ensuring it is easy to share, access and cite, wherever you are.

Find out more about our institutional offering, Digital Commons Data

Search the repository

Search results powered by Data Monitor

Recently published

Unprocessed SARS-CoV-2 spike Nucleotide Sequences

Rojas Chavez, Roberth
Published 19 March 2024 | Mendeley Data

1. Sequence GenBank IDs of all 615,374 nucleotide spike sequences isolated from samples collected between December 2019 and July 2021. 2. Nucleotide alignment of the 16,808 unique spike sequences derived from the above. 3. Baseline Sequence IDs collected up to July 2021 4. B.1.1.7 Sequences IDs collected up to March 2022 5. P.1 Sequences IDs collected up to February 2022 6. AY.4 Sequences IDs collected up to February 2022 7. AY.4.2 Sequences IDs collected up to February 2022 8. BA.1 Sequences IDs collected up to February 2022 9. BA.1.1 Sequences IDs collected up to March 2022 10. BA.2 Sequences IDs collected up to March 2022 11. Biosample accession of deep...

Evaluation of the anticancer activity of RIN-1, a Notch signaling modulator, in head and neck squamous cell carcinoma

The Generalist Repository Ecosystem Initiative

Elsevier's Mendeley Data repository is a participating member of the National Institutes of Health (NIH) Office of Data Science Strategy (ODSS) GREI project. The GREI includes seven established generalist repositories funded by the NIH to work together to establish consistent metadata, develop use cases for data sharing, train and educate researchers on FAIR data and the importance of data sharing, and more.

GREI Collaborative Webinar

What is a Data Repository?

What is a Data Repository?

A data repository is a large database infrastructure that collects, manages, and stores data sets for analysis and sharing.

Key Characteristics

The NSTC has guidelines for desirable characteristics structured in three major categories. To evaluate a data repository, evaluate based on:

1. Organizational Infrastructure
2. Digital Object Management
3. Technology

The National Science and Technology Council, *Desirable Characteristics of Data Repositories for Federally Funded Research,* 2022, DOI: https://doi.org/10.5479/10088/113528

Key Characteristics

1. Organizational Infrastructure:

- Free and Easy Access
- Clear Use Guidance
- Risk Management
- Retention Policy
- Long-Term Organization Sustainability

The National Science and Technology Council, *Desirable Characteristics of Data Repositories for Federally Funded Research,* 2022, DOI: https://doi.org/10.5479/10088/113528

Key Characteristics

2. Digital Object Management:

- Unique Persistent Identifiers (DOIs)
- Metadata
- Curation and Quality Assurance
- Broad and Measured Reuse
- Common Format
- Provenance

The National Science and Technology Council, *Desirable Characteristics of Data Repositories for Federally Funded Research,* 2022, DOI: https://doi.org/10.5479/10088/113528

Key Characteristics

3. Technology

- Authentication
- Long-term Technical Sustainability
- Security and Integrity

The National Science and Technology Council, *Desirable Characteristics of Data Repositories for Federally Funded Research,* 2022, DOI: https://doi.org/10.5479/10088/113528

Additional Considerations

Additional Considerations for Repositories Storing Human Data:

- Fidelity to Consent
- Security
- Limited Use Compliant
- Download Control
- Request Review
- Plan for Breach
- Accountability

The National Science and Technology Council, *Desirable Characteristics of Data Repositories for Federally Funded Research,* 2022, DOI: https://doi.org/10.5479/10088/113528

Two types of Repositories

Two types of Repositories

Discipline-specific repositories: provide options that generalist repositories do not: file previews, analysis and visualization tools, discipline specific metadata standards, larger file size support. NIH-supported repositories are discipline-specific repositories.

Generalist Repositories: store and preserve a wide variety of data types and research outputs and usually accept data regardless of the type, format, content, disciplinary focus, or research institution affiliation.

Discipline-Specific Repositories

Two major databases for discipline-specific repositories:

NIH-supported Scientific Data Repositories:

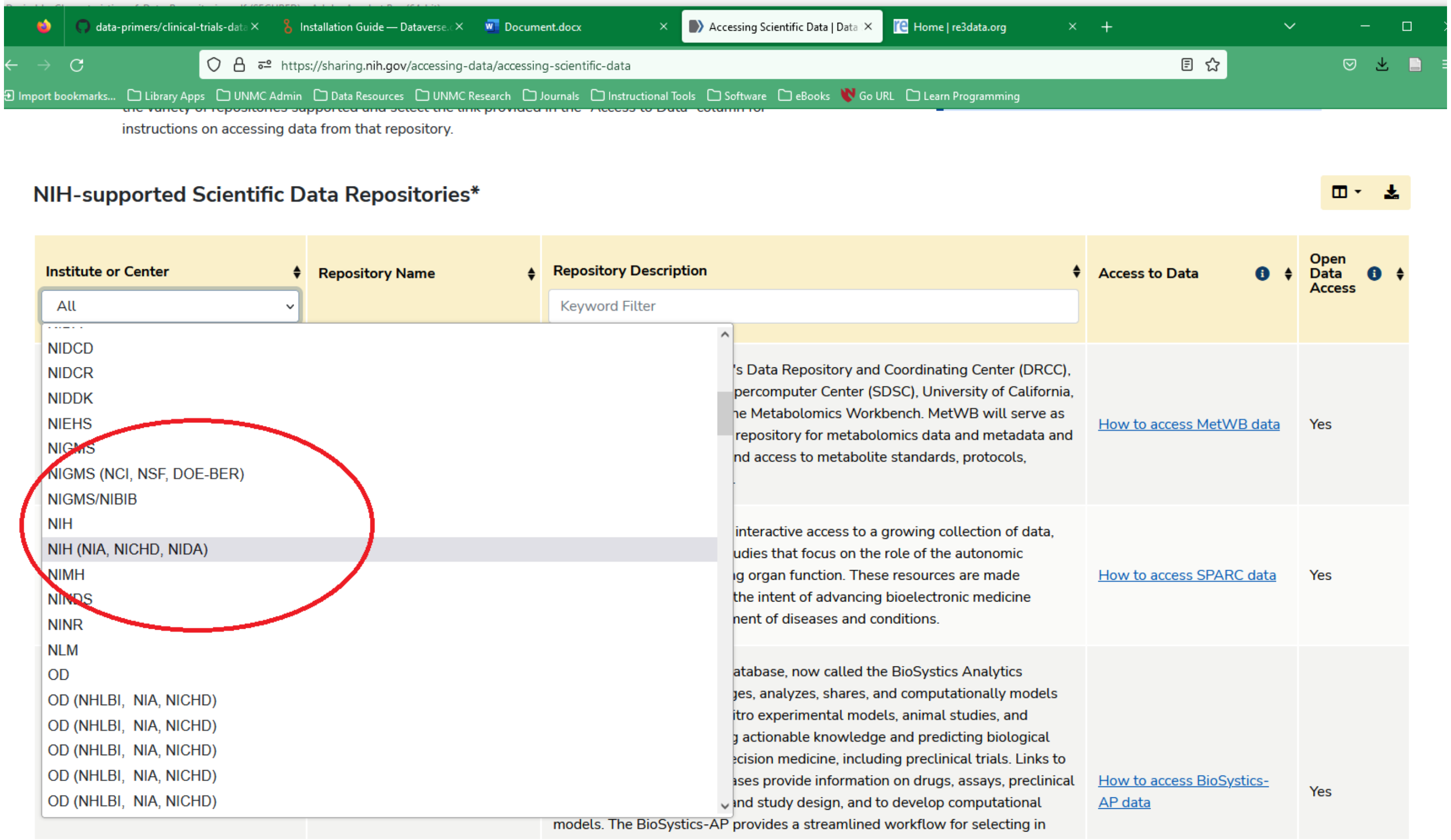
https://sharing.nih.gov/accessing-data/accessing-scientific-data

Registry of Research Data Repositories:

https://www.re3data.org/

NIH-Supported Repositories

https://sharing.nih.gov/accessing-data/accessing-scientific-data



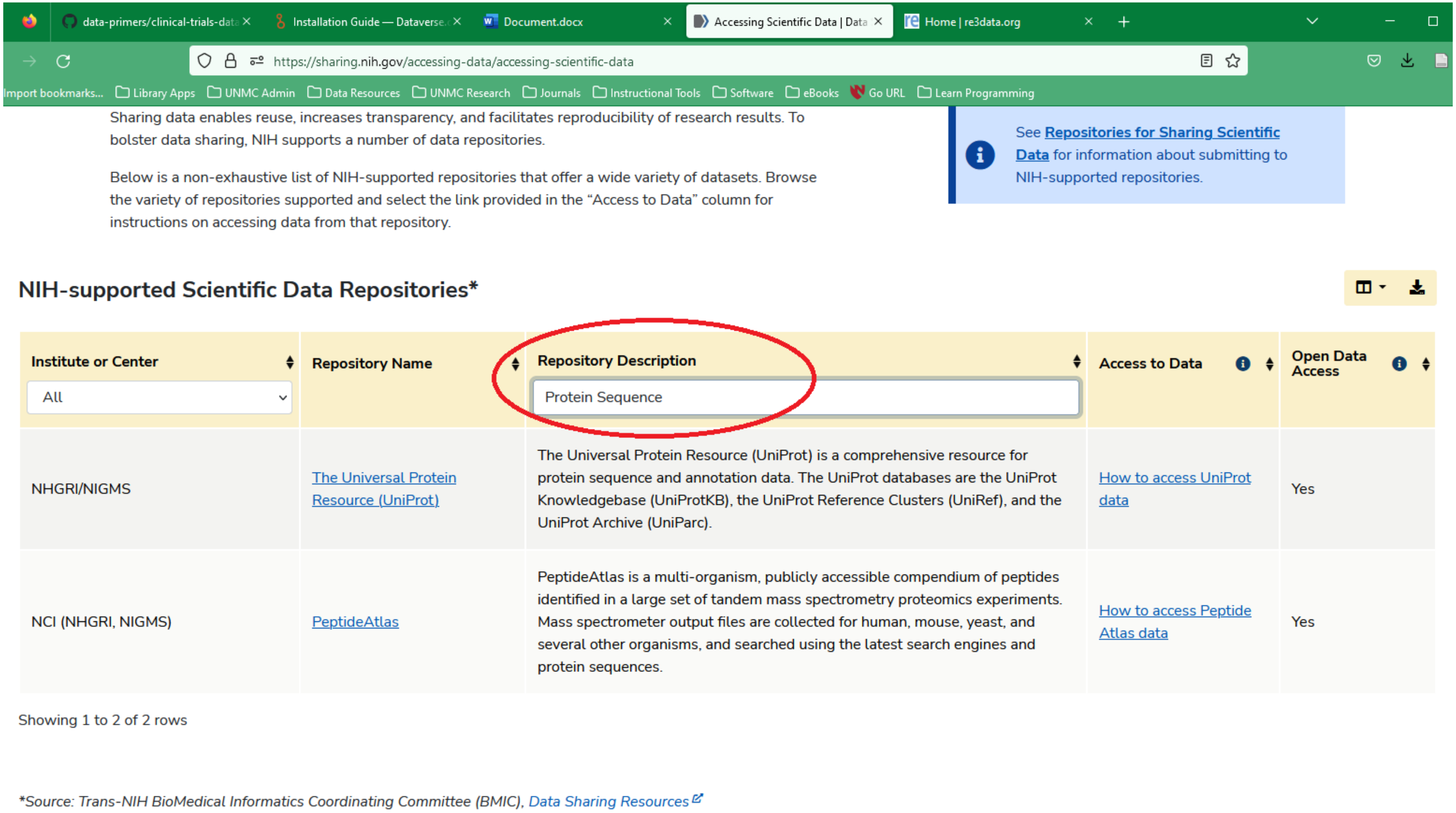
instructions on accessing data from that repository.

NIH-supported Scientific Data Repositories*

Institute or Center	Repository Name	Repository Description	Access to Data	Open Data Access
All		Keyword Filter		
		's Data Repository and Coordinating Center (DRCC), percomputer Center (SDSC), University of California, he Metabolomics Workbench. MetWB will serve as repository for metabolomics data and metadata and nd access to metabolite standards, protocols,	How to access MetWB data	Yes
		interactive access to a growing collection of data, udies that focus on the role of the autonomic g organ function. These resources are made the intent of advancing bioelectronic medicine nent of diseases and conditions.	How to access SPARC data	Yes
		atabase, now called the BioSystics Analytics ges, analyzes, shares, and computationally models itro experimental models, animal studies, and g actionable knowledge and predicting biological ecision medicine, including preclinical trials. Links to ases provide information on drugs, assays, preclinical nd study design, and to develop computational models. The BioSystics-AP provides a streamlined workflow for selecting in	How to access BioSystics-AP data	Yes

NIDCD
NIDCR
NIDDK
NIEHS
NIGMS
NIGMS (NCI, NSF, DOE-BER)
NIGMS/NIBIB
NIH
NIH (NIA, NICHD, NIDA)
NIMH
NINDS
NINR
NLM
OD
OD (NHLBI, NIA, NICHD)
OD (NHLBI, NIA, NICHD)
OD (NHLBI, NIA, NICHD)
OD (NHLBI, NIA, NICHD)
OD (NHLBI, NIA, NICHD)

NIH-Supported Repositories



Sharing data enables reuse, increases transparency, and facilitates reproducibility of research results. To bolster data sharing, NIH supports a number of data repositories.

Below is a non-exhaustive list of NIH-supported repositories that offer a wide variety of datasets. Browse the variety of repositories supported and select the link provided in the "Access to Data" column for instructions on accessing data from that repository.

See **Repositories for Sharing Scientific Data** for information about submitting to NIH-supported repositories.

NIH-supported Scientific Data Repositories*

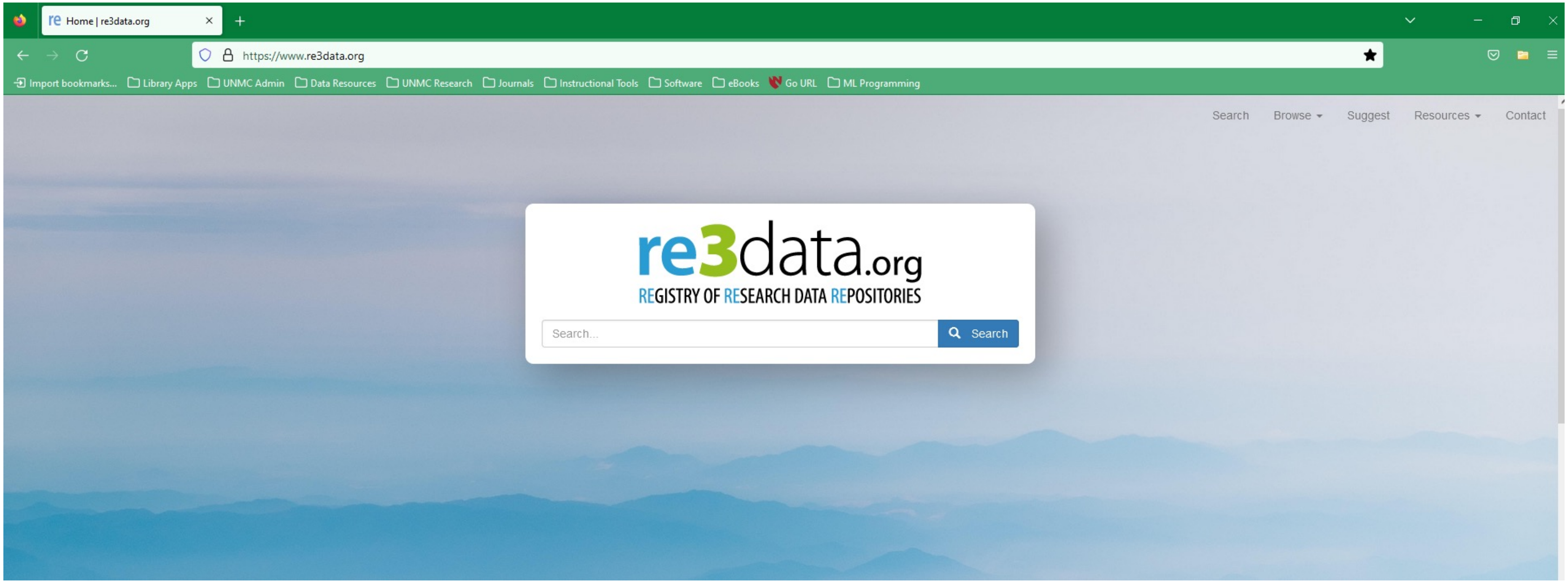
Institute or Center All	Repository Name	Repository Description Protein Sequence	Access to Data	Open Data Access
NHGRI/NIGMS	The Universal Protein Resource (UniProt)	The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc).	How to access UniProt data	Yes
NCI (NHGRI, NIGMS)	PeptideAtlas	PeptideAtlas is a multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments. Mass spectrometer output files are collected for human, mouse, yeast, and several other organisms, and searched using the latest search engines and protein sequences.	How to access Peptide Atlas data	Yes

Showing 1 to 2 of 2 rows

Source: Trans-NIH BioMedical Informatics Coordinating Committee (BMIC), Data Sharing Resources

Registry of Research Data Repositories

www.re3data.org



Other Discipline-Specific Resources

Wiki list of data repositories hosted by Simmons University:

https://oad.simmons.edu/oadwiki/Data_repositories

Data repository guidance from *Nature's Scientific Data* (journal dedicated to publishing solely datasets):

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

Generalist Repositories

Supported by UNMC:

DataVerse

Dryad

figshare

Zenodo

Generalist Repositories

My recommendations:

DataVerse

The Dataverse® Project

Zenodo

zenodo

Evaluating Repositories for Scientific Data

Choosing a Repository

Assigns DOIs

Choosing a Repository

Long-term sustainability

Choosing a Repository

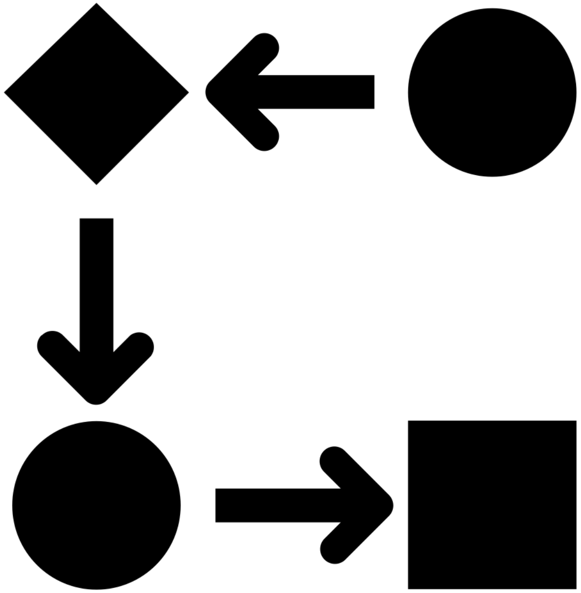
Curation and quality assurance services

Choosing a Repository

Free and easy access

Choosing a Repository

Allows broad and measured reuse



Choosing a Repository

Provides clear use guidance

Choosing a Repository

Security and integrity

Choosing a Repository

Maintains confidentiality

Choosing a Repository

Supports common file formats

Choosing a Repository

Records data provenance
(e.g., tracks data versions)

Choosing a Repository

Documented retention policies

Additional Considerations: Human Subjects Research

- Fidelity to consent
- Restricted use compliance
- Privacy
- Plan for breach
- Download control
- Procedures for violations
- Request review

Modified from: National Institutes of Health, *Supplementary Information to the NIH Policy for Data Management and Sharing: Selecting a Repository for Data Resulting from NIH-Supported Research*, 2020, https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-016.html.

Clinical Trials Repository:

Vivli:

Vivli

CENTER FOR GLOBAL CLINICAL RESEARCH DATA

Reflection:

What data do you collect, store, and use for analysis?

Given the options discussed, can you find at least one repository that might work for your data?

Questions?

Connect with me!

Lisa Chinn, PhD, MLIS, Research Data Services, McGoogan Health Sciences Library
lchinn@unmc.edu

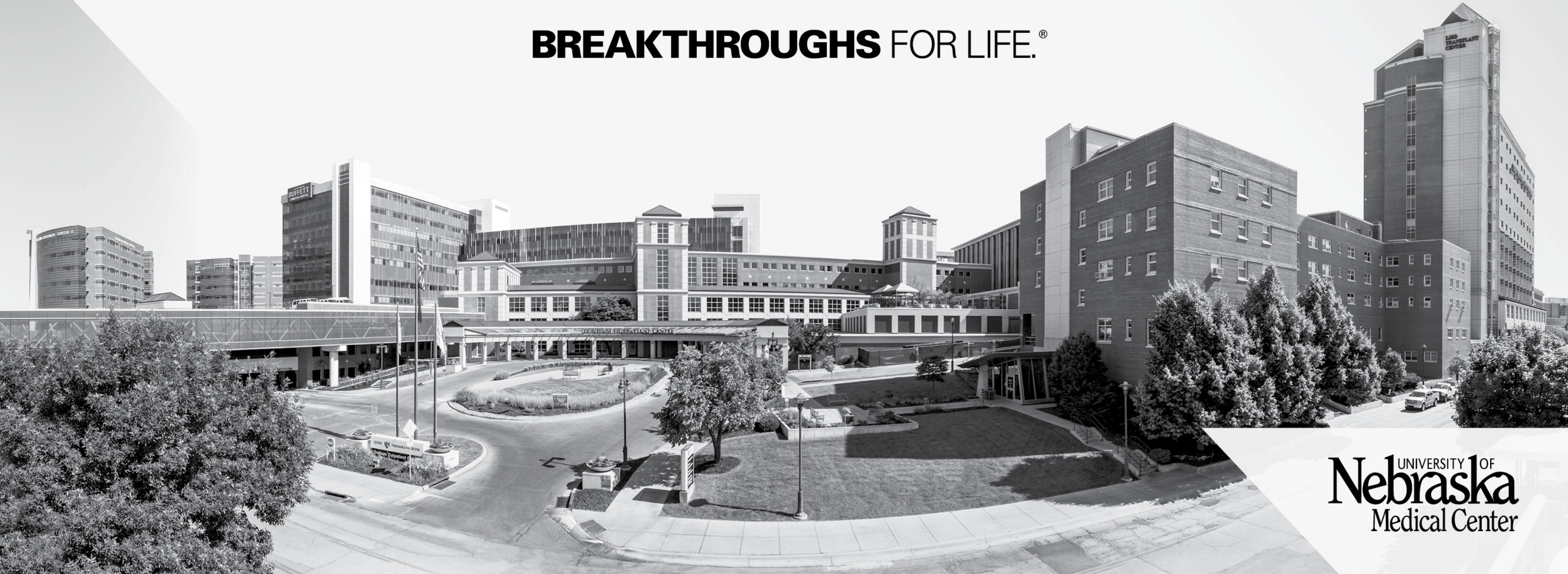
Research Data Services email
researchdata@unmc.edu

Book an Appointment with me:
https://go.unmc.edu/veb3

Upcoming Events:
https://www.unmc.edu/library/services/instruction.html

University of Nebraska
Medical Center℠

BREAKTHROUGHS FOR LIFE.®



UNIVERSITY OF
Nebraska
Medical Center