National Institutes of Health
Office of Data Science Strategy

Partnerships Contributing to Data Management and Sharing Policy Implementation Through FAIR Data Sharing

Dr. Susan K. Gregurick
Associate Director of Data Science

May 20, 2022

Office of Data Science Strategy

Vision

- The complexity and volume of basic, translational, and clinical research data generated by NIH-supported investigators continues to rapidly increase. To take full advantage of these data, NIH must integrate the collection, storage, analysis, use, and sharing of these data according to FAIR practices and foster a talented and diverse data science workforce.

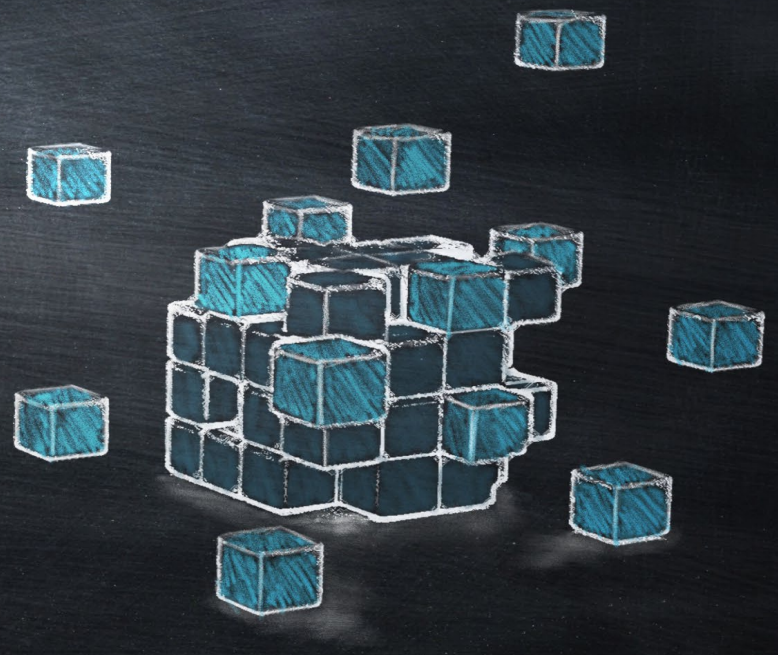
Mission

- To catalyze new capabilities in biomedical data science by providing trans-NIH leadership and coordination for modernization of the NIH data resource ecosystem, development of a diverse and talented data science workforce, and building strategic partnerships to develop and disseminate advanced technologies and methods.


We recognize the challenges ahead

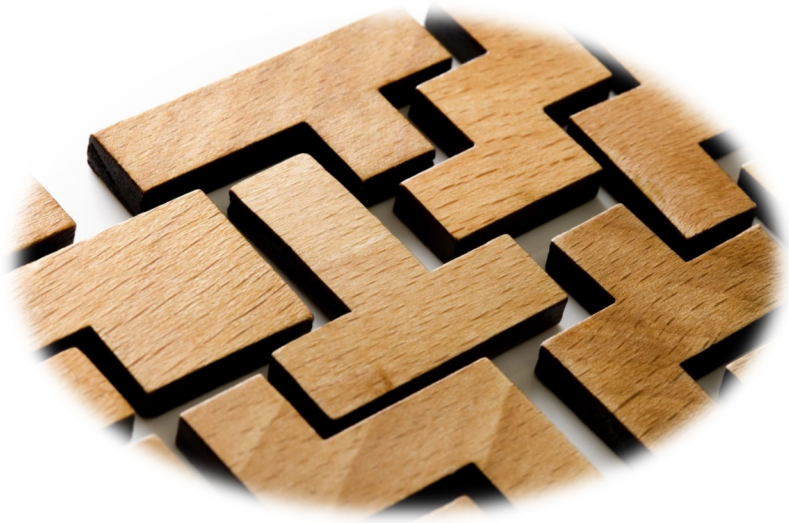
Sharing Data…

- FAIR & AI Ready Data
- Support for Data Resources
- Creating a Data Ecosystem
- Reaching a Broad Community



2021 Making Data AI-Ready NOSI (NOT-OD-21-094)

Artificial intelligence and machine learning (AI/ML) have the potential to significantly advance biomedical research. NIH makes a wealth of biomedical data available and reusable to research communities however, not all of these data are able to be used efficiently and effectively by AI/ML applications.



AI/ML-readiness should be guided by a concern for human and clinical impact and requires attention to ethical, legal, and social implications of AI/ML including, but not limited to:

- biases in datasets, algorithms, and applications
- issues related to identifiability and privacy
- impacts on disadvantaged or marginalized groups and health disparities
- unintended, adverse social, individual, and community consequences of research and development

Collaborations to Make Data FAIR and AI/ML Ready

ODSS supported collaboration, bringing together expertise in biomedicine, data management, and artificial intelligence and machine learning (AI/ML) to make NIH-supported data AI-ready for AI/ML analytics.

36 Awards:

- 2 IDeA States
- 12 address specific ethics challenges in AI
- 2 early-stage/new investigators
- 13 female investigators

Most common biomedical focus areas: Alzheimer's disease, cardiovascular disease, and aging

Most common data types: imaging, EHRs, -omics, speech

NHGRI | NIA | NIBIB | NIDA | NIDCD | NIDCR | NIEHS | NIGMS | NIMH | NINDS

Improving the AI/ML-Readiness of Data

Rutvik Desai, University of South Carolina

3-R01-DC017162-02S1

Goal: Study of how concepts are represented and processed in the brain.

Research: Improving neuroimaging datasets — using sematic language techniques to create machine readable metadata format

John Gilmore, University of North Carolina Chapel Hill

3-R01-MH123747-01A1S1

Goal: Study imaging and image analysis methodologies to identify children at high risk for schizophrenia.

Research: Bridge missing timepoint imaging data (data imputation) using Out-of-Distribution Detection (ML) from existing data at different timepoints.

Carl Kesselman, University of Southern California

3-U01-DE028729-02

Goal: The FaceBase consortium is a distributed network of researchers investigating craniofacial development and dysmorphology. FaceBase Hub infrastructure stores, represents, and serves relevant data to the research community

Research: Streamline curation using ML approaches to improve metadata descriptive elements while maintaining required restrictions on data handling.

The NIH Data Sharing Landscape

NIH strongly encourages
open access Data Sharing Repositories
as a first choice.

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

Datasets up to **2 gigabytes**

PubMed Central

Stores publication-related supplemental materials and datasets directly associated publications.

Datasets up to **20 gigabytes**

Generalist Repositories

Datasets associated with publications or otherwise and links to PubMed.

High priority datasets **petabytes**

Cloud Partners
(STRIDES Program)

Store and manage large scale, high priority NIH datasets.

BMIC Repository lists

Trans-NIH
BioMedical
Informatics
Coordinating Committee
(BMIC)

To help researchers locate an appropriate resources for sharing their data, as well as to promote awareness of resources where datasets can be located for reuse, BMIC maintains lists of several types of data sharing resources:

- **Open NIH-supported domain-specific repositories that house data of a specific type or related to a specific discipline; Identified by NIH ICOs as key repositories**
- Other NIH-supported domain-specific resources, including repositories and knowledgebases, that have limitations on submitting and/or accessing data; and
- Generalist repositories that house data regardless of type, format, content, or subject matter.

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

PubMed Central Article Datasets are Now Available on the Cloud

To enhance machine access to biomedical literature and drive impactful analyses and reuse, PubMed Central (PMC) Article Datasets are available on Amazon Web Services (AWS) as part of AWS's Open Data Sponsorship Program (ODP).

These datasets collectively span 4 million of PMC's 7 million (total) full-text scientific articles.

https://registry.opendata.aws/

Support for NIH data repositories

NIH supports a variety of data repositories and knowledgebases (with data repository functionality) of **differing sizes** and **complexity** and at **different levels of maturity**

- Each has the **potential** to bring **value** to a given research area, but tend to be at **different stages** of maturity demonstrating that they have the appropriate practices in place to reliably manage the data they ingest and make available

- **Spectrum** of **ability** and **readiness** to adhere to the characteristics that are desirable for a data repository that are aligned with **FAIR** (**F**indable, **A**ccessible, **I**nteroperable, and **R**eusable) and **TRUST** (**T**ransparency, **R**esponsibility, **U**ser focus, **S**ustainability, and **T**echnology) principles

- **Developing metrics** for evaluating the **usage**, **utility**, and **impact** of a given repository is **evolving** and likely a function of several aspects

Positioning Repositories for Data Sharing

Support for existing data repositories to align with FAIR and TRUST principles and evaluate usage, utility, and impact

ALLIANCE of GENOME RESOURCES

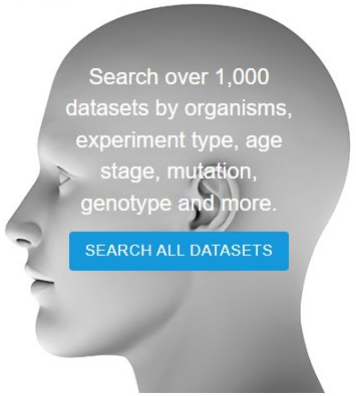
Mouse Phenome Database at The Jackson Laboratory

UniProt

DBAASP v3.0

Database of antimicrobial activity and structure of peptides

Explore our repository:



IMMPORT

BIOINFORMATICS FOR THE FUTURE OF IMMUNOLOGY

WORLDWIDE PDB

PROTEIN DATA BANK

17 Awards in 2021:

- 2 IDeA States
- 7 Female PI's
- 5 intramural

- 8 addressing FAIR and TRUST
- 6 addressing FAIR, TRUST, and Metrics
- 2 addressing FAIR
- 1 addressing TRUST

Biomedical focus areas: traumatic brain injuries, obesity nutrition, mental health, immune response

Data types: imaging, behavioral measures, clinical, EHRs, -omics, speech and language

Findable Accessible Interoperable Reusable

TRUST

Transparency Responsibility User focus Sustainability Technology

NHGRI | NIA | NIAID | NICHD | NIDA | NIDCD | NIDCR | NIDDK | NIEHS | NIMH | NINDS

Optimized Funding for NIH Data Repositories and Knowledgebases
PAR Funding Opportunities

- Data resources are important research tools
- Historically funded through research grants
- Funding mechanism should be optimal for type of resource
- **End goal:** researcher confident in data and information integrity

- **Solution:** New Funding Announcement for data repositories and knowledgebases
- Resource plan requirement

Scientific Impact	**Community Engagement**
Quality of Data and Services and Efficiency of Operations	**Governance**

PAR-20-089 and PAR-20-097

Data Repository (DR) & Knowledgebase (KB) Program

An NIH program to support investigator-initiated, sustainable data resource development driven by critical research needs

- Fill a scientific need or gap
- Encourage adoption of good data management practices
- Engage the research community to contribute and use data
- Govern data life-cycle and preservation

In 2020-2021: 29 applications reviewed & 7 awarded

BioPortal

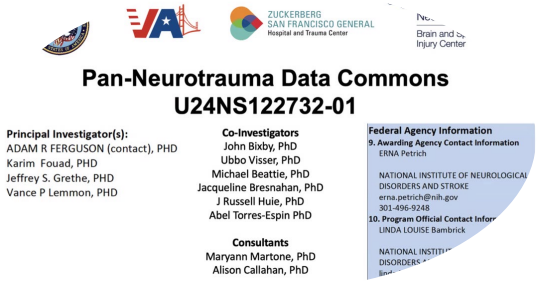
UniProt

ICTV

ctd

RED FLY

SHARC
SOUTHERN HIV & ALCOHOL RESEARCH CONSORTIUM



NEW: The Generalist Repository Ecosystem Initiative

Solicit applications from generalist repositories working together to:

- Implement consistent capabilities (NOT-OD-21-016)
- Create better access to & discovery of NIH funded data
- Conduct outreach & train on FAIR data practices
- Engage the research community

Expected Outcomes

- Make data sharing easier
- Improve discoverability
- Increase reproducibility of research
- Encourage secondary use of data

The Dataverse Project · DRYAD · figshare · MENDELEY DATA · OSF · Vivli CENTER FOR GLOBAL CLINICAL RESEARCH DATA

GREI Objectives

generalist repository ecosystem initiative

- Align with Desirable Characteristics for Data Repositories
- Implement browse & search for NIH funded data
- Develop consistent metadata models
- Conduct limited Q/AC of the NIH funded data
- Enable connectivity of digital objects
- Use case support including (x-repository use cases)
- Implement open metrics
- Develop educational materials
- Conduct broad outreach (workshops)
- Commit to "Co-opetition"

Openly share software & work products developed under the award

GREI generalist repository ecosystem initiative Awardees

Generalist Repository	Website
The Dataverse Project	https://dataverse.org
DRYAD	https://datadryad.org
figshare	https://figshare.com
MENDELEY DATA	https://data.mendeley.com
OSF (Open Science Framework)	https://osf.io
Vivli CENTER FOR GLOBAL CLINICAL RESEARCH DATA	https://vivli.org

DATAWorks! FASEB Prize

$500,000 Total Available

Up to 12 monetary prizes recognizing team achievement in data sharing or reuse practices

Entries Open: May 11, 2022
Entries Close: July 19, 2022

Highlighting the Power of Data Sharing and Reuse in the Biological & Biomedical Sciences

Learn More & Enter
www.herox.com/dataworks

DataWorks! Prize is a partnership between FASEB and NIH

NIH National Institutes of Health
Office of Data Science Strategy

DATAWorks! FASEB
Advancing Data Sharing and Reuse

(New) Data Curation Network – Event Series (ODSS, NLM)

https://datacurationnetwork.org/

Event 1: Kick-off Webinar for Researchers (date – Apr 6 2022)

Role of Librarians at Universities, Service offerings, Curation Resources

Event 2: Virtual (half-day) Workshop for Program Officers (~Jul 2022)

DMPs – Review/Evaluation, & Metrics of Review

Event 3: Virtual (half-day) Workshop for Curators (~ Oct 2022)

Train librarians, repository owners, others on curation of data type/format

Event 4: In-person Workshop for Curators (2 day) (~ Feb 2023)

Train librarians/curators on biomedical data types/formats – BYOD workshop

Check files
Understand or try to
Request missing information
Augment the submission
Transform the format
Evaluate for FAIRness
Document throughout

Make **Self-Paced Training Content** Available to Researchers, POs, Repository Owners & Other Curators

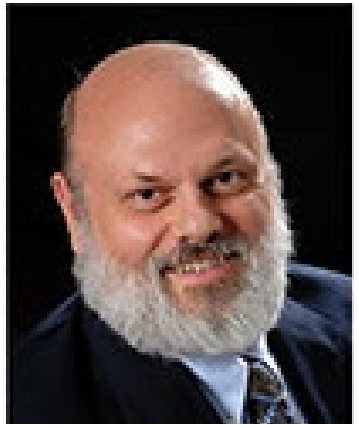
ODSS Data Sharing and Reuse Seminar Series

The Office of Data Science Strategy (ODSS) hosts a seminar series to highlight exemplars of data sharing and reuse on the second **Friday of each month at noon ET**. The monthly series highlights researchers who have taken existing data and found clever ways to reuse the data or generate new findings. *A different NIH institute or center (IC) will also share its data science activities each month.*

Past Speakers:

Karen E. Adolph, Ph.D.
Databrary: Secure and Ethical Sharing of Research Video as Data and Documentation

Purvesh Khatri, Ph.D.
Adventures of a Data Parasite: Accelerating Clinical Translation Using Heterogeneity in Public Data



Alexander Ropelewski
The Brain Image Library:
A Resource for Sharing Microscopy Data

https://datascience.nih.gov/nih-data-sharing-and-reuse-seminar-series

National Institutes of Health
Office of Data Science Strategy

Thank You!