Towards Better Data Management & Sharing at the NIH

Federal Data Partnership (FDP) Meeting
May 12 2022

Ishwar Chandramouliswaran
Program Director, Office of Data Science Strategy, NIH
ishwarc@nih.gov
Office of Data Science Strategy

www.datascience.nih.gov

A modernized, integrated, FAIR biomedical data ecosystem
Office of Data Science Strategy

The NIH **Office of Data Science Strategy (ODSS)**, in the Office of the Director:

- Provides **leadership and coordination** on the strategic plan for data science.
- Develops and implement NIH’s vision for a **modernized and integrated** biomedical data ecosystem.
- Enhances a **diverse and talented** data science workforce.
- **Builds strategic partnerships** to develop and disseminate advanced technologies and methods.
Strategic Plan for Data Science: Goals and Objectives

1. Data Infrastructure
   - Optimize data storage and security
   - Connect NIH data systems

2. Modernized Data Ecosystem
   - Modernize data repository ecosystems
   - Support storage and sharing of individual datasets
   - Better integrate clinical and observational data into biomedical data science

3. Data Management, Analytics, and Tools
   - Support useful, generalizable, and accessible tools
   - Broaden utility of, and access to, specialized tools
   - Improve discovery and cataloging resources

4. Workforce Development
   - Enhance the NIH data science workforce
   - Expand the national research workforce
   - Engage a broader community

5. Stewardship and Sustainability
   - Develop policies for a FAIR data ecosystem
   - Enhance stewardship

https://datascience.nih.gov/
Catalyzing Data Science Across NIH

More than 190 NIH staff from 23 ICOs contributed to these activities
NIH requires researchers to prospectively plan for how scientific data will be preserved and shared through submission of a Data Management and Sharing Plan.

Submission of a Data Management and Sharing Plan outlining how scientific data and any accompanying metadata will be managed and shared, taking into account any potential restrictions or limitations.

Plan is part of the budget Justification section of the application for extramural awards and as part of the technical evaluation for contracts.

The DMS Policy applies to all research, funded or conducted in whole or in part by NIH, that results in the generation of scientific data. This includes research funded or conducted by extramural grants, contracts, Intramural Research Projects, or other funding agreements regardless of NIH funding level or funding mechanism. The DMS Policy does not apply to research and other activities that do not generate scientific data, including training, infrastructure development, and non-research activities.
FAIR Principles

Findable
Accessible
Interoperable
Reusable

Image Courtesy: Dr. Patti Brennan
NIH supports a variety of data repositories and knowledgebases 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** (Findable, Accessible, Interoperable, and Reusable) and **TRUST** (Transparency, Responsibility, User focus, Sustainability, and Technology) principles

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

Data resources are key enablers of modern biomedical research

Preparing the NIH Data Ecosystem for more effective sharing and the new NIH Data Management and Sharing Policy (NOT-OD-21-013)

1. Data Infrastructure
   - Optimize data storage and security
   - Connect NIH data systems

2. Modernized Data Ecosystem
   - Modernize data repository ecosystems
   - Support storage and sharing of individual datasets
   - Better integrate clinical and observational data into biomedical data science

Active Funding Opportunity Announcements
- Biomedical Data Repository FOA (PAR20-089)
- Knowledgebase FOA (PAR20-097)

One-year Administrative Supplements
- Implement desirable characteristics for data repositories
- Improve the AI/ML-Readiness of NIH-Supported Data
- Enhancement of Software Tools for Open Science

Incorporating Generalist Repositories into the NIH Data Ecosystem

Icons courtesy Flaticon.com
Supplements to Support Existing Repositories

Better enable data discoverability, interoperability, and reuse

Goals

• Implement the relevant portions of the Desirable Characteristics for Repositories (i.e., NOT-OD-21-016) with the aim to strengthen adoption of the FAIR principles

• Strengthen the adoption of the TRUST principles, which may include a structured plan to pursue certifications

• Adopt, enhance, or contribute to community-based metrics standards or best practices of metrics to evaluate the usage, utility, and impact of the data resource throughout its lifecycle

NOT-OD-22-069
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

Why is the program important?

- Reduce or eliminate siloes; lower barriers for data sharing
- Allow discovery and use of data
- Disentangle data resources from research projects
- Optimize efficiency of operations and costs

In 2020-2021: 29 applications reviewed & 7 awarded
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
Use biomedical and behavioral research grand challenges to generate **flagship data sets**

Emphasize **ethical** best practices

Prepare **AI/ML-friendly data**

Promote **diverse teams**

https://commonfund.nih.gov/bridge2ai
STRIDES Initiative (The Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability) provides:

- State-of-the-art data storage and computational capabilities
- Training and education for researchers
- Innovative technologies such as artificial intelligence and machine learning
- Professional engineering and technical support
Helping advance biomedical research by delivering access to industry-leading cloud providers.

The STRIDES Initiative aims to help NIH and its institutions accelerate biomedical research by reducing barriers in utilizing commercial cloud services. This initiative aims to harness the power of the cloud to accelerate biomedical discovery. NIH and NIH-funded researchers can take advantage of STRIDES benefits.

- Discounts on partner services
- Professional services consultations
- Access to training
- Potential collaborative engagements

Enroll Now

>163
Petabytes of Data

201M
Compute Hours

>693
NIH & NIH-funded Research Programs/Projects

>4081
People Trained

$28M
Cost Savings

https://datascience.nih.gov/strides
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/
The NIH Cloud Platform Interoperability (NCPI) effort aims to establish and implement guidelines and technical standards to empower end-user analyses across participating NIH cloud platforms.
Researcher Workflows After RAS

Internal NIH and External researchers sign in with preferred credentials

Secure access to NIH-funded datasets and tools

Auditing & Logging of Events

Identity Providers

Key Features

- **More efficient** - Authorizations from NIH dbGaP Data Access Committee (DAC) decisions are centralized and provisioned only upon login
- **Simple** – Log in or link accounts from multiple identity providers for a single sign on experience across systems
- **Secure** – Use multi-factor authentication (MFA) for data repositories that require a higher level of access security

https://datascience.nih.gov/researcher-auth-service-initiative
Smart and Connected Health (SCH)

Accelerate innovations in computer and information science and engineering to support the transformation of health and medicine

These supplements will

• support robustness, sustainability, and scalability of existing biomedical research software tools and workflows

• invest in research software tools with recognized value in a scientific community to enhance their impact by leveraging best practices in software development and advances in cloud computing

• support collaborations between biomedical scientists and software engineers to enhance the design, implementation, and “cloud-readiness” of research software
Best Practices for Sharing Research Software

Provides information on:

• How to make software open and citable
• How to allow developed (research) software for use in medical practice or clinical settings
• Software vulnerabilities
• Metadata
• Benchmarking, provenance etc

https://datascience.nih.gov/tools-and-analytics/best-practices-for-sharing-research-software-faq
Graduate Data Science Summer Programs

- Led by Office of Intramural Training and Education as part of Summer Internship Program
- 10-week summer program
- All interns placed in intramural research labs
  
  https://www.training.nih.gov/data_science_summer

Coding it Forward is a non-profit focused on developing the next generation of technology leaders through the summer Civic Digital Fellowship. These fellows spend 10 weeks at NIH channeling their computational expertise toward hands-on experience with biomedical data-related challenges.

https://www.codingitforward.com/summer-fellowships
DATA Scholars Applications Now Open!

- DATA Scholars will optimize and accelerate data science in biomedicine to improve human health and well-being.

- Open application period now through **May 27th, 2022**

- Now accepting DATA Scholar applications for 9 IC Projects
  - AI-Ready Data Ecosystems for Pandemic Preparedness
  - Automating Review and Update of Consumer Health Information
  - Building an Interoperable Autoimmunity, Inflammation, and COVID-19 Data Ecosystem
  - Creating Multi-modal Cancer Data Integration Solutions from Cross-atlas Datasets
  - Democratization of NIDDK Knowledgebases
  - Enhancing Interoperability of Multi-modality Medical Image and Multi-format Clinical Data Repositories for AI/ML Algorithm Development for Clinical Applications
  - Exploration of Wearable Device Data in Predicting Clinical Outcomes
  - Sharing Language Models for Biological Sequences in Biomedical Repositories
  - Vision for Improving Interoperability for Eye Heath Data

Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD)

Goals

• to enhance the participation and representation of researchers and communities currently underrepresented in the development of artificial intelligence and machine learning (AI/ML) models

• to address health disparities and inequities using AI/ML

• to improve the capabilities of this emerging technology, beginning with the use of electronic health record (EHR) and extending to other diverse data

https://aim-ahead.net/
https://datascience.nih.gov/artificial-intelligence/aim-ahead
Data Curation Network – Event Series (ODSS, NLM)


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

Make **Self-Paced Training Content** Available to Researchers, POs, Repository Owners & Other Curators
Data Sharing and Reuse Seminar

Monthly series on the second Friday at 12 noon EST and highlights:

- Data Science and Programs activities at the NIH Institutes and Centers
- Data reuse by NIH funded researchers to answer novel research questions
- Illustrate value of data sharing

May 13, 2022, 12 p.m. ET
Dr. Jayashree Kalpathy-Cramer
Associate Professor of Radiology
Harvard Medical School
Sponsored by the National Eye Institute

Learn more and watch recordings:
$500,000 Prize Purse
Up to 12 monetary prizes recognizing team achievement in data sharing or reuse practices

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

DataWorks! Prize is a partnership between FASEB and NIH

Learn More & Enter www.herox.com/dataworks
NIH strongly encourages open access data sharing repositories as a first choice


**Scaled implementation options for sharing datasets**

<table>
<thead>
<tr>
<th>Datasets up to 2 gigabytes</th>
<th>Datasets up to 20*gigabytes</th>
<th>High priority datasets petabytes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PubMed Central</strong></td>
<td><strong>Generalist Repositories</strong></td>
<td><strong>Cloud Partners</strong></td>
</tr>
<tr>
<td>• Stores publication-related supplemental materials and datasets directly associated publications.</td>
<td>• Datasets associated with publications or otherwise and links to PubMed.</td>
<td>• Store and manage large scale, high priority NIH datasets.</td>
</tr>
</tbody>
</table>

Cloud Partners (STRIDES Program)

- Store and manage large scale, high priority NIH datasets.
EXPECTED OUTCOMES

Include but are not limited to:

• Make it easier for researchers to share data
• Improve discoverability of NIH-funded data across resources
• Make data associated with publications readily available (reproducibility)
• Avoid duplication of the data
• Encourage NIH-funded researchers to be contributors and consumers
Office of Data Science Strategy

www.datascience.nih.gov

A modernized, integrated, FAIR biomedical data ecosystem