How the NLM creates Trust in Data
Patricia Flatley Brennan, RN, PhD
NLM Director
Overview

- NLM Strategic Plan
- NIH Strategic Plan for Data Science
- How NLM is preparing for a future of data-driven discovery
- NIH Training and Policy Options
NLM Strategic Plan
Transforming Data into Knowledge

Accelerate discovery and advance health through data-driven research

Reach more people in more ways through enhanced dissemination and engagement

Build a workforce for data-driven research and health
Accelerate discovery and advance health through data-driven research
Fostering a ecosphere of discovery

digital research objects
Data Science & Open Science
Reach more people in more ways through enhanced dissemination and engagement
Enhance information delivery
New users, New ways
Biomedical & health information access methods
& information dissemination strategies
The Human Enterprise:
Reaching People Wherever
Build a workforce for data-driven research and health
Expand & enhance research training for biomedical informatics & data science
Training across society
Foster distinctiveness of NLM as a reliable, trustable source of health information & biomedical data
The National Institutes of Health
27 Institutes and Centers – $37B
Requested by Congress: In its fiscal year 2017 report, the Committee directed NIH to work with its external partners and stakeholders to provide the Committees on Appropriations of the House of Representatives and the Senate a detailed strategic plan by May 5, 2018, spelling out how it intends to make big data sustainable, interoperable, accessible, and usable.

**GOALS**

- **Data Infrastructure**
  - Optimize data storage and security
  - Connect NIH data systems

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

- **Data Mgmt, Analytics, & Tools**
  - Support useful, generalizable, & accessible tools and workflows
  - Broaden utility of and access to specialized tools
  - Improve discovery & cataloging resources

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

- **Stewardship and Sustainability**
  - Develop policies for a FAIR data ecosystem
  - Enhance stewardship
NIH Strategic Plan for Data Science: 
Cross-Cutting Themes

- Support **common infrastructure and architecture** on which more specialized platforms can be built and interconnected
- **Leverage commercial tools, technologies, services, and expertise**; adopt and adapt tools and technologies from other fields
- Enhance the nation’s biomedical **data-science research workforce**: improved training programs, novel partnerships
- Ensure NIH-supported data resources are **FAIR**
- Ensure **information security** of patient and participant data
- Improve the ability to capture, curate, validate, store, and analyze **clinical data** for biomedical research
- With community input, develop, promote—and refine as needed—**data standards**
- **Coordinate and collaborate** with federal, private and international funding agencies and organizations to promote **economies of scale and synergies and prevent unnecessary duplication**
Create a sustainable infrastructure

Federated Data Commons Model

**NIH Data Commons**
- GTEx
- TOPMed

**Interoperability Tools**
- Common user authentication system
- Shared APIs for data access & computing
- Adoption of FAIR Principles
- Docker containers
- Workflows management
- Digital objects catalogues and IDs
- Data standards and ontologies

**NHGRI Genomic AnVIL**

**All of Us**
NIH strongly encourages open access Data Sharing Repositories as a first choice.

Options for scaled implementation for orphan datasets

Datasets up to 2 gigabytes

**PubMed Central**
- PMC stores publication-related supplemental materials and datasets directly associated with publications. Up to 2 GB.
- Generate Unique Identifiers for the stored supplementary materials and datasets.

Datasets up to 20*gigabytes

**Dryad and FigShare, or other**
- Assign Unique Identifiers (DOI) to datasets associated with publications and link to PubMed
- Store and manage datasets associated with publication, up to 20* GB.

High priority Datasets T-Pbyte range

**NIH managed repositories**
- Store and manage large scale, high priority NIH datasets (Partnership with STRIDES)
- Assign Unique Identifiers (dbGap), implement authentication, authorization & access control (eRA Commons)

PubMed Central
- PMC stores publication-related supplemental materials and datasets directly associated with publications. Up to 2 GB.
- Generate Unique Identifiers for the stored supplementary materials and datasets.

Dryad and FigShare, or other
- Assign Unique Identifiers (DOI) to datasets associated with publications and link to PubMed
- Store and manage datasets associated with publication, up to 20* GB.

NIH managed repositories
- Store and manage large scale, high priority NIH datasets (Partnership with STRIDES)
- Assign Unique Identifiers (dbGap), implement authentication, authorization & access control (eRA Commons)
A Platform for Biomedical Discovery and Data-Powered Health
The 21st Century Collection

- Innovative attribution
- Automated indexing
- Personalized presentation & delivery

CUSTODIAN | CONNECTOR | DISCOVERER
Creating the 21st Century Collection of BIOMEDICAL KNOWLEDGE

• Permanent, discoverable archive of text and data
• Listening to trends in Science & Scientific Communication
  – Open Science
  – Preprints & other interim products of research
  – Data Sharing
• Preparing for (and leading) new directions in scientific communications through collaboration
NLM Collections Development Policy is set by the Board of Regents

PubMed
- 28 M records
- Only bibliographic records

MEDLINE
- ~5600 selected journal titles
- Only bibliographic records
- 93% of PubMed

National Library of Medicine Collection
- ~17,000 serial titles
- All Serials – journals, annuals, statistics, etc.
- Discoverable in NLM Catalog & LocatorPlus
- NLM provides ILL and ensures archiving

PMC = PubMed Central Archive
- ~2,000 full participation journals
- 5 million full text articles
- ~2.8 M federally funded public access articles
- Bibliographic records display in PubMed

PubMed • ~28 M records • Only bibliographic records
MEDLINE • ~5600 selected journal titles • Only bibliographic records • 93% of PubMed
National Library of Medicine Collection • ~17,000 serial titles • All Serials – journals, annuals, statistics, etc. • Discoverable in NLM Catalog & LocatorPlus • NLM provides ILL and ensures archiving
PMC = PubMed Central Archive • ~2,000 full participation journals • 5 million full text articles • ~2.8 M federally funded public access articles • Bibliographic records display in PubMed
NIH Public Access Policy

The NIH Public Access Policy implements Division F Section 217 of PL 111-8 (Omnibus Appropriations Act, 2009). The law states:

The Director of the National Institutes of Health ("NIH") shall require in the current fiscal year and thereafter that all investigators funded by the NIH submit or have submitted for them to the National Library of Medicine's PubMed Central an electronic version of their final, peer-reviewed manuscripts upon acceptance for publication, to be made publicly available no later than 12 months after the official date of publication: Provided, that the NIH shall implement the public access policy in a manner consistent with copyright law.
Reporting Preprints and Other Interim Research Products

Notice Number: NOT-OD-17-050

Key Dates
Release Date: March 24, 2017
Effective date for application: Applications submitted for the May 25, 2017 due date and thereafter
Effective date Research Performance Progress Report (RPPR): RPPRs submitted on or after May 25, 2017

Related Announcements
NOT-OD-17-006: Request for Information (RFI): Including Preprints and Interim Research Products in NIH Applications and Reports
Issued by
National Institutes of Health (NIH)

Purpose
The NIH encourages investigators to use interim research products, such as preprints, to speed the dissemination and enhance the rigor of their work. This notice clarifies reporting instructions to allow investigators to cite their interim research products and claim them as products of NIH funding.

Definitions
Interim Research Products are complete, public research products that are not final.
A common form is the preprint, which is a complete and public draft of a scientific document. Preprints are typically unreviewed manuscripts written in the style of a peer-reviewed journal article. Scientists issue preprints to speed dissemination, establish priority, obtain feedback, and offset publication bias.
Another common type of interim product is a preregistered protocol, where a scientist publicly declares key elements of their research protocol in advance. Preregistration can help scientists enhance the rigor of their work.
NIH Data Sharing Policy

- Data sharing plan in proposals (unscored)
  - Publishing
  - Researcher Efforts
  - Data Enclave
  - Data Archive
  - Mixed Mode

- NIH funds can be used to support data sharing
Data Discovery in PMC & PubMed
Journal Article as a Nexus for Discovery

PubMed Central (PMC) contains full text records of publisher’s versions and author manuscripts.

PubMed contains citations, abstracts, and associated metadata.

https://www.crossref.org/blog/the-article-nexus-linking-publications-to-associated-research-outputs/
NLM Strategic Plan, Objective 1.1

“NLM will stimulate new forms of scientific communication and become the library of the future, one of connections between and among literature, data, models, and analytical tools. Creating efficient ways to link the literature with associated datasets enables knowledge generation and discovery.”

Leverage existing NCBI resources to provide users with new ways to deposit, find, access, and re-use (when applicable) data.

Support NIH efforts to enhance rigor and further support research that is reproducible, robust, and transparent.
Data Availability Statements in PubMed Central (PMC)

Examples of what you will find in these sections:

- “Data available upon request”
- References to supplementary material
- NCBI database accession numbers (GEO, GenBank, dbGap, SRA)
- General data repository record DOIs (figshare, Dryad)
- Government data repository links
- Personal servers

Data availability

The datasets and computer code produced in this study are available in the following databases: Hi-C and RNA-seq data: Bioproject PRJNA464299.
The fasta sequences of the Syn-HiC region and its native counterpart are available as a supplementary file (Dataset EV1).
Previously published datasets used in this work are described in Table 3.

PMC Search: has data avail[filter]
Supplementary Data in PubMed Central (PMC)

Examples of what you will find in supplementary material:

- Computer code
- Mathematical or computational models
- Audio or video clips
- Oversized tables
- Intervention protocols
- Primary or supplementary data sets
- Expanded methodology sections
- Additional figures

PMC Search: has suppdata[filter]

Source: Publication Manual of APA (6th ed.)
Data Citations in PMC

“Data citations should facilitate access to the data and any associated metadata, code, or related materials.”

PMC Search: has data citations[filter]

http://jats4r.org/data-citations
~20% of articles in PMC (~930,000) have associated supplementary material, a data availability statement, and/or data citation(s)

**Challenge:**
Inconsistent (or non-existent) policies on how to make data available and cite it at the funder and journal level result in inconsistent data association practices

Source: PubMed Central, 16 July 2018
Data Links in PubMed: Secondary Source ID

Secondary Source ID data sources:
- Publishers
- NLM indexers
- PMC

Challenges:
- NLM indexing resources were reduced in 2016, leading NLM to explore alternative options
- No incentives for journals/publishers to supply these metadata to NLM

Example PubMed Searches: 
- hasdatabanklist
- genbank[si] OR figshare[si]
Data Links in PubMed: LinkOut

Links to the materials directly supporting the research discussed in the cited article, including data sets from experiments/studies accessory graphics, images, sound, and multimedia files related to the article.

- American Society for Microbiology - videos and podcasts (asmmedia)
- DataMed.org (DataMed)
- Dryad Digital Repository (dryadb)
- figshare (figshare)
- NCI caNanoLab Data Portal (ncicananolab)
- NHGRI Bioinformatics and Scientific Programming Core (nhgribspc)
- NIMH Data Archive (NDA)
- Nuclear Receptor Signaling Atlas (NURSA) (nursa)
- Princeton University MicroArray database (PUMAdb)
- Surface Management System Database (SumsDB)
Future Direction: Data Citations in PubMed Labs

Future Direction: Exposing Data in PMC

Secondary source IDs

Secondary source IDs
figshare/10.6084/m9.figshare.c.4009681
Dryad/10.5061/dryad.6cp6f

Data accessibility

Data for spectral properties of ocular sparks, copepod eye reflectance and all data of the two experiments are available at http://dx.doi.org/10.5061/dryad.6cp6f [48].

Future Direction: Linking Publications to Data in My Bibliography
Additional Data Resources at NLM
ClinicalTrials.gov is a database of privately and publicly funded clinical studies conducted around the world.

Explore 279,151 research studies in all 50 states and in 204 countries.

ClinicalTrials.gov is a resource provided by the U.S. National Library of Medicine.

IMPORTANT: Listing a study does not mean it has been evaluated by the U.S. Federal Government. Read our disclaimer for details.

Before participating in a study, talk to your health care provider and learn about the risks and potential benefits.
ClinicalTrials.GOV: Informational Scaffold

- Journal publications
- Results database entries
- Conference abstracts
- CSRs
- Other Information (e.g., press releases, news articles, editorials)

ClinicalTrials.gov Record

- IPD
- Full protocols
- SAPs
- Other study documents
The NIH Common Data Elements (CDE) Repository has been designed to provide access to structured human and machine-readable definitions of data elements that have been recommended or required by NIH Institutes and Centers and other organizations for use in research and for other purposes. Visit the NIH CDE Resource Portal for contextual information about the repository.

The Repository is a platform for identifying related data elements in use across diverse areas, for harmonizing data elements, and for linking CDEs to other existing standards and terminologies, including the value sets in the Value Set Authority Center (VSAC).

Search
Search for individual data elements, by definition, users or sources. Search for sets of data elements ("boards") identified by a particular group for a particular use (e.g. particular research solicitation).

Compare / Harmonize
Analyze and resolve differences between data elements. Assure that your forms are using variables that will be usable by certified EHRs.

Create
Draw upon the experience of colleagues and others to design unique data elements and measures.
Common Data Elements
Make data findable, interoperable

- Structured human & machine readable definitions of NIH CDEs allowing
  - Search for individual CDE or sets per programs
  - Compare & harmonize similar but distinct CDEs
  - Select or create CDEs with minimal duplication
NIH Data Sharing Repositories

This table lists NIH-supported data repositories that make data accessible for reuse. Most accept submissions of appropriate data from NIH-funded investigators (and others), but some restrict data submission to only those researchers involved in a specific research network. Also included are resources that serve as a portal for information about biomedical data and information sharing systems. The table can be sorted by repository name and by NIH Institute or Center and may be searched using keywords so that you can find repositories more relevant to your data. Links are provided to information about submitting data to and accessing data from the listed repositories. Also, we have asked all repositories to indicate whether or not they have—current NIH funding support; open data submission; open data access; open time frame for data submission; and sustained support. Their Yes or No responses are included in the five columns to the right of the table. The information in the BMIC list is also available in a downloadable Excel version.

In our query to the repositories, we acknowledged that some repositories may have restrictions or limitations on data submission and requirements for data access. Additional information about the repositories and points-of-contact for further information or inquiries can be found on the websites or links shown for each. Are we missing a data sharing repository that should be included in this list? Contact us.

<table>
<thead>
<tr>
<th>ICO Fund</th>
<th>Repository Name</th>
<th>Repository Description</th>
<th>Data Submission Policy</th>
<th>Access to Data</th>
<th>Current NIH funding support</th>
<th>Open data submission</th>
<th>Open data access</th>
<th>Open time frame for data deposit</th>
<th>Sustained support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common Fund</td>
<td>Epigenomics</td>
<td>Epigenomic, 6 histone modification marks, DNAase I, DNA methylation, transcriptome for wide variety of cell types and tissues.</td>
<td>Not applicable</td>
<td>How to access Epigenomics data</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Common Fund</td>
<td>exRNA Atlas</td>
<td>Includes exRNA profiles derived from various bioluid profiles and conditions. Currently stores data profiled from small RNA sequencing assays.</td>
<td>Not applicable</td>
<td>How to access exRNA Atlas data</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Common Fund</td>
<td>GTEx</td>
<td>The Genotype-Tissue Expression (GTEx) project aims to provide the scientific community a resource with which to study human gene expression and regulation and its relation to human biology and disease.</td>
<td>Project data only</td>
<td>How to access GTEx data</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Office of Data Science Strategy
NIH Office of the Director

Susan Gregueric, PhD
Senior Advisor to the Director for Data Science
NIH Data Science Fellowships – Happening NOW

Graduate Data Science Summer Program

- OITE in partnership with ODSS
- Consortium of local universities, but students from all universities can apply
  - UVA, George Mason, George Washington, UMD
- Pilot driven by discussion with local universities consortium
- Masters students for summer program at NIH
- Intramural placement
- ~15 interns for summer 2019
- Call for projects – late November

https://www.training.nih.gov/data_science
NIH Data Science Fellowships – Happening NOW

Coding it Forward

- Civic Digital Fellowship program
- Student-led non-profit to place tech-savvy students in federal agencies; pipeline to public service for technology students
- Mentor/project call for 2019
- 10-12 week summer program
- Placement in administrative/extramural funding offices at NIH
- ODSS will coordinate some central NIH activities so fellows connect with each other on campus
- ~10 fellows for summer 2019

https://www.codingitforward.com/
NIH Data Science Fellowships – Happening SOON

<table>
<thead>
<tr>
<th>Overview</th>
<th>PLAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>• One or two year positions in high-impact NIH programs</td>
<td>• Goal: first cohort starts fall 2019</td>
</tr>
<tr>
<td>• Data science experts, not necessarily familiar with biomedical research</td>
<td>• Launch multiple onboarding/training mechanisms to apply</td>
</tr>
<tr>
<td>• Opportunity to work with large volumes of biomedical research data; impact on public health</td>
<td>• Advantages – flexibility in salary, no one-size fits all</td>
</tr>
<tr>
<td>• Learn from NIAID’s program</td>
<td>• Program evaluation in 2024</td>
</tr>
<tr>
<td>• ~2-5 fellows in first cohort</td>
<td></td>
</tr>
</tbody>
</table>

NIH Data and Technology Advancement (DATA) Fellows
Provisions for Data Management and Sharing Policy

Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research

Notice Number: NOT-OD-19-014

Key Dates

Release Date: October 10, 2018
Response Date: December 10, 2018

Purpose

Introduction

The purpose of this Notice is to solicit public input on proposed key provisions that could serve as a foundation for a future NIH policy for data management and sharing. NIH has a longstanding commitment to making the results and accomplishments of the research that it funds and conducts available to the public. In NIH's view, data should be made as widely and freely available as possible while safeguarding the privacy of participants and protecting confidential and proprietary data. Increasing access to scientific data results in new knowledge, producing and informing future research pathways, and increases the return on investment of scientific research data sets, enabling hard-to-generate data, and exploring new research results that sustain and inform research in key scientific areas. Moreover, sharing data enhances research reproducibility, improves access to research data, and enables data sharing to more effectively test the validity of research findings, strengthen analyses, develop new research areas, and explore new research directions.

Background

In 2015, NIH established a strategic plan for increasing access to scientific publications and digital scientific data from NIH-funded research. The plan includes developing data sharing policies that are consistent with the principles of FAIR (Findable, Accessible, Interoperable, and Reusable) data. The plan also establishes a strategic framework for NIH data management and sharing. NIH in 2018 released a Request for Information on Strategies for NIH Data Management, Sharing, and Citation and, in 2017 co-hosted a workshop on data management and sharing with the National Science Foundation. These efforts have helped to inform NIH's thinking about key provisions for a future NIH policy for the management and sharing of data.
What Do We Need to Know?

• Before we can start talking about a NIH policy for sharing and managing data, we need to think about what the key components are, such as:
  – The **definition** of scientific data
  – The **requirements** for data management and sharing plans
  – The optimal **timing** to consider in implementing various parts of a new policy
    • Phased adoption?
    • Incremental implementation?
Ongoing Implementation Considerations

- Budgeting for data management and sharing
- Use of existing repositories; criteria for identifying
- Community expectations for shared data
Questions and discussion
Reaching NLM

@NLM_news
@NLMdirector

patti.brennan@nih.gov