Priorities for Research Data and Information: Perspectives from the National Library of Medicine (NLM) and National Institutes of Health (NIH)

Dina N. Paltoo, Ph.D., M.P.H.

Assistant Director for Policy Development
Office of the Director, U.S. National Library of Medicine
National Institutes of Health
U.S. Department of Health & Human Services

Board on Research Data and Information/US CODATA
National Academies of Science, Engineering, and Mathematics
May 8, 2019

National Library of Medicine

- A component of the NIH (1968) and a leader of research in biomedical informatics and data science

- **The world's largest biomedical library (1836)**
  - NLM makes almost 300 databases and online services freely available to support health care, public health, disease prevention and wellness, biomedical research, and innovation
  - Every day, NLM
    - Serves more than 5 million users
    - Receives up to 15 terabytes of new data
    - Provides more than 115 terabytes of information

- Facilitate open science and scholarship by making digital research objects Findable, Accessible, Interoperable, & Reusable (FAIR), and Attributable & Sustainable

NLM lives at the intersection of Data Science and Open Science

NLM's Systems and Services

PubMed

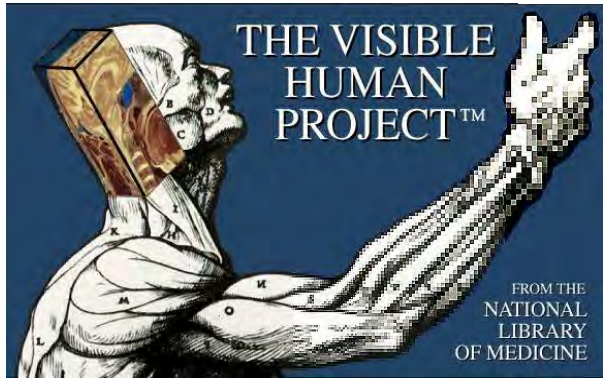
MeSH on Demand

WISER

Search DRUG INFORMATION PORTAL

Pillbox

ClinicalTrials.gov
A service of the U.S. National Institutes of Health



NCBI

MedlinePlus
Trusted Health Information for You

PEOPLE LOCATOR

Profiles in Science

PubMed Health

MEDLINE

NIH NLM NNLM

Genetics Home Reference

NIH Manuscript Submission System

PubChem

itk

MTI NLM Medical Text Indexer
Biomedical Literature → MTI MTIFL → MeSH Suggestions

Haz-Map @NIH

ACCESS GUDID

LHNCBC

PubMed labs

TOXNET TOXICOLOGY DATA NETWORK

OPEN

PubMed Central

GTR Genetic Testing Registry

ClinVar Clinically relevant variation

MedGen Conditions with a genetic component

TOXMAP classic

HISTORY OF MEDICINE

DOCLINE®

db GaP GENOTYPES and PHENOTYPES

Protein BLAST
protein ▸ protein

Challenges for Data Stewardship

- Time and effort
  - Determine which data to preserve
  - Clean data, put in accessible format (consistency; standardized elements)
  - Provide metadata
  - Limited training in data management and sharing
- Infrastructure
  - Sustainability and long-term preservation
  - Procedures for providing data access
- Human resources and burden (for all stakeholders, including librarians, data managers, data scientists, federal staff)
- Value assessment
- Curation at scale
- Lack of rewards/incentives
  - Citations/publications used for academic credit
  - Carrot and stick vs. benefits
- Considerations for ethical, legal, and social implications, human participant protections, privacy and trust
- Continuous advances in technology
- Proprietary interests
  - Researchers want to analyze & publish first
  - Institutions/Individuals want to protect competitive advantage
  - Licensing for data reuse
- Utility of large datasets is limited; data are:
  - Disconnected
  - Incompatible/lack of interoperability
  - Difficult for users to find and access
  - Expensive to generate, store, download, and compute on
- Compliance and enforcement
- Policy Coordination (e.g., across agencies, funders, publishers, journals)

Challenges for Data Stewardship

- Incentives – Establish and align incentives to promote open science practices (e.g., sharing data, adopting standards, using appropriate repositories)
  - Strategically align incentives across entire ecosystem to maximize impact
  - Likely best done domain-by-domain

- At-scale Curation and Provenance – Rapid increase in number of digital research objects (DROs) and the need to find, associate, and monitor their versions is outstripping the ability to apply consistent, useful metadata to them. Move from applying metadata to having DROs imply their metadata
  - Move from search to learning, and from learning to awareness
  - Draw from other approaches (e.g., artificial intelligence/machine learning, blockchain)

- Sustainability – Assure return on investment (ROI) by assessing the value of particular investments in the ecosystem (e.g., in infrastructure, data acquisition, preservation, policy changes, etc.)
  - Rigorous cost vs benefit analyses
  - Metrics and models

ADDRESSING THE CHALLENGES…


NIH U.S. National Library of Medicine
A Platform for Biomedical Discovery and Data-Powered Health
Strategic Plan 2017-2027
NIH NCI
NIH CC
NIDCR
NIAID
NLM
NIAMS
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
NIH U.S. National Library of Medicine

NLM Implementation Activities

- Blue Ribbon Panel Review of Intramural Research
- Data Science Research RFI
- NSF-NLM Data Science MOU
- Reproducibility Workshop
- Data Science Drivers Workshop
- Chief Data Science Innovator Initiative
- Assessment of NIH Data Science Training
- Data Science Core Skills Analysis
- Data Science Librarians Needs Assessment
- Open Science Staff Initiative
- Aligning Curation Across Data & Literature
- Dataset Metadata Model Initiative
- Outreach Audit & Outreach Future Initiative
- User Experience / User Development Initiative
- Assessment of Tools to Evaluate Resources
- Assessment of Comparative Web Metrics
- Assessment of IT – 5 Teams
- Assessment of Products, Services, and Resources
- Assessment of Data Center & Cloud Use
- Assessment of Trans-NLM Central Functions
- Workspace Audit & Initiative
- Project Managemnet for Implementation Initiative
- Internal and External Communications Plans & Information Resources

Accelerate discovery & advance health through data-driven research

Goal 1

1.1 Connect the resources of a digital research enterprise

1.2 Advance research and development in biomedical informatics and data science

1.3 Foster open science policies and practices

1.4 Create a sustainable institutional, physical, and computational infrastructure

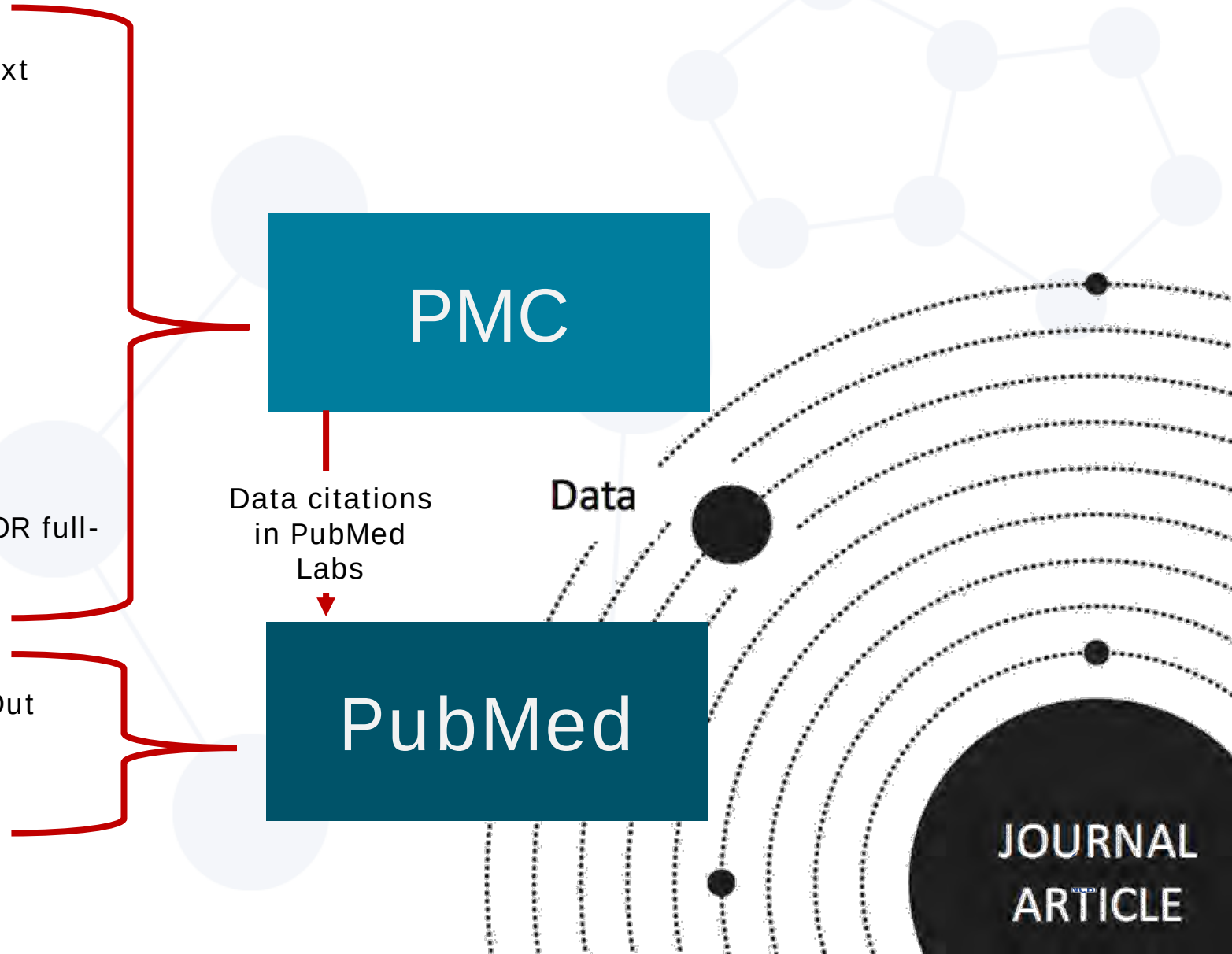
Building Publication-Data Links

Supplementary data
- Files stored and made available with full-text article
- Provided by
  - Publishers / journals
  - Authors via NIHMS

Data availability statements
- Text within full-text article
- Provided by
  - Publishers / journals
  - Authors via NIHMS

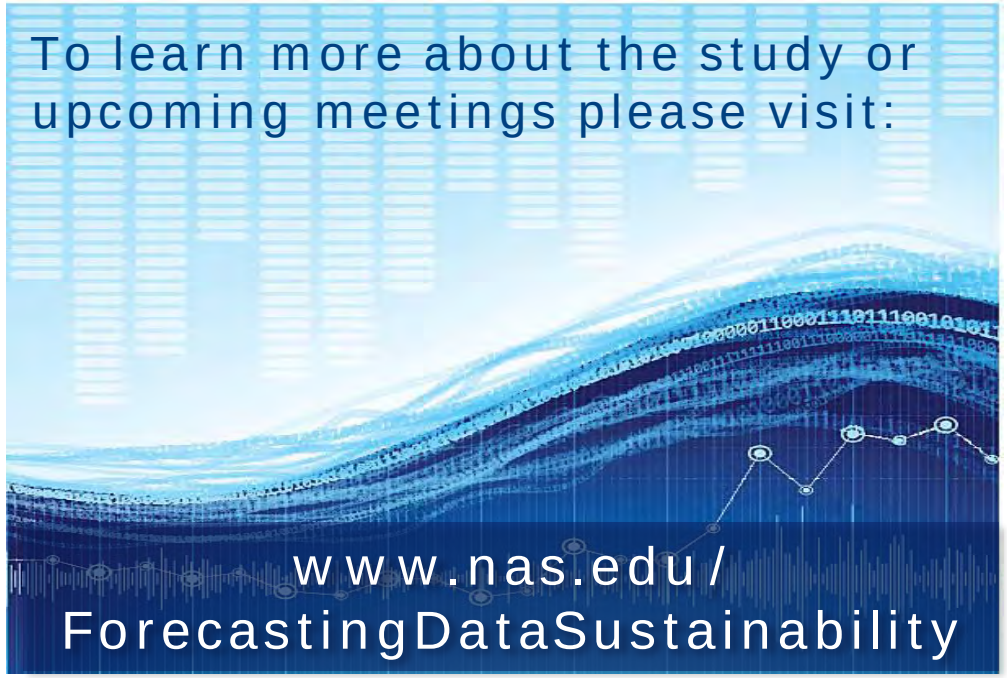
Data citations
- Machine-readable metadata in references OR full-text
- Provided by publishers/journals

Other data links
- Repository-provided dataset links via LinkOut
- NLM-indexer supplied dataset identifiers
- Publisher-supplied dataset identifiers



NASEM Study on Forecasting Costs for Preserving, Archiving, and Promoting Access to Biomedical Data

- *Commissioned by NLM*
- *A cross-disciplinary committee of experts to develop and demonstrate a framework for forecasting long-term costs of data, examining:*
  - *Economic factors of data set life-cycle costs*
  - *Cost consequences of (de-)accessioning data*
  - *Economic factors for designating data as high value*
  - *Data collection and modeling assumptions*
  - *Anticipated technology developments & disruptors*
  - *Critical factors for researcher adoption*



TIMELINE:	Sept 2018	June 2019	Fall 2019	Spring 2020	Fall 2020
	Study begins	Public workshop	Workshop report released	Final report released	Dissemination activities

The NIH Strategic Plan for Data Science

Requested by Congress, the NIH Strategic Plan will:

- Modernize the data resource ecosystem to increase utility for researchers
- Enhance data sharing, access and interoperability
- Modernize infrastructure, increase capacity

Data science is an interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data

Overarching goals:

Support Highly Efficient and Effective Data Infrastructure for Biomedical Research	Promote the Modernization of the Research Data Resources Ecosystem	Support the Development and Dissemination of Advanced Management, Analytics, and Visualization Tools	Enhance Workforce Development for Biomedical Data Science	Enact Appropriate Policies to Promote Stewardship and Sustainability

National Institutes of Health
Office of Strategic Coordination - The Common Fund

Search Common Fund

Common Fund Programs | Common Fund Research Funding | News & Media | Common Fund Highlights | About Common Fund

New Models of Data Stewardship

New Models of Data Stewardship

- Highlights
- Frequently Asked Questions
- Funded Research

NIH Data Commons Pilot Phase

Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES)

Amazon Web Services joins NIH's STRIDES Initiative to harness latest cloud technologies for biomedical researchers

Learn More

Program Snapshot

The New Models of Data Stewardship (NMDS) program is designed to enhance biomedical discovery and improve efficiency through new digital data management strategies. These strategies contribute to NIH efforts to develop and sustain a modern biomedical data ecosystem as described in the NIH Strategic Plan for Data Science. They also aim to make data for research findable

Announcements

The STRIDES Initiative announces new agreement with Amazon Web Services! The New Models of Data Stewardship program is announcing its new agreement with Amazon Web Services (AWS) through the STRIDES initiative. This is the second agreement with a cloud service provider, following the first with Google Cloud. Both agreements will enable NIH to make high-value data sets more accessible to researchers, help optimize technology-intensive research, and lower economic barriers for research. Read the blog

NIH Policy Development Process: Data Management and Sharing

- Oct. 2018 NIH solicited stakeholder feedback on proposed provisions for a data management and sharing policy (NOT-OD-19-014)
  - Two public webinars with ~800 participants (combined)
  - 189 submissions from national and international stakeholders
  - Considerations for:
    - The definition of Scientific Data
    - Requirements for Data Management and Sharing Plans
    - Optimal timing and phased adoption to consider for future policy implementation
- Next steps:
  - Consider public comments and release draft policy for public input
  - Release final policy
  - Policy ≠ Implementation: consider guidance to accompany future NIH policy for data management and sharing

Options for Sharing Data

NIH strongly encourages use of existing NIH repositories as a first choice for sharing data

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

Options of scaled implementation for sharing datasets

Datasets up to 2 gigabytes	Datasets up to 20*gigabytes	High Priority Datasets petabytes
PubMed Central • PMC stores publication-related supplemental materials and datasets directly associated publications (up to 2 GB) • Generate Unique Identifiers for the stored supplementary materials and datasets.	**Use of commercial and non-profit repositories** • Assign Unique Identifiers to datasets associated with publications and link to PubMed • Store and manage datasets associated with publication, up to 20* GB.	**STRIDES Cloud Partners** • Store and manage large scale, high priority NIH datasets (Partnership with STRIDES) • Assign Unique Identifiers, implement authentication, authorization & access control

Characteristics of appropriate repositories?


TRUST
Concepts
Opening remarks and mini sessions on the creations of trust and repositories
Examples
Sessions on previously established TRUST repositories
Challenges
Group sessions on applying requirements to biomedical sciences repositories
Community
What are the challenges and how can we help each other?
Networking Sessions
NIH Workshop on Data Repositories for Biomedical Sciences
2019
Our speakers include Robert Downs, Jared Lyle, John Westbrook, as well as
Susan Gregurick
Senior Advisor at ODSS
Division Director NIGMS
Dawei LIN
Senior Advisor NIAID
Ingrid Dillo
Deputy Director DANS
Johnatan Crabtree
Director
Cyberinfrastructure
UNC
April 8th : 9am - 5:15pm
April 9th: 9am-12:15pm
5601 Fishers Lane 1D06AB Rockville MD
PDB, ICPSR, IDA,TCIA, NIF/dkNET, ImmPort, PhysioNet, ZEBrA, TalkBank, FITBIR, WormBase, UniProt, dbSNP, DASH, GEO, BioLINCC, GlyGen, OncoMX, eyeGENE, ICE
Health and Human Services
NIH-ODSS-NIAID
Webnar info at
https://datascience.nih.gov/community

Other NIH Activities

- NIH Advisory Committee to the Director (ACD) Artificial Intelligence Working Group

- NIH and HHS implementation of the *Open, Public, Electronic, and Necessary (OPEN) Government Data Act* – Section II of the Foundations for Evidenced-Based Policy Making Act of 2018 (Public Law 115-435)
  - Applies to data maintained by the government (i.e., administrative/enterprise data)
  - Requires federal agencies to publish their open government data assets, using machine-readable data formats.
  - Each agency shall develop and maintain an inventory for all data assets created by, collected by, under the control or direction of, or maintained by the agency

- NIH Graduate Data Science Summer Program

A Future for Data Stewardship

What we need to do to get there…

- Models for data stewardship and FAIRness
- Citation and incentivization
  - National Academies of Sciences, Engineering, and Medicine (NASEM) Roundtable on Aligning Incentives for Open Science
- Value assessment
  - 2017 NIH- NSF Science of Science Innovation Policy (SciSIP) Workshop on The Value of Data Sharing
  - NASEM study on "Forecasting Costs for Preserving, Archiving, and Promoting Access to Biomedical Data: A Study and Workshop for the National Library of Medicine"
- At-scale curation and provenance
- Policy and implementation
- Coordinate and partner with other funders and organizations
  - Interagency activities in open science

THANK YOU!

Dina.Paltoo@nih.gov