Assessing dataset and data resource value and reach

Susan Gregurick, Ph.D.

Associate Director for Data Science and Director, Office of Data Science Strategy

February 19, 2020

National Institutes of Health
Office of Data Science Strategy

Thank You!
For all that you have done and will do as a result of this workshop

Co-Chairs:

Daniella Lowenberg, California Digital Library

Dr. Warren Kibbe, Chief Data Officer for the Duke Cancer Institute

NIH Planning Committee:

Kim Pruitt, Fenglou Mao, Dawei Lin, Jennie Larkin, Elaine Collier, Susan Wright, Lisa Federer, Matthew McAuliffe, Christine Melchior, and Minghong Ward

We all have similar goals

Better health, longer life, reduced illness and disability

Enabled through new research, development of cutting edge technologies and through the **useful applications of data**

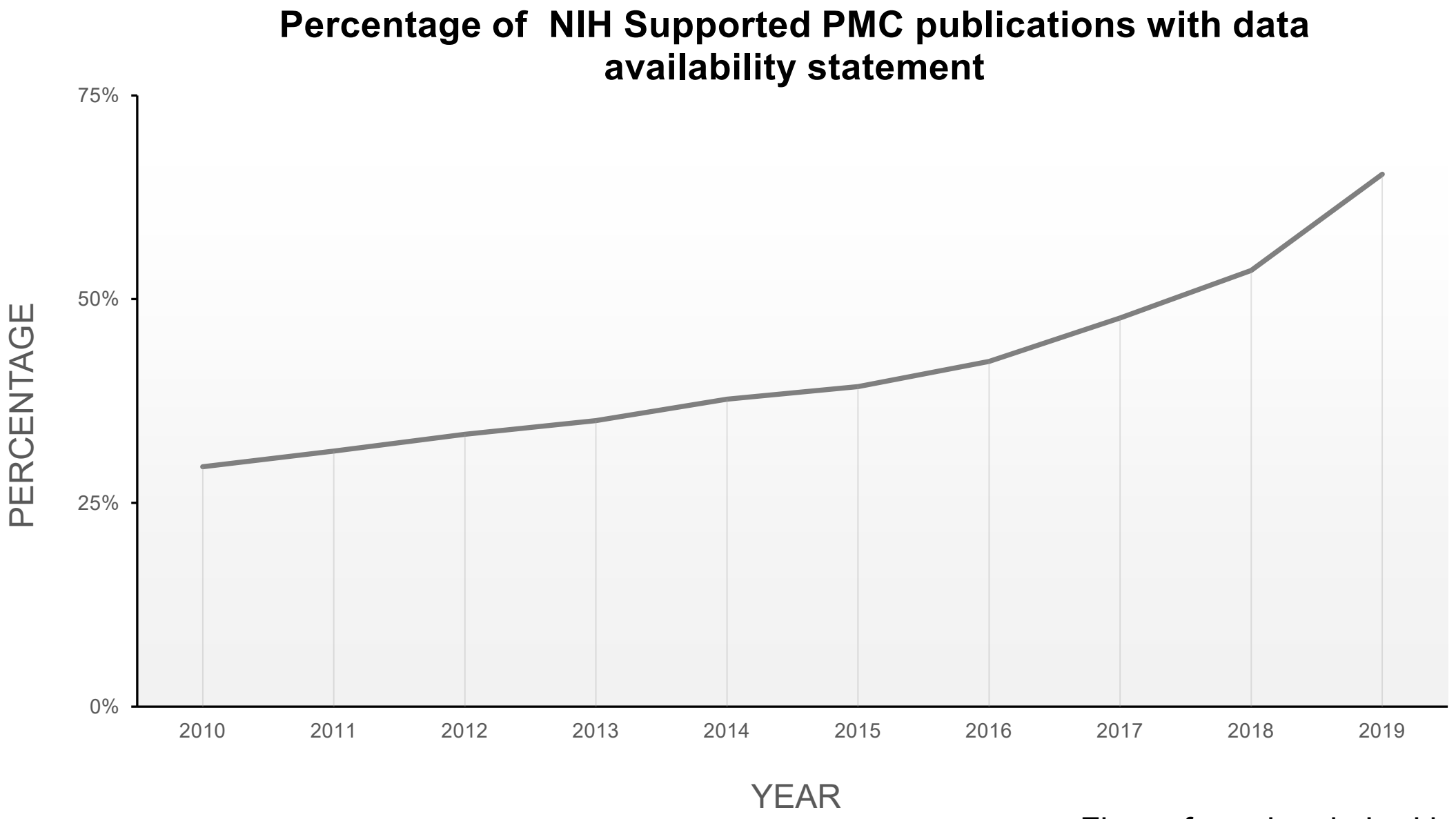


Figure from Jennie Larkin

NIH Data Management and Sharing Policy Development

- **Researchers** with NIH-funded or conducted research projects resulting in the generation of scientific data will be required to submit a Plan
- **Plans** should explain how scientific data generated by a research study will be managed and which of these scientific data will be shared

Community Input Solicited
- 189 submissions from national and international stakeholders

→

Identified need for appropriate infrastructure
- policy and implementation to go 'hand-in-hand'

→

Develop draft policy for data management and sharing and related guidance

→

Released draft for community input

→

Release final policy (spring 2020)

FAIR and data sharing

Researchers understand the concepts behind FAIR but need guidance on how to put FAIR into practice

The FAIR principles (Findable, Accessable, Interoperable and Reusable) are familiar to many.

However, there is confusion about what FAIR means in practice.

It can be time consumingto create FAIR datasets.

Challenges

Prioritizing dataset annotation and curation when it is time consuming and perceived as an added burden

Selecting metadata to annotate their data that is compatible with other datasets and tools in the ecosystem

Where to put the data so it can be stored for the long term and securely accessed by authorized users (as appropriate)

Researchers with different needs requires multiple options

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

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

Options of scaled implementation for sharing datasets

Datasets up to **2 gigabytes**

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.

Datasets up to **20*gigabytes**

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.

High Priority Datasets **petabytes**

STRIDES Cloud Partners

- Store and manage large scale, high priority NIH datasets. (Partnership with STRIDES)
- Assign Unique Identifiers, implement authentication, authorization and access control.

NIH supports many repositories for biomedical data sharing

DASH Data and Specimen Hub

NIAGADS

CELL IMAGE LIBRARY

RCSB PDB PROTEIN DATA BANK

GEO Gene Expression Omnibus

PhysioNet the research resource for complex physiologic signals

AphasiaBank

dbSNP Short Genetic Variations

AMP-AD Knowledge Portal

OpenfMRI

ImmPort Bioinformatics For The Future Of Immunology

BioLINCC Biologic Specimen and Data Repository Information Coordinating Center Web Access to NHLBI Biospecimens and Data

FITBIR Federal Interagency Traumatic Brain Injury Research INFORMATICS SYSTEM

NIMH National Institute of Mental Health

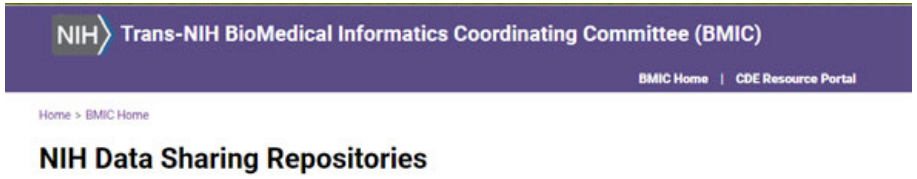
Other Widely Used Repositories

- **Dryad**
- **Elsevier Mendeley**
- **FigShare**
- **Zenodo**

How to find Data Repositories?

- **BMIC Data Repository Listing**

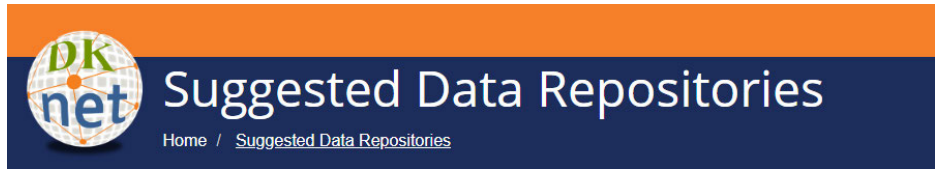
https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html



- **SciCruch/dkNET**

  Organized by repository type and scientific area.

https://dknet.org/about/Suggested-data-repositories



- **FAIRsharing**

https://fairsharing.org/

FAIRsharing.org
standards, databases, policies

- **DataMed**

https://datamed.org/

dataMED
BETA version

Optimized Funding for NIH Data Repositories and Knowledgebases

- 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

Optimized Funding for NIH Data Repositories and Knowledgebases

Funding Opportunities

- NIH released two funding opportunities on Jan. 17 to support biomedical data repositories and knowledgebases:
- Biomedical Data Repository (PAR-20-089)
- Biomedical Knowledgebase (PAR-20-097)

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

Motivation & Objectives of Proposed Desirable Characteristics

- NSTC **Subcommittee on Open Science**: Improve the consistency of guidelines and best practices Federal agencies provide to implement Public Access Plans
- Proposed set of desirable characteristics of data repositories for data resulting from Federally funded research
- Characteristics can apply to repositories operated by government or non-governmental entities

Slide provided by Jerry Sheehan

Characteristics intended to:

- Support data discoverability, management, and sharing in a user-friendly manner, consistent with FAIR principles
- Be consistent with certification criteria (e.g., CoreTrustSeal), but achievable by many more repositories
- Be enduring, but evolve over time

NOT Intended to:

- Describe an exhaustive set of design features, functional requirements or implementation details for data repositories
- Override other requirements, e.g., Federal IT security, privacy
- Be used by Federal agencies to certify data repositories

Slide provided by Jerry Sheehan

Now Accepting Comments!

Comments will help the NIH and Subcommittee refine and develop a common set of characteristics to support efforts to improve the management, sharing, and long-term preservation of data

- **OSTP/STPI** will assist in reviewing comments
- **URL:** https://www.federalregister.gov/documents/2020/01/17/2020-00689/request-for-public-comment-on-draft-desirable-characteristics-of-repositories-for-managing-and
- **Closing Date:** March 6, 2020
- **Email:** OpenScience@ostp.eop.gov

Slide provided by Jerry Sheehan

Research articles

Survival of retinal ganglion cells after damage to the occipital lobe in humans is activity dependent

Colleen L. Schneider, Emily K. Prentiss, Ania Busza, Kelly Matmati, Nabil Matmati, Zoë R. Williams, Bogachan Sahin and Bradford Z. Mahon

Published: 27 February 2019 | https://doi.org/10.1098/rspb.2018.2733

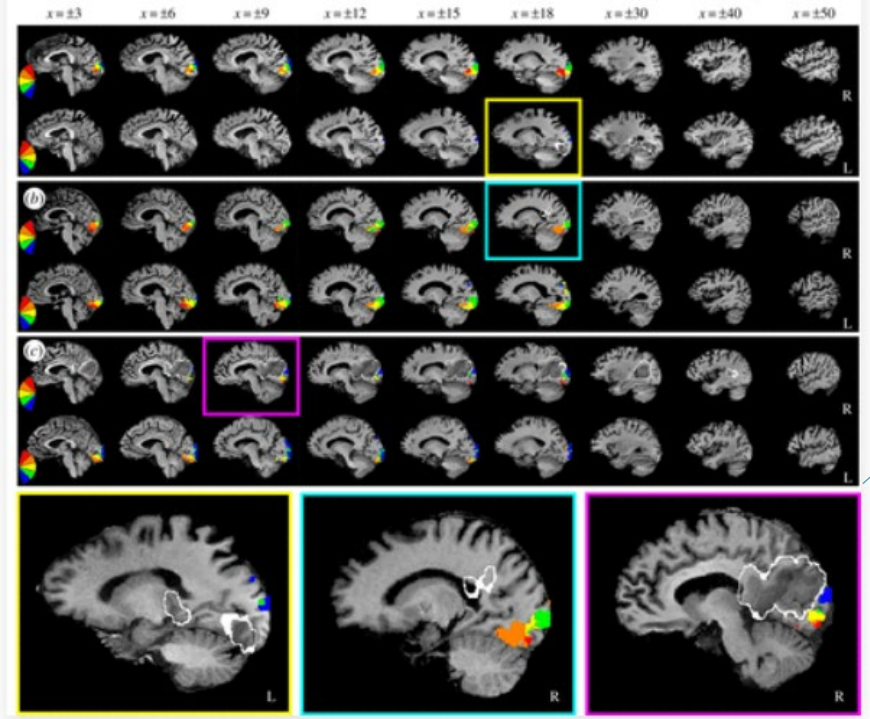


Figure 1. Overview of key measures. (*a*) Example measures from participant 5 collected at the final time point. Winner map of fMRI activity to flickering checkerboard wedges (stimulus example shows random order, lesion outlined from clinical T2 FLAIR or diffusion-weighted image *DWI shown in white; left panel), GCC thickness averaged over both eyes

ParticipantID	fMRI	Behavior	wedge	Patient_Age	TimePoint	deltaTofscan	nVoxTC_cont	deltaTofOCT	sector	MacularT	deltaTofHum	sensitivity	total_dev
1	365	86	1	55	2	NaN	NaN	NaN	5	NaN	63	20.17	-9.5
1	365	86	2	55	2	NaN	NaN	NaN	4	NaN	63	25.5	-5.1666667
1	365	86	3	55	2	NaN	NaN	NaN	3	NaN	63	26.17	-3.3333333
1	365	86	4	55	2	NaN	NaN	NaN	2	NaN	63	28.67	-2.5
1	365	86	5	55	2	NaN	NaN	NaN	1	NaN	63	27.5	-3.6666667
1	365	86	6	55	2	NaN	NaN	NaN	12	NaN	63	26	4.8333333

What if:

- Journal articles could link to repository data sets
- Metadata were computable so that a search for similar datasets was possible
- Analysis tools were linked to datasets, via Github, Bioconductor, Galaxy or other….

Importance of Data in Publications, Repositories & Research

- Incorporate data management in research plans
- Openly accessible and computable metadata, minimal metadata for open access sharing
- Data Quality Pipelines within repositories
- Collaborations between journals and repositories-at time of submission of articles
- Cite the data generator, the repository & acknowledge data users

Strategic Plan for Data Science: Goals and Objectives

Data Infrastructure

- Optimize data storage and security
- Connect NIH data systems

Modernized Data Ecosystem

- Modernize data repository ecosystems
- Support storage and sharing of individual datasets
- Better integrate clinical and observational data into biomedical data science

Data Management, Analytics, and Tools

- Support useful, generalizable, and accessible tools
- Broaden utility of, and access to, specialized tools
- Improve discovery and 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 Data and Technology Advancement (DATA) National Service Scholar Program

- One- or two-year national service program with high-impact NIH projects
- Seeking industry data and computer scientists, experts from related fields
- Expecting 5+ fellows in first cohort starting in summer 2020
- Submit CV and cover letter including vision statement and projects of interest to datascience@nih.gov.
- Eligibility: doctoral degree (required) and industry experience (strongly preferred)
- Women and individuals from underrepresented groups are encouraged to apply.

Office of Data Science Strategy

- Provide leadership and catalyze trans-NIH activities to support the NIH strategic plan for data science.
- Develop and implement NIH's vision for a **modernized** and **integrated** biomedical data ecosystem.
- Enable a diverse and talented data science workforce.
- In coordination with the CIO, build strategic partnerships for advanced technologies and methods.

Stay Connected

@NIHDataScience

/NIH.DataScience

www.datascience.nih.gov