Open & FAIR Data @ NIH

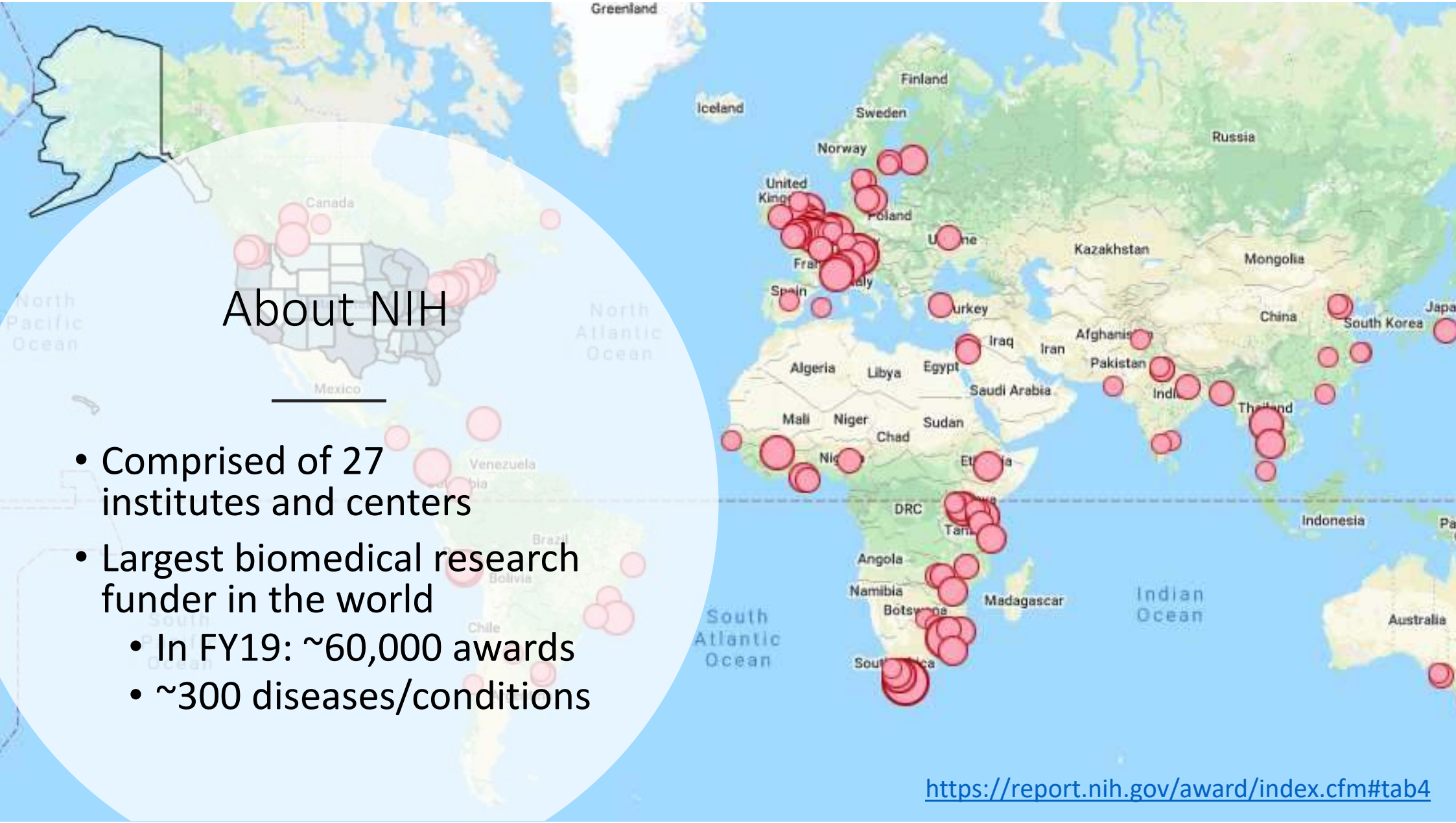
Perspectives on data sharing from a funder (& funder repository)

Kathryn Funk, MLIS, Program Manager, PubMed Central

US National Library of Medicine, National Institutes of Health

About NIH

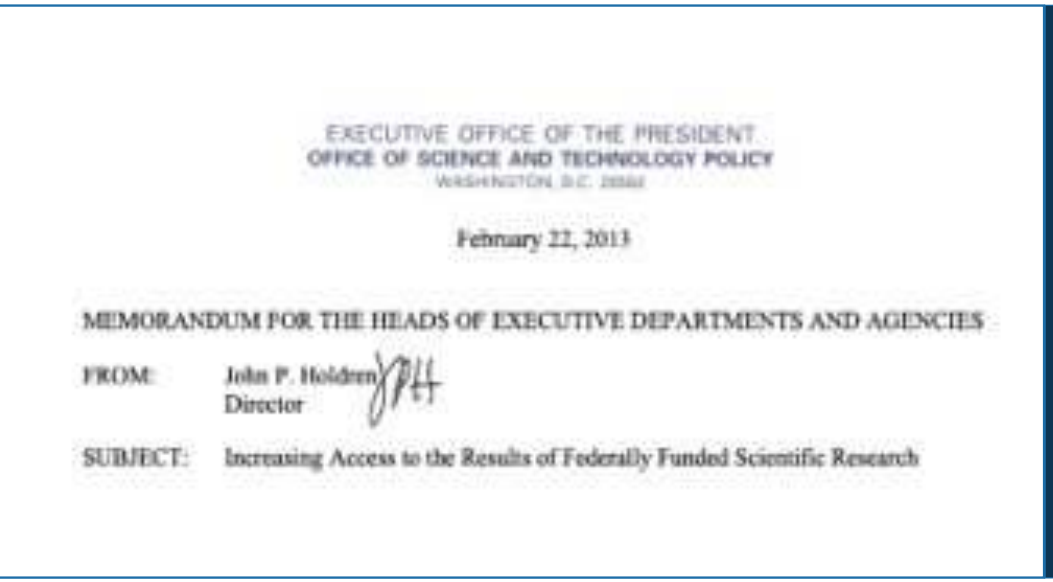
- Comprised of 27 institutes and centers
- Largest biomedical research funder in the world
  - In FY19: ~60,000 awards
  - ~300 diseases/conditions



https://report.nih.gov/award/index.cfm#tab4

NIH has a longstanding commitment to making research results and accomplishments available to the public.

Why share data? The public funder perspective



EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

"To achieve the Administration's commitment to increase access to federally funded published research and digital scientific data, Federal agencies investing in research and development must have clear and coordinated policies for increasing such access."

The NIH Public Access Policy

Centralized database with unique PID

Publicly accessible within 12 months or less

Machine-readable XML format

Text Mining Collections to facilitate reuse

How do we do the same for data across NIH?

NIH DRAFT POLICY FOR DATA MANAGEMENT AND SHARING

Current Proposal

Scope.

All research, funded or conducted by NIH, that results in generation of scientific data

Requirements.

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

Compliance.

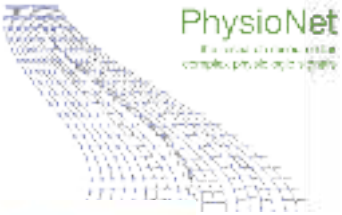
Failure to comply with the approved Plan may affect future NIH funding decisions

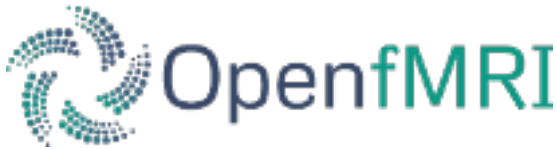
1. Where to share data

2. How to integrate data into the publication record

NIH Supports Many Repositories for Biomedical Data Sharing

NIH strongly encourages
Open-Access Data Sharing Repositories
as a first choice.





NIH Strategic Plan for Data Science (2018)

Or, how to establish infrastructure for a modernized, integrated, FAIR biomedical data ecosystem.

Modernize data repository ecosystem

Support storage and sharing of individual datasets

https://www.federalregister.gov/documents/2020/01/17/2020-00689/request-for-public-comment-on-draft-desirable-characteristics-of-repositories-for-managing-and

I. Desirable Characteristics for All Data Repositories

A. Persistent Unique Identifiers

B. Long-term sustainability

C. Metadata

D. Curation & Quality Assurance

E. Open access

F. Easy to Access and Reuse

G. Track Reuse

H. Secure

I. Privacy

J. Common Format

K. Provenance

NIH Data Repository Ecosystem: Domain-Specific Repositories

Findable, Accessible < **Interoperable, Reusable**

NIH Data Repository Ecosystem: Generalist Repositories

Findable, Accessible > Interoperable, Retrievable

NIH Figshare Instance: A Data Sharing Resource

https://nih.figshare.com/f/faq

NIH Figshare

SHARE	DISCOVER	CITE
• Self-publish any data type and file format • Link grant or project identifier • Bulk-upload with API • 100GB storage per user	• Access open, de-identified data • Search and filter on metadata • Indexed in Google Dataset Search • Track usage metrics	• Get a DOI • Attach a license • Ability to embargo • Secure storage on FedRAMP AWS S3

Data ID Services

The Department of Energy (DOE) Office of Scientific and Technical Information (OSTI) offers two services for registering datasets to help increase access to scientific research data – the DOE Data ID Service and the Interagency Data ID Service (IAD).

Principles and Guidelines for Reporting Preclinical Research (2014)

Or, how to enhance rigor and reproducibility of NIH research.

All datasets on which the conclusions of the paper rely must be made available upon request

Recommend deposition of datasets in public repositories

Encourage presentation of all other data values in machine readable format in the paper or its supplementary information.

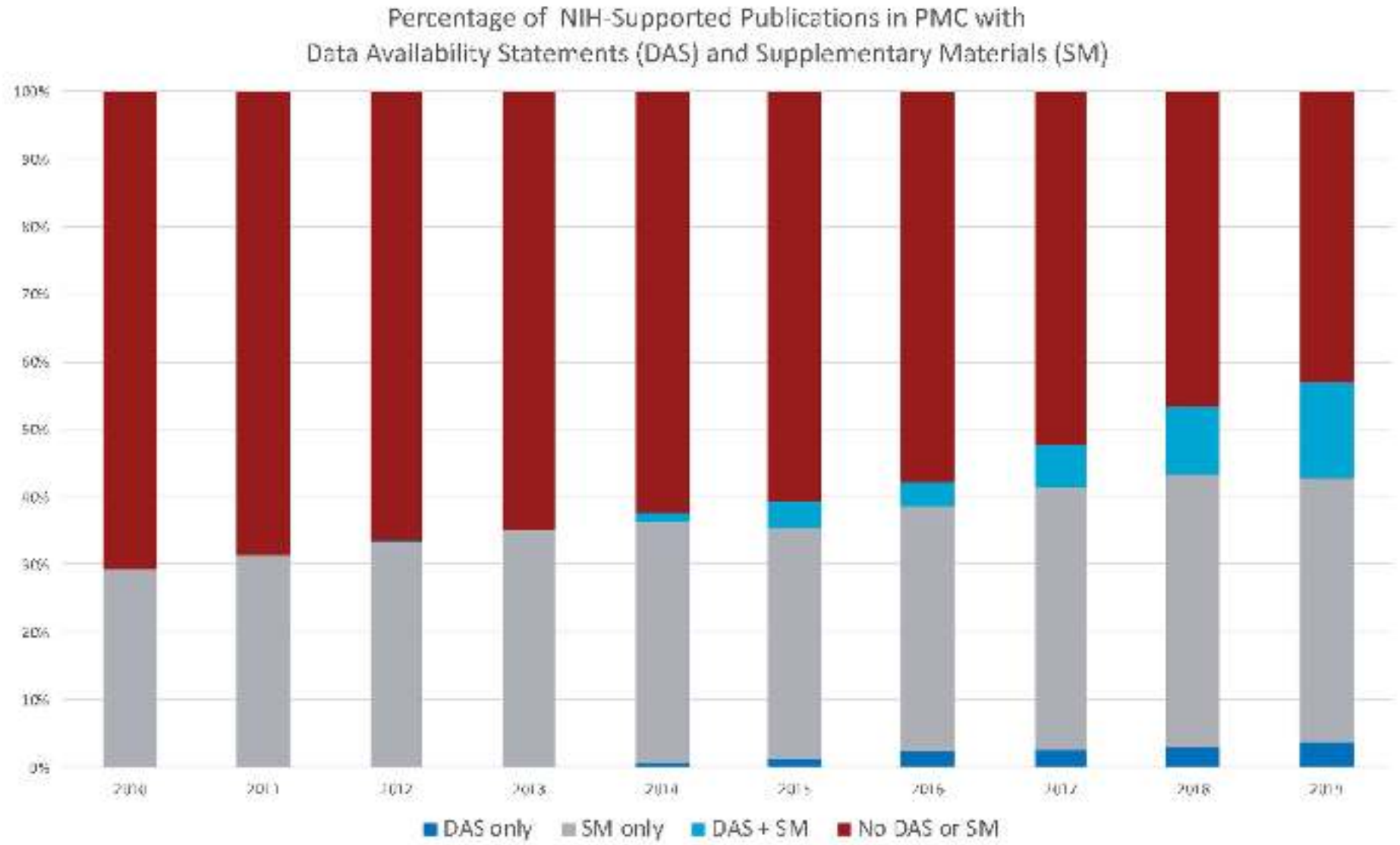
Encourage sharing of software and require a statement in the manuscript describing how it can be obtained

Associated Data Box

Released November 2018

Exposes any data citations, supplementary materials, or data availability statements in an article

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351104/

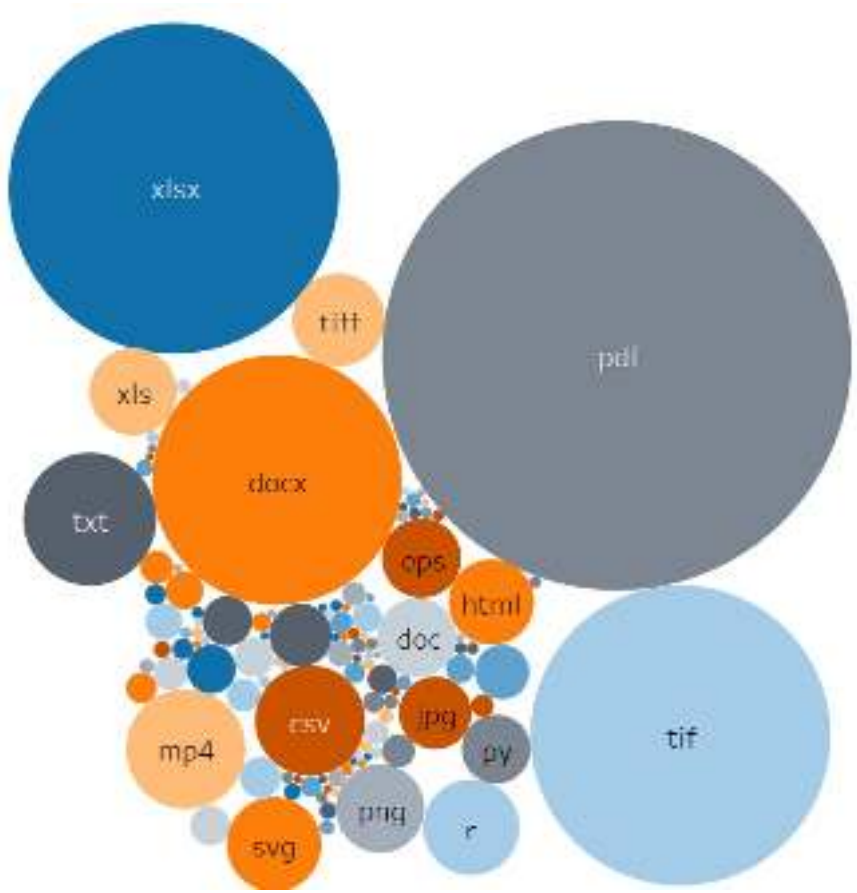


NIH U.S. National Library of Medicine
National Center for Biotechnology Information

What's Up With The Supp?

An Analysis of Supplementary Materials in PubMed Central

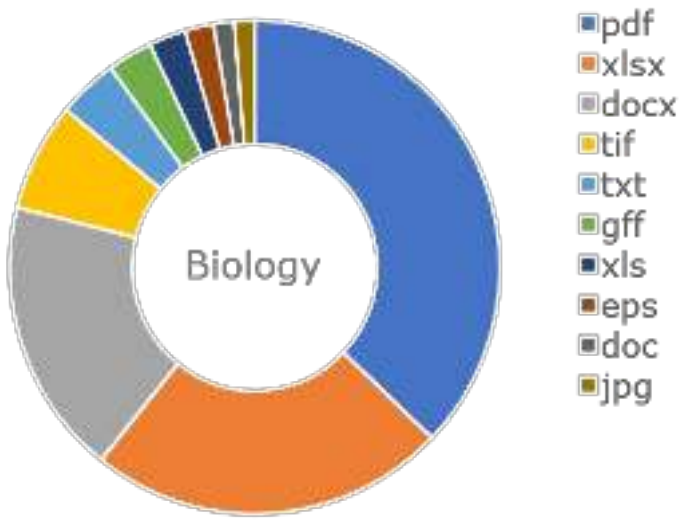
Associate Fellows 2019-2020 Project

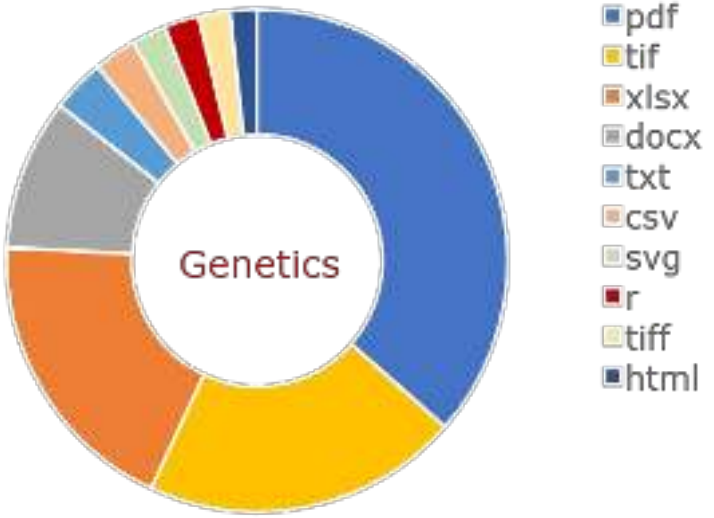


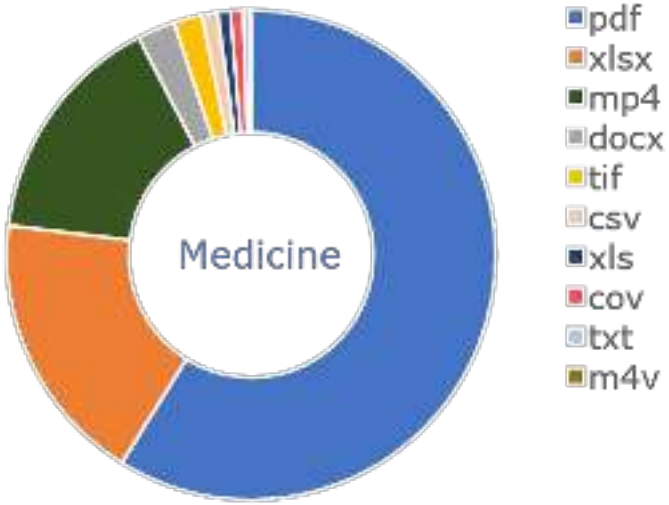
- Research Questions
  - What types of supplemental materials are found in PMC?
  - How do supplemental materials differ across subjects?
- Our Dataset
  - 20 Journals in 4 Broad Subjects
    - Biology, Genetics, Medicine, Neoplasms
  - 1,466 articles
  - 8,765 supplemental files
  - 100+ different file formats

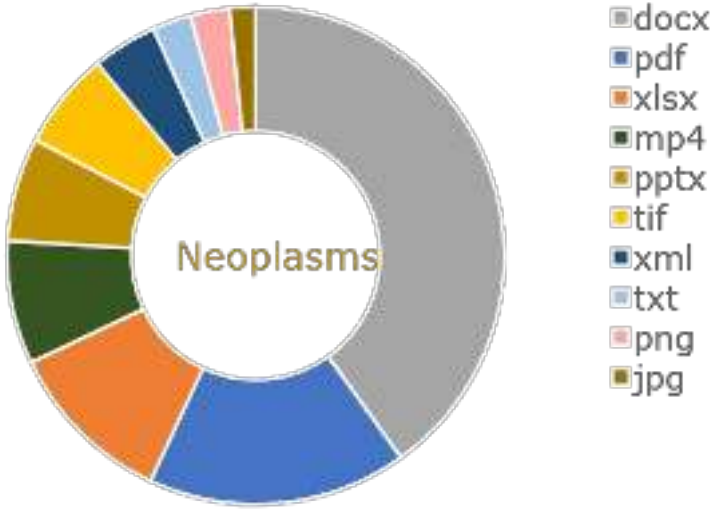
This project was supported in part by an appointment to the Science Education Programs at National Institutes of Health (NIH), administered by ORAU through the U.S. Department of Energy Oak Ridge Institute for Science and Education.

Top 10 File Formats by Subject

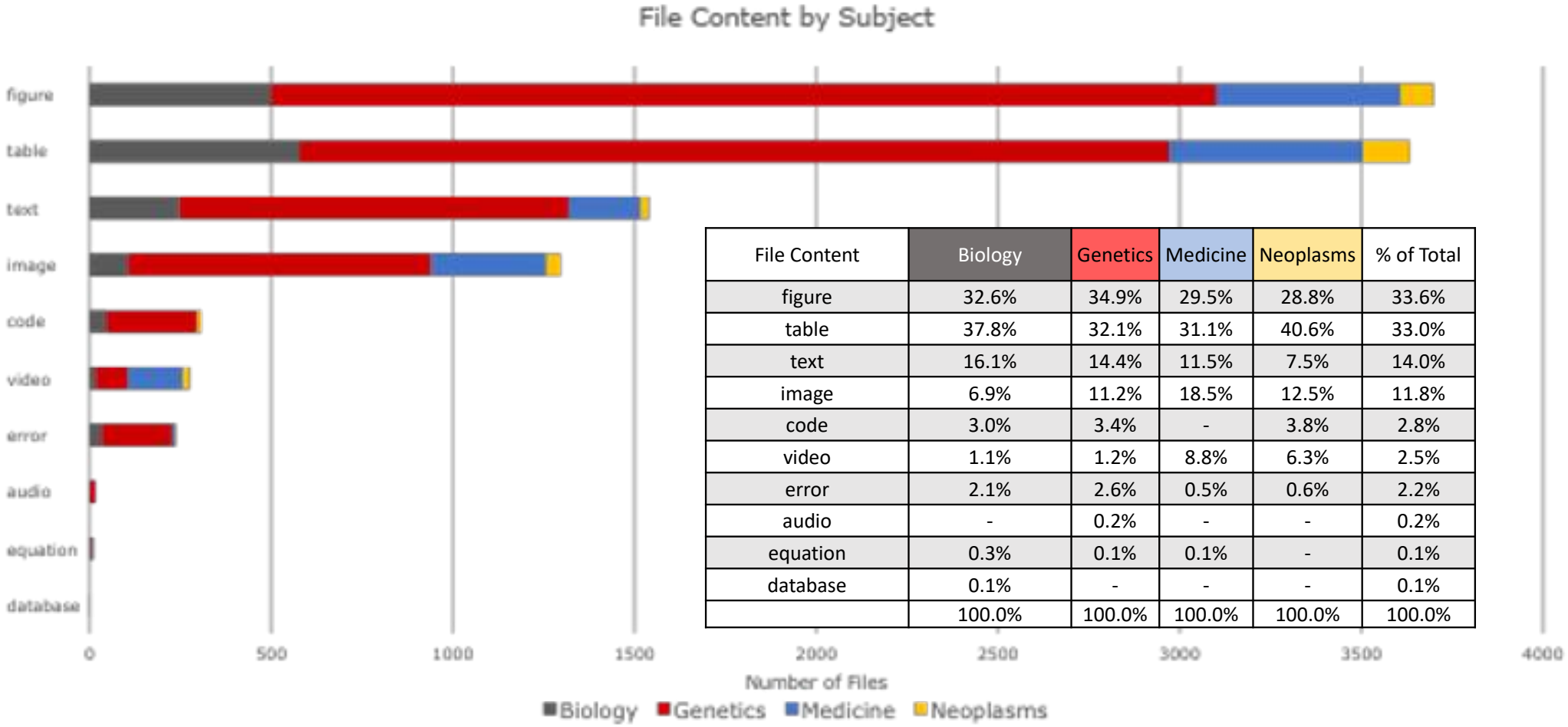








File Content by Subject



File Content	Biology	Genetics	Medicine	Neoplasms	% of Total
figure	32.6%	34.9%	29.5%	28.8%	33.6%
table	37.8%	32.1%	31.1%	40.6%	33.0%
text	16.1%	14.4%	11.5%	7.5%	14.0%
image	6.9%	11.2%	18.5%	12.5%	11.8%
code	3.0%	3.4%	-	3.8%	2.8%
video	1.1%	1.2%	8.8%	6.3%	2.5%
error	2.1%	2.6%	0.5%	0.6%	2.2%
audio	-	0.2%	-	-	0.2%
equation	0.3%	0.1%	0.1%	-	0.1%
database	0.1%	-	-	-	0.1%
	100.0%	100.0%	100.0%	100.0%	100.0%

Increased use of data citations in standardized format.

More meaningful Data Availability Statements.

Thank you!

kathryn.funk@nih.gov