Strategic Approaches to Data Science & Open Science: Research Data Management

Research Data Management Symposium - December 5, 2019

Michael F. Huerta, PhD
Director, Office of Strategic Initiatives
Associate Director, National Library of Medicine, NIH

All of Us RESEARCH PROGRAM

- Goal is to enroll 1 million participants in a research program
- Participant-engaged, data-driven enterprise
- Collecting data about behavior, genetics, physiology, environmental exposures, social factors, and much more

Opportunities

New ways to measure risk for diseases based on environmental exposures, genetic factors & interactions

Correlate activity, physiological, genetic, & environmental measures with **health outcomes**

Discover **biomarkers** that signal increased or decreased risk of developing diseases

Develop **new disease classifications** & relationships

Identify the causes of **individual differences in response** to commonly used drugs & other treatments

Challenges

How to coordinate large scale **data collection** across many very different sites?

How to **deliver data** to the right users, at the right time, in the right way?

How to protect and preserve the **privacy** of participants, while maximizing research opportunities?

How to **synthesize** large volumes of diverse data types to power discovery?

How to help diverse publics **harness the power** of their personal health data?

Challenges

How to coordinate large scale **data collection** across many very different sites?

How to **deliver data** to the right users, at the right time, in the right way?

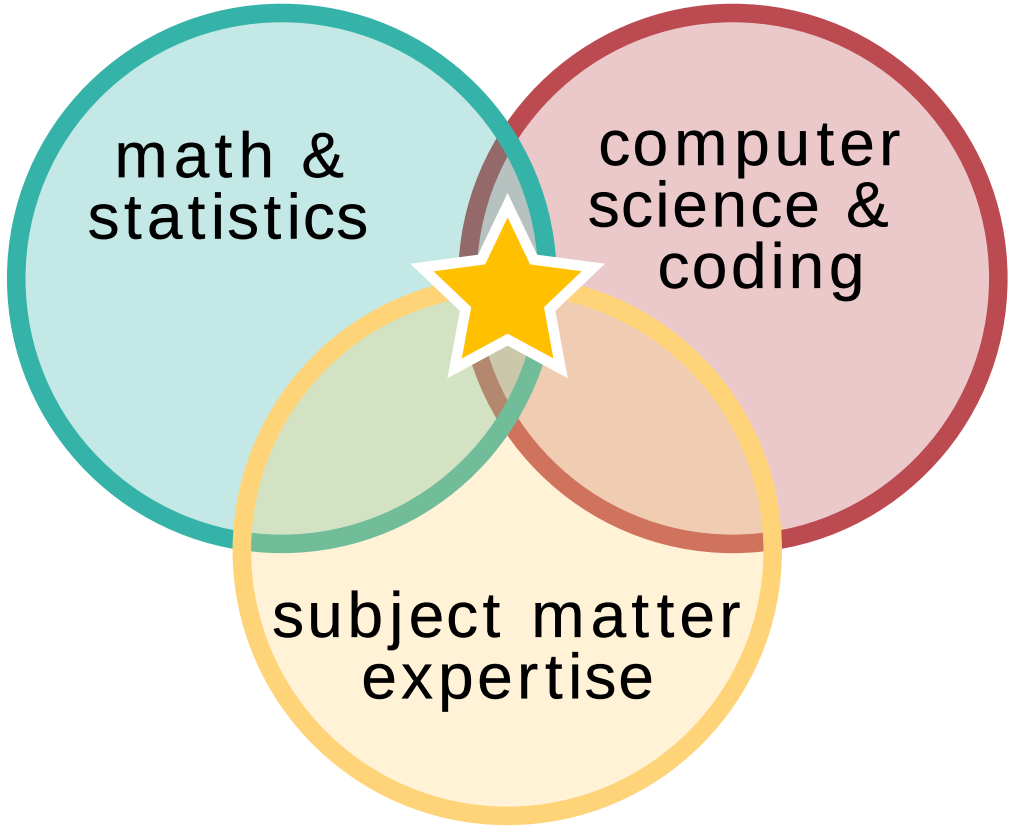
How to protect and preserve the **privacy** of participants, while maximizing research opportunities?

How to **synthesize** large volumes of diverse data types to power discovery?

How to help diverse publics **harness the power** of their personal health data?

By working at the intersection of data science & open science

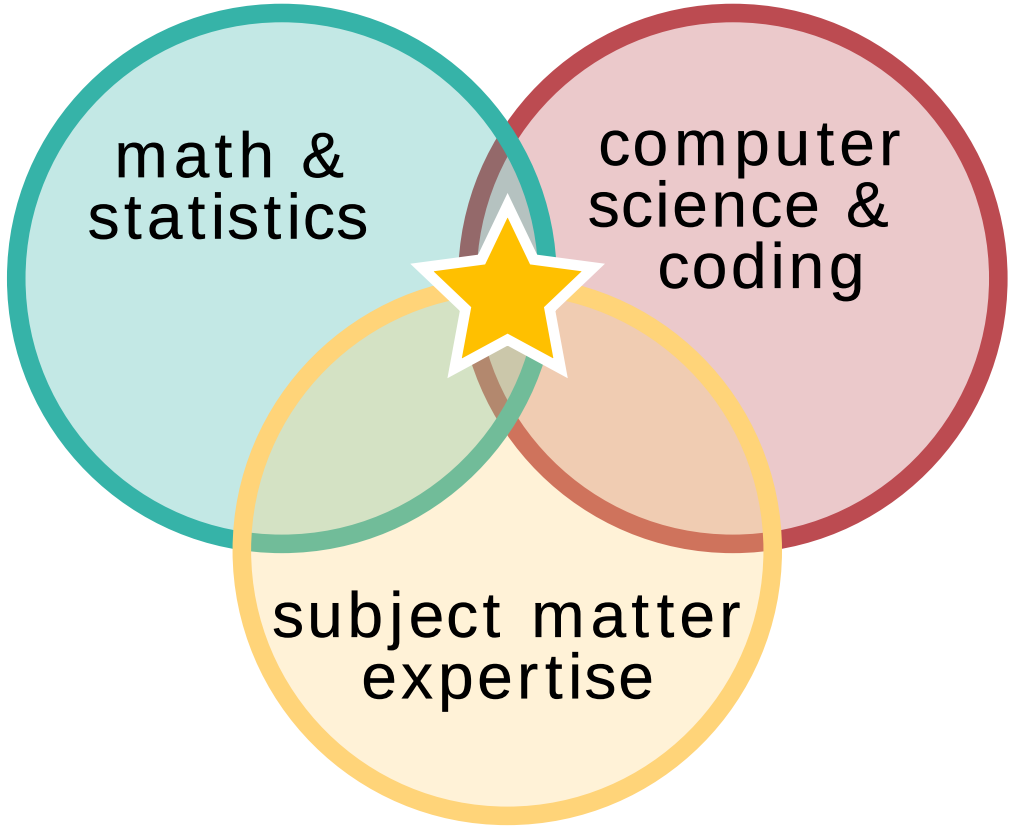
Data Science



Data Science is a **scientific & methodologic approach** to understanding data

(Just as molecular biology is a scientific & methodologic approach to understanding disease)

Data Science



Data Science is a **scientific & methodologic approach** to understanding data

(Just as molecular biology is a scientific & methodologic approach to understanding disease)

New tools for new insights

“New directions in science are launched by new tools much more often than by new concepts.”

Freeman Dyson
Imagined Worlds
(1997)
Harvard University Press

Open Science

Open Science is a **new paradigm**,
a different way of doing science
in which the **products and processes of research** (e.g., datasets) are
broadly **available & usable**

Open Science

Open Science is a **new paradigm**,
a different way of doing science
in which the **products and processes of research** (e.g., datasets) are broadly **available & usable**

Open science is facilitated when **research data management** abides by the **FAIR principles**

Findable

Accessible

Interoperable

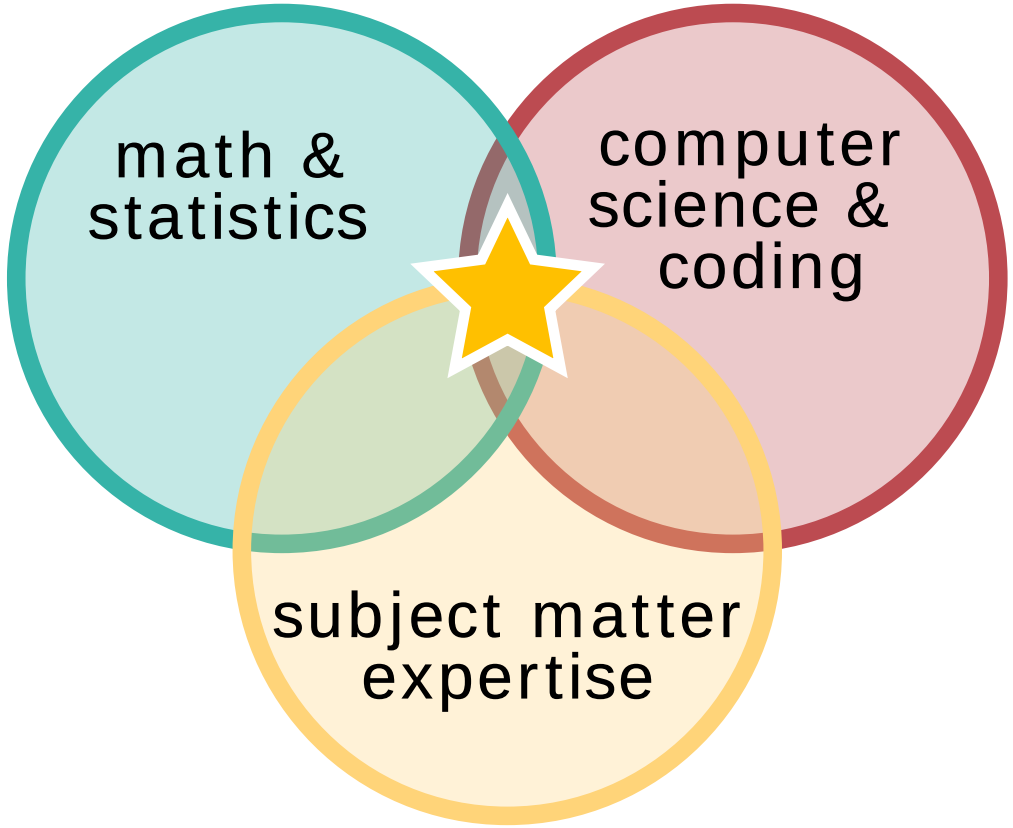
Reusable

"...paradigm changes do cause scientists to see the world of their research engagements differently."

Thomas Kuhn
The Structure of Scientific Revolutions
(1962)
University of Chicago Press

DS + OS = new tools and new paradigm

Data Science



Open Science

Findable

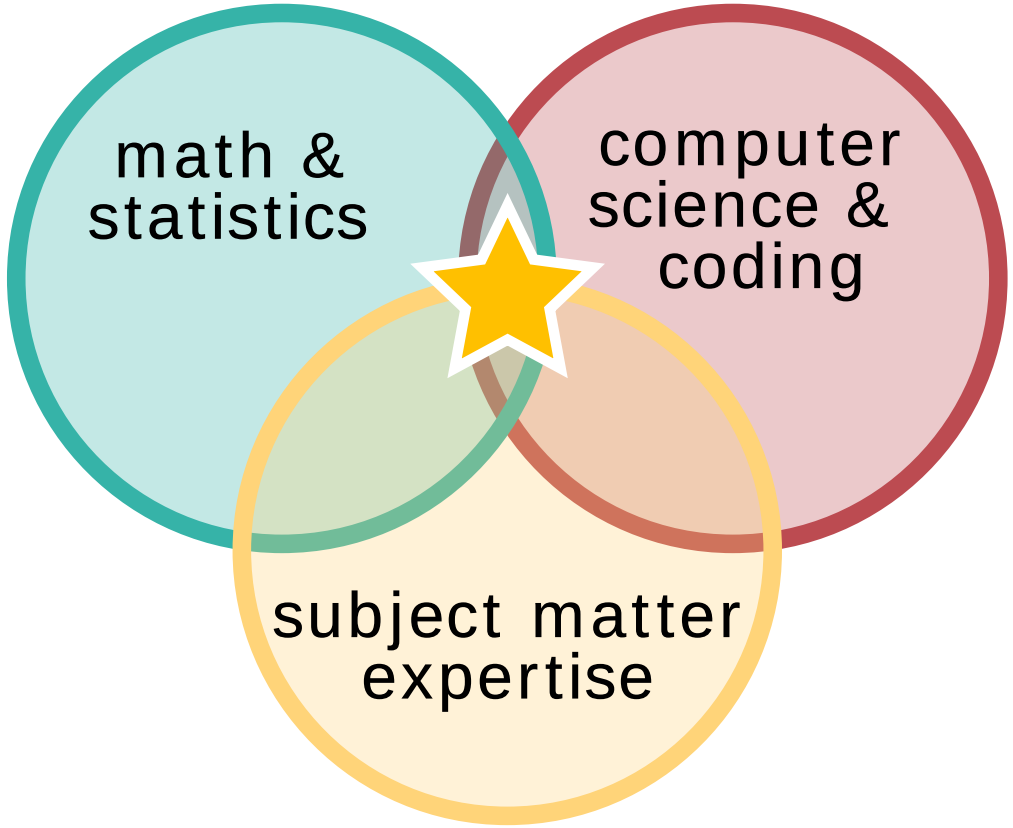
Accessible

Interoperable

Reusable

DS & OS are very powerful when paired

Data Science



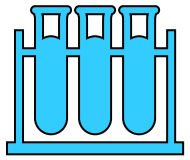
Open Science

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

The ***National Library of Medicine***
lives at the intersection of
data science *&* ***open science***

National Library of Medicine

National Library of Medicine



- An Institute of the NIH (1968)
  - Lead, conduct, and support research and training in biomedical:
    - Information science
    - Informatics
    - Data science

National Library of Medicine

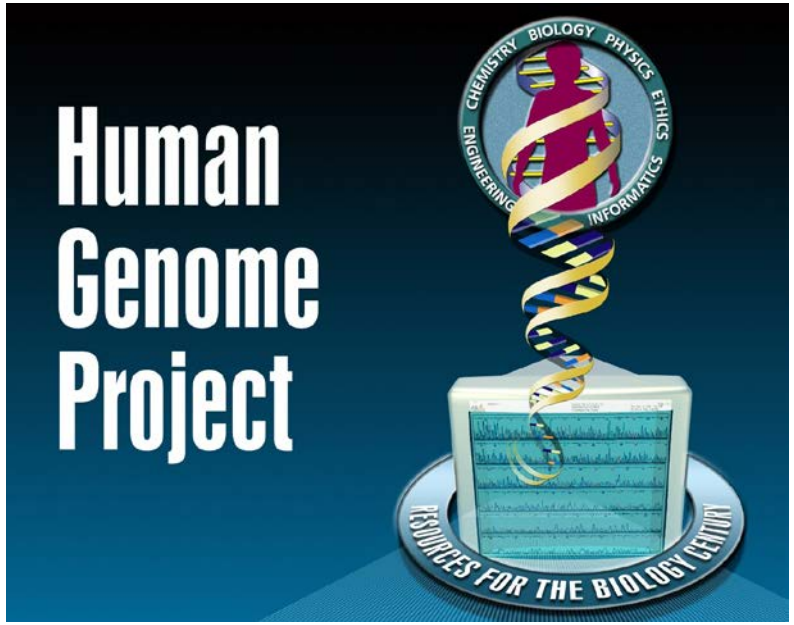
- An Institute of the NIH (1968)
  - Lead, conduct, and support research and training in biomedical:
    - Information science
    - Informatics
    - Data science
- The world's largest biomedical library (1836)
  - Create & host major resources, tools, & services for literature, data, standards, & more
    - Send > 115 terabytes of data to > 5 million users daily
    - Receive > 15 terabytes of data from > 3,000 users daily

National Library of Medicine

- An Institute of the NIH (1968)
  - Lead, conduct, and support research and training in biomedical:
    - Information science
    - Informatics
    - Data science
- The world's largest biomedical library (1836)
  - Create & host major resources, tools, & services for literature, data, standards, & more
    - Send > 115 terabytes of data to > 5 million users daily
    - Receive > 15 terabytes of data from > 3,000 users daily
  - Facilitate open science & scholarship by making digital research objects:
    - Findable, Accessible, Interoperable, & Reusable (FAIR)
    - As well as Attributable & Sustainable

NIH has paired DS & OS to tackle significant biomedical research questions through large scale, data-centric, and open initiatives

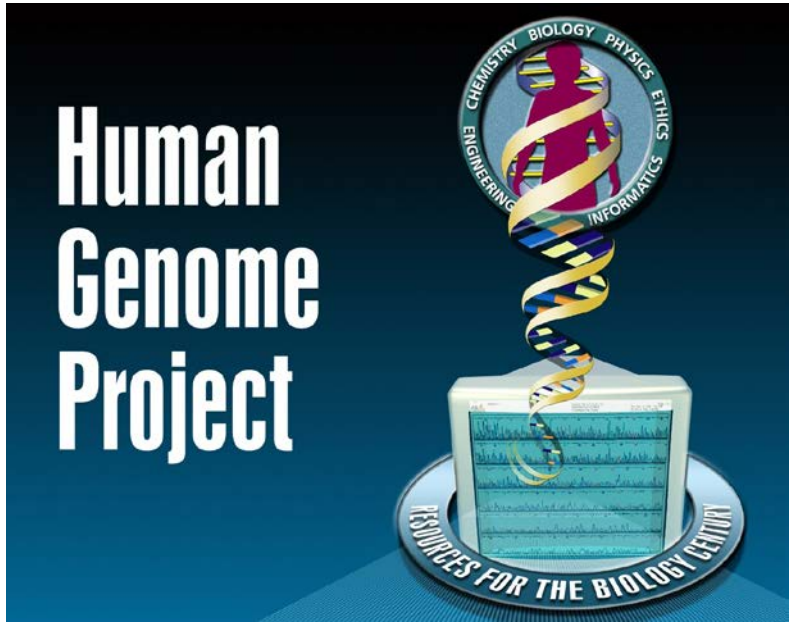
NIH has paired DS & OS to tackle significant biomedical research questions through large scale, data-centric, and open initiatives



NIH
The Common Fund

GTE

NIH HUMAN MICROBIOME PROJECT



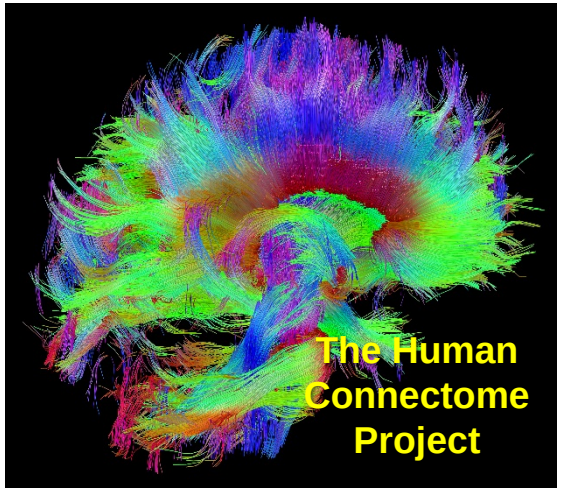
FlyBase

MGI

ZFIN
The Zebrafish Model Organism Database

SGD Saccharomyces GENOME DATABASE

WormBase

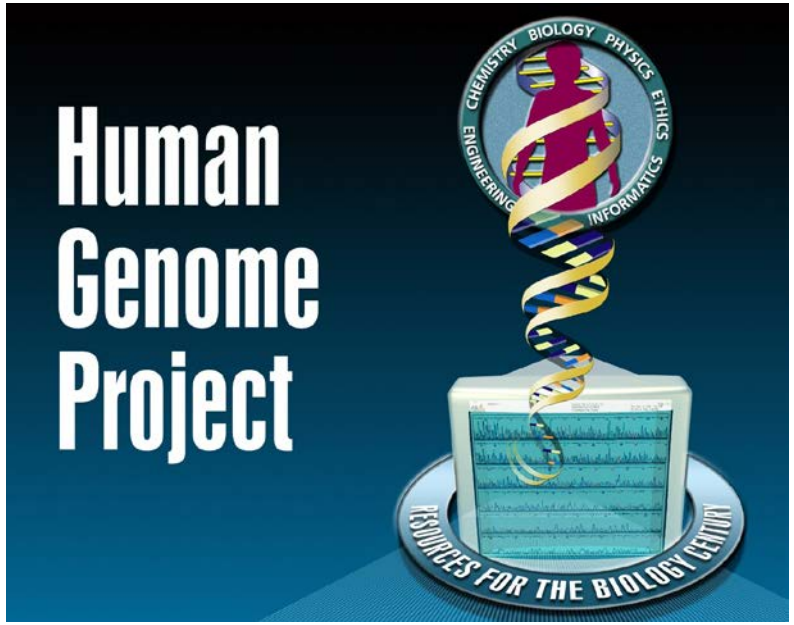


NIH
The Common Fund

GTE

ECHO
Environmental Influences on Child Health Outcomes

HMP
NIH HUMAN MICROBIOME PROJECT



ADNI
Alzheimer's Disease Neuroimaging Initiative

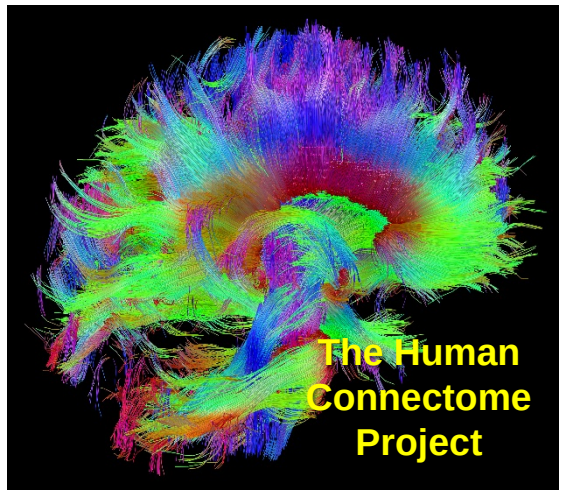
FlyBase

MGI

ZFIN
The Zebrafish Model Organism Database

SGD Saccharomyces GENOME DATABASE

WormBase



NIH
The Common Fund

GTE

ECHO
Environmental Influences on Child Health Outcomes

HMP
NIH HUMAN MICROBIOME PROJECT

CANCER MOONSHOT

Human Genome Project

ADNI
Alzheimer's Disease Neuroimaging Initiative

FlyBase

MGI

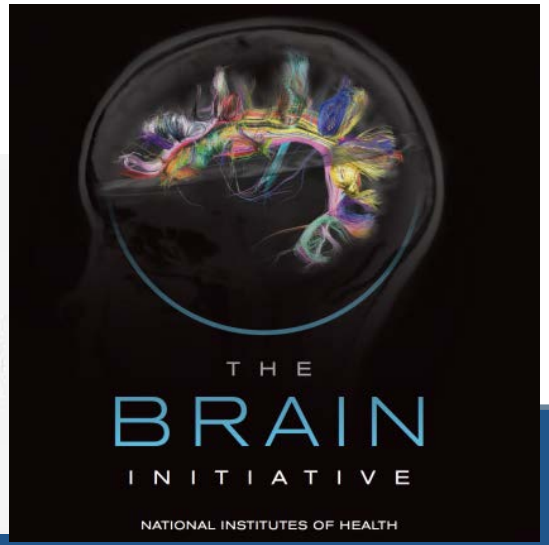
ZFIN
The Zebrafish Model Organism Database

SGD Saccharomyces Genome Database

WormBase

Adolescent Brain Cognitive Development
Teen Brains. Today's Science. Brighter Future.

All of Us℠
THE FUTURE OF HEALTH BEGINS WITH YOU
The Precision Medicine Initiative®

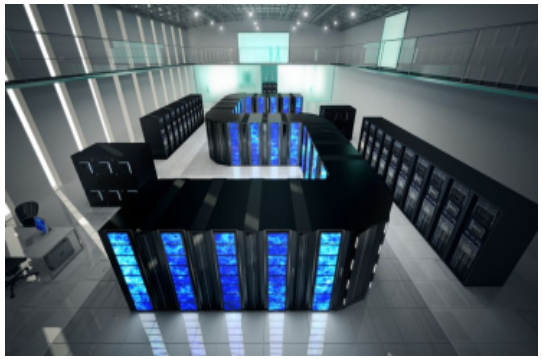


NIH research to be even MORE data-centric & open

Societal Expectations

NIH research to be even MORE data-centric & open

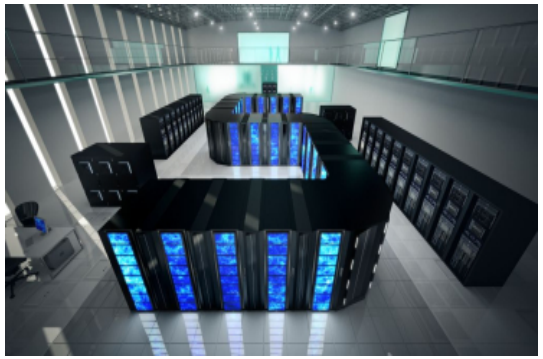
Societal Expectations



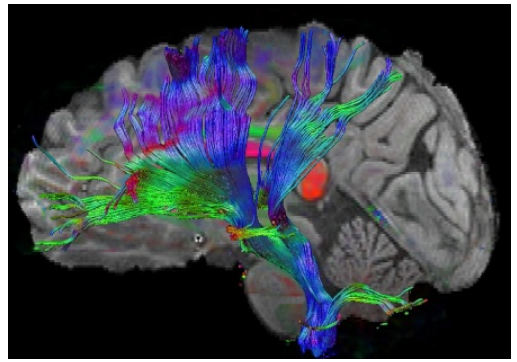
NIH research to be even MORE data-centric & open

Technical Capabilities

Societal Expectations



NIH research to be even MORE data-centric & open



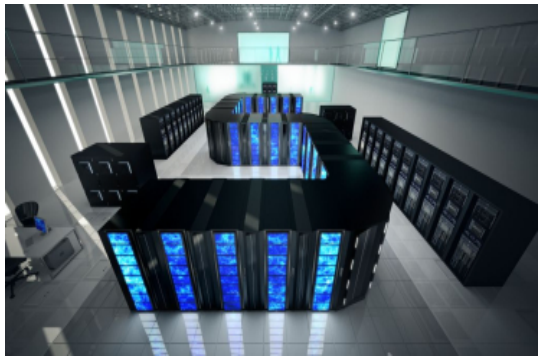
Technical Capabilities

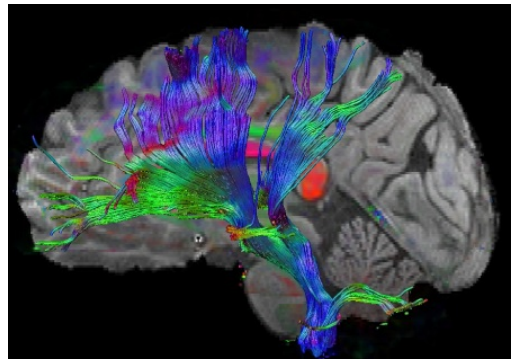
Scientific Opportunities

Societal Expectations

Laws & Policy Directives

NIH research to be even MORE data-centric & open





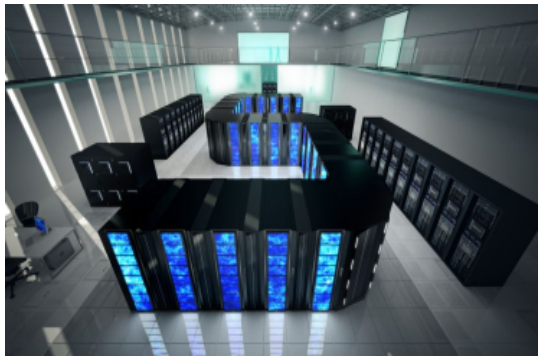
Technical Capabilities

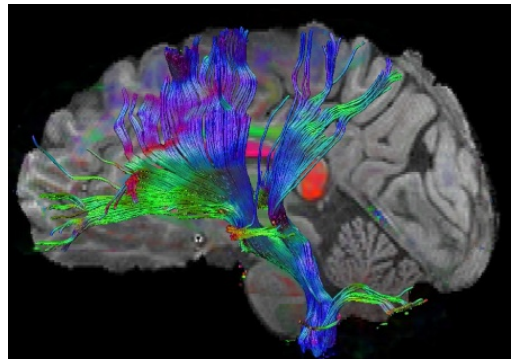
Scientific Opportunities

Societal Expectations

Laws & Policy Directives

NIH research to be even MORE data-centric & open





Technical Capabilities

Scientific Opportunities

Opening Science by Memo

- **February 2013** - OSTP memo to federal agencies to increase public access to research results
  - Publications
  - Data

- **February 2015** - NIH plan posted

- **Today** - Implementation is underway

Opening Science by Memo

- **February 2013** - OSTP memo to federal agencies to increase public access to research results
  - Publications
  - Data

- **February 2015** - NIH plan posted

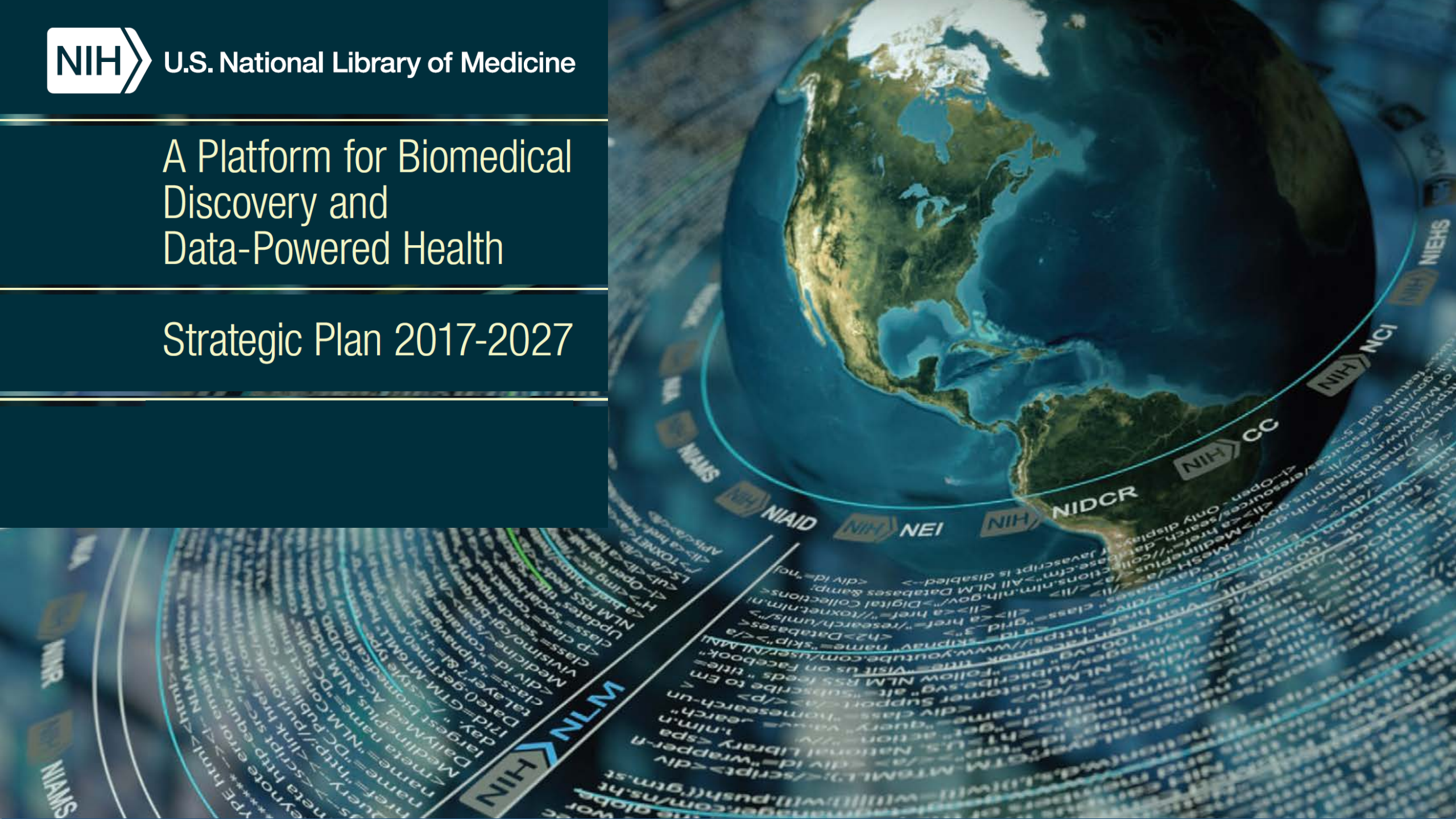
- **Today** - Implementation is underway

Requires strategic & coordinated approaches

U.S. National Library of Medicine

A Platform for Biomedical Discovery and Data-Powered Health

Strategic Plan 2017-2027



Innovate, build, & sustain an open digital ecosystem for health info, science, & scholarship

Innovate, build, & sustain an open digital ecosystem for health info, science, & scholarship

Optimize user experience with, and use of, NLM digital resources

Innovate, build, & sustain an open digital ecosystem for health info, science, & scholarship

Optimize user experience with, and use of, NLM digital resources

Assure a data-savvy biomedical workforce and a data-ready public

NIH Strategic Plan for Data Science

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**

Successful DS & OS Strategies Depend on Good RDM

Policy implementation for open science

Enhancing the repository ecosystem

Connecting repositories through metadata

Successful DS & OS Strategies Depend on Good RDM

Policy implementation for open science

Enhancing the repository ecosystem

Connecting repositories through metadata

Open Science Begins With Sharing Data

Advances science & the public good in many ways

- Deeper **understanding** of the research & conclusions
- Enables **validation** of scientific results
- Allows **reanalysis** of data with different methods
- Increases **statistical power & corpus size** via data aggregation
- Allows **big data approaches** by combining heterogeneous data
- Facilitates **reuse** of hard-to-generate data
- Promotes research **collaboration**

- Good stewardship **increases return on taxpayer investment**
- Fosters **transparency and accountability**
- Maximizes research **participants' contributions**

NIH Draft Policy for Data Management & Sharing
(*seeking public comment*)

NIH Draft Policy for Data Management & Sharing
(*seeking public comment*)

- **Draft policy reflects the importance of:**
  - Data stewardship
  - Good research data management
  - FAIR principles
  - Flexibility in Plans
  - Research participants' autonomy & protection

Current DRAFT Policy for Data Management & Sharing

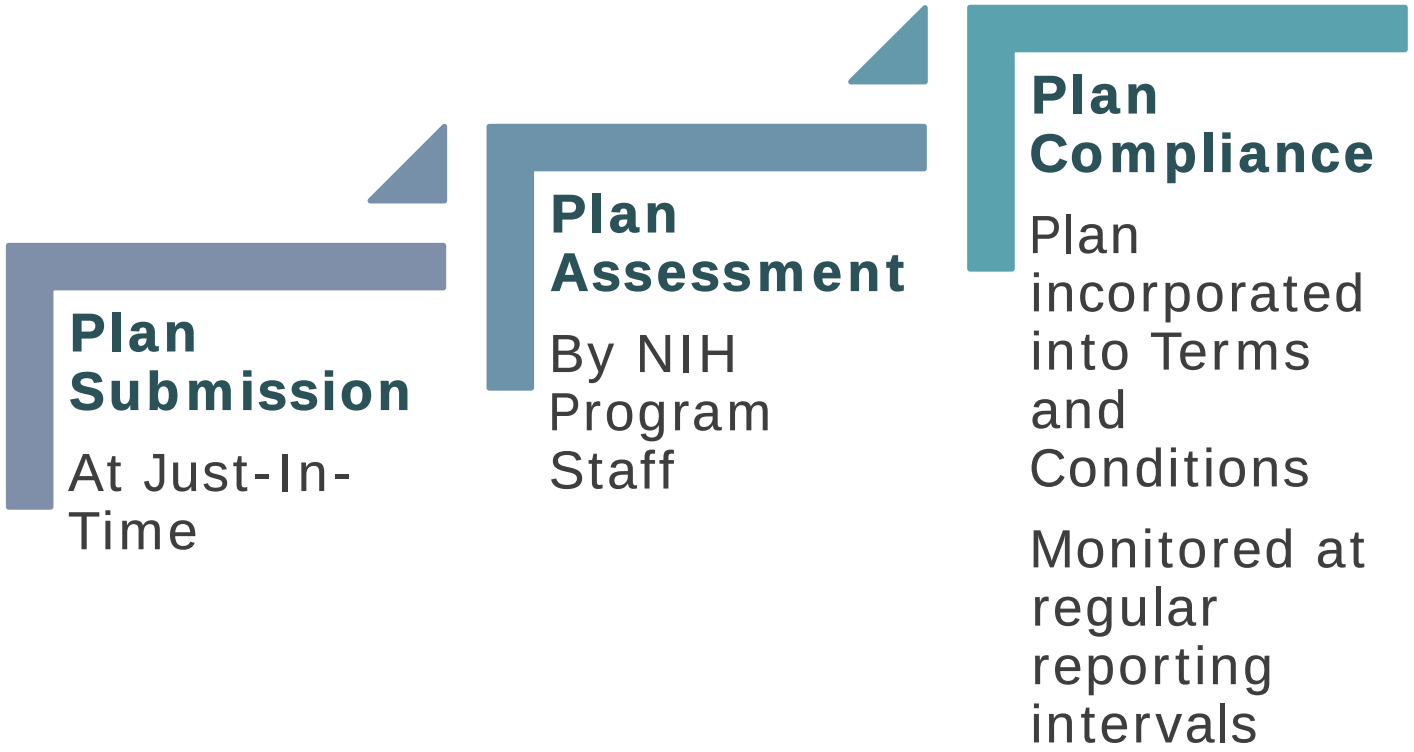
- **Scope**: Applies to **all research** funded or conducted by NIH that results in generation of scientific data
- **Requirements**: Submission of and compliance with a Data Management and Sharing **Plan outlining how scientific data will be managed and shared**, taking into account any potential restrictions or limitations
- **Compliance**: Failure to comply with the approved Plan may **affect future NIH funding** decisions
- **Paying for Data Management & Sharing**: Costs of data management & sharing activities are **allowable** budget requests

DRAFT Policy Guidance – Plan Elements

- **Proposed Plan Elements**
  - Data type (including metadata/documentation)
  - Related tools, software, and/or code
  - Standards
  - Data preservation, access, & their associated timelines
  - Data sharing agreements, licenses, & other use limitations
  - Oversight of data management

DRAFT Policy - Plan Flow

Extramural Grant Awards

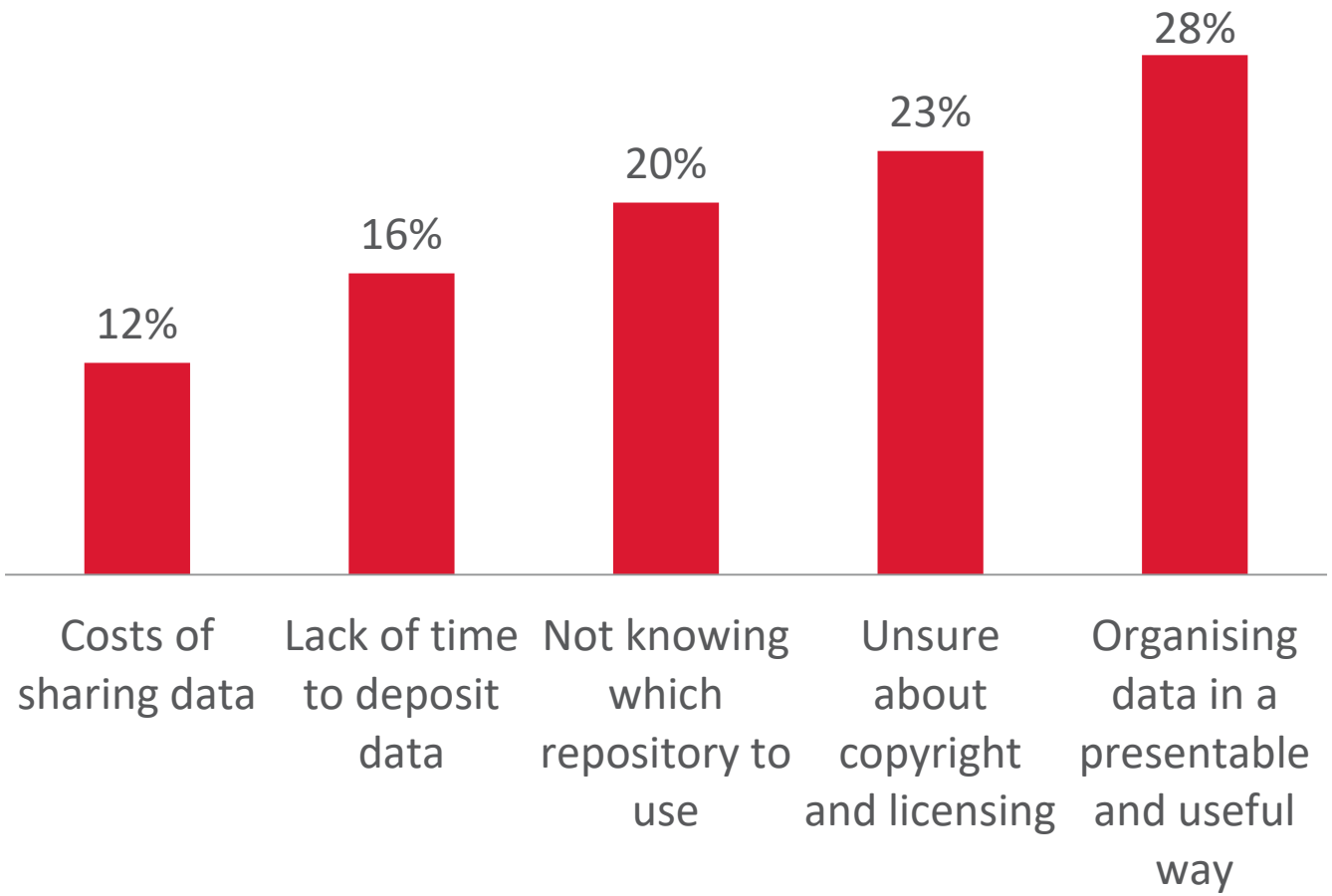


NIH is Seeking Public Comment on DRAFT Policy for Data Management & Sharing

- Comments are being accepted until **January 10, 2020**
- Will be used to inform a final policy

https://osp.od.nih.gov/scientific-sharing/nih-data-management-and-sharing-activities-related-to-public-access-and-open-science/

What problems do authors have in sharing datasets?

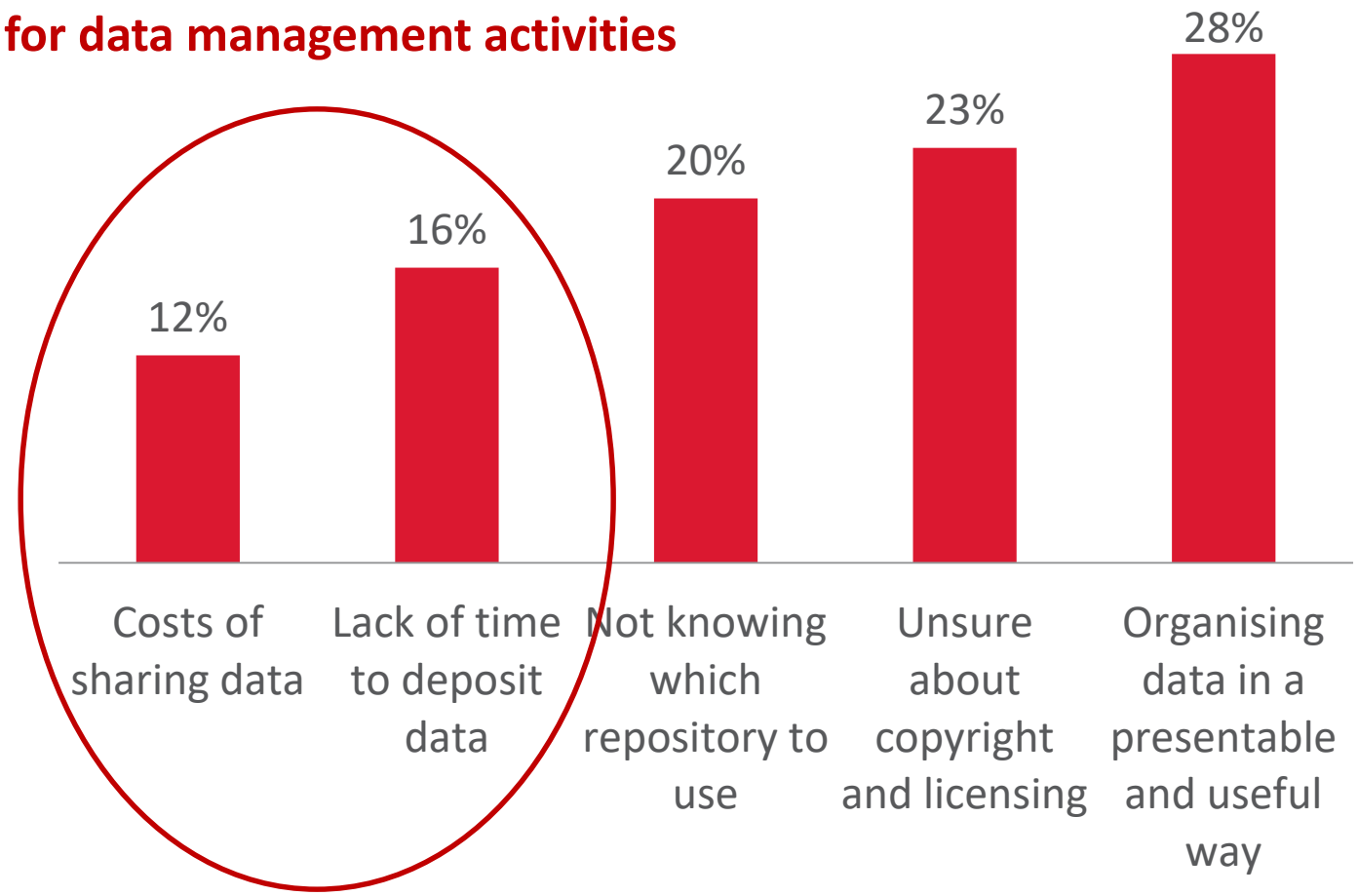


From a Springer Nature researcher survey. Total respondents: 7719

https://doi.org/10.6084/m9.figshare.5975011

What problems do authors have in sharing datasets?

NIH draft policy provides cost recovery for data management activities

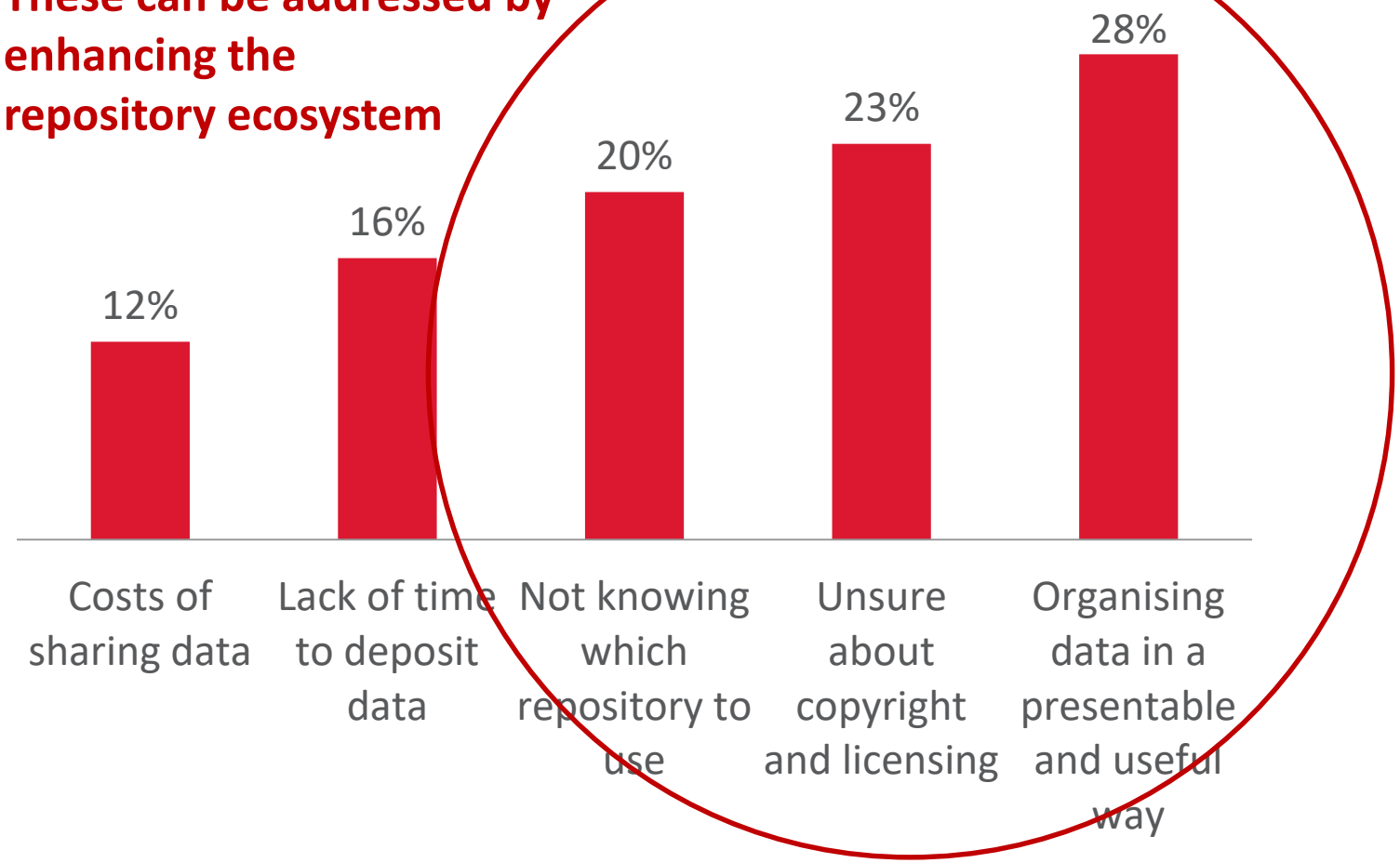


From a Springer Nature researcher survey. Total respondents: 7719

https://doi.org/10.6084/m9.figshare.5975011

What problems do authors have in sharing datasets?

These can be addressed by enhancing the repository ecosystem



From a Springer Nature researcher survey. Total respondents: 7719

https://doi.org/10.6084/m9.figshare.5975011

Successful DS & OS Strategies Depend on Good RDM

Policy implementation for open science

Enhancing the repository ecosystem

Connecting repositories through metadata

Sharing Publication-Related Data

NIH strongly encourages use of open domain-specific repositories as a first choice

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

BMIC List of >80 Domain-Specific Repositories

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

NIH Trans-NIH BioMedical Informatics Coordinating Committee (BMIC)

BMIC Home | CDE Resource Portal

ICO	Repository Name	Repository Description	Data Submission Policy	Access to Data	Current NIH funding support	Open data submission	Open data access	Open time frame for data deposit	Sustained support
NICHD	Xenbase	Xenbase is a Xenopus laevis and Xenopus tropicalis biology and genomics resource.	How to submit data to Xenbase	How to access Xenbase data	Yes	Yes	Yes	Yes	Yes
NHGRI	WormBase	WormBase is an international consortium of biologists and computer scientists dedicated to providing the research community with accurate, current, accessible information concerning the genetics, genomics and biology of C. elegans and related nematodes.	How to submit data to WormBase	How to access WormBase data	Yes	Yes	Yes	Yes	Yes
NIAID	Virus Pathogen Research (ViPR)	The Virus Pathogen Resource (ViPR) provides a searchable public repository of genomic, proteomic and other important research data for more than 500,000 strains of pathogenic viruses along with a suite of	How to submit data to ViPR	How to access ViPR data	Yes	Yes	Yes	Yes	Yes

Sharing Publication-Related Data

NIH strongly encourages use of open domain-specific repositories as a first choice

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

Options of scaled implementation for sharing datasets

Sharing Publication-Related Data

NIH strongly encourages use of open domain-specific 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 GB**

PubMed Central (PMC)

Datasets associated with PMC articles

Separate unique identifiers assigned to datasets & articles

Generalist / Institutional Repositories

CYVERSE

zenodo

DRYAD

figshare

The Dataverse® Project

MENDELEY DATA

Sharing Publication-Related Data

NIH strongly encourages use of open domain-specific 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 GB**

Terabytes

PubMed Central (PMC)

Datasets associated with PMC articles

Separate unique identifiers assigned to datasets & articles

Sharing Publication-Related Data

NIH strongly encourages use of open domain-specific 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 GB**

Terabytes

PubMed Central (PMC)

Datasets associated with PMC articles

Separate unique identifiers assigned to datasets & articles

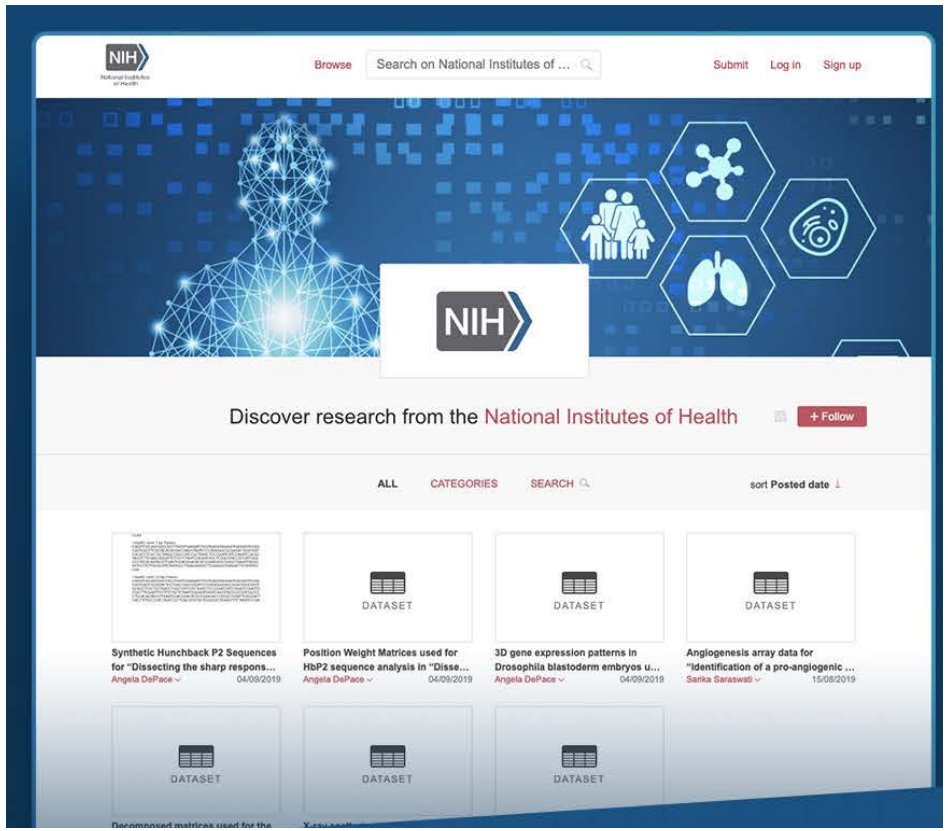
Generalist / Institutional Repositories

figshare pilot project 100 GB

Datasets assigned unique identifiers & linked to PubMed citations

NIH figshare Pilot

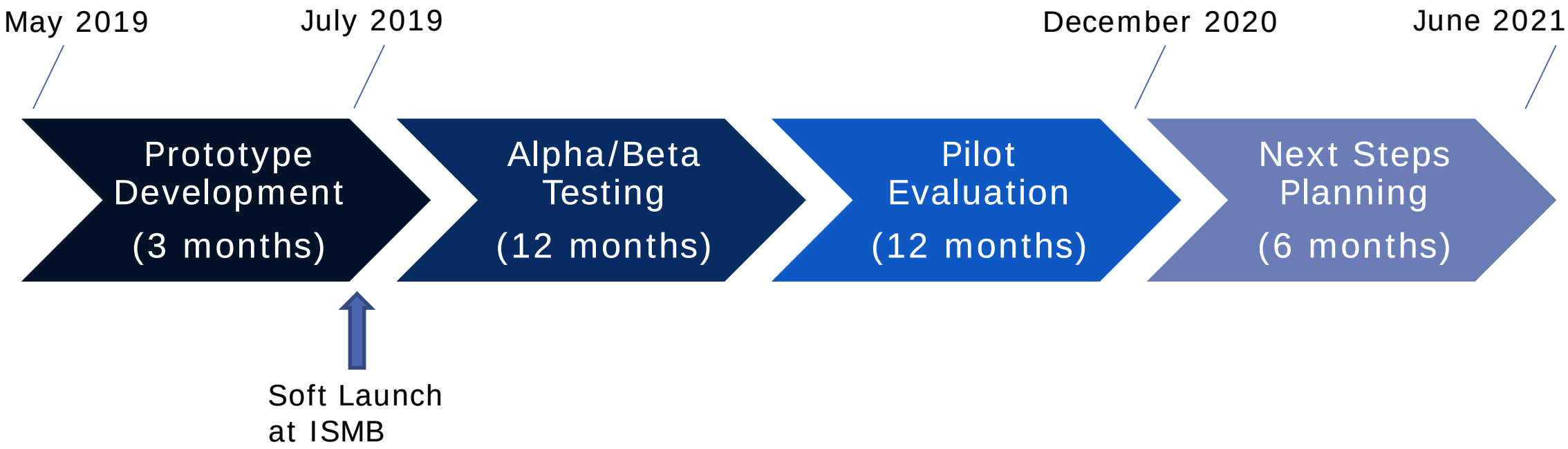
https://nih.figshare.com/



Make NIH-supported research data **citable**, **shareable**, and **discoverable** in a few easy steps.

- Available for all NIH-supported researchers
- Allows researchers to share:
  - Research data that do not have a **domain-specific repository**
  - Data underlying publication figures and tables
  - Useful data not associated with publications

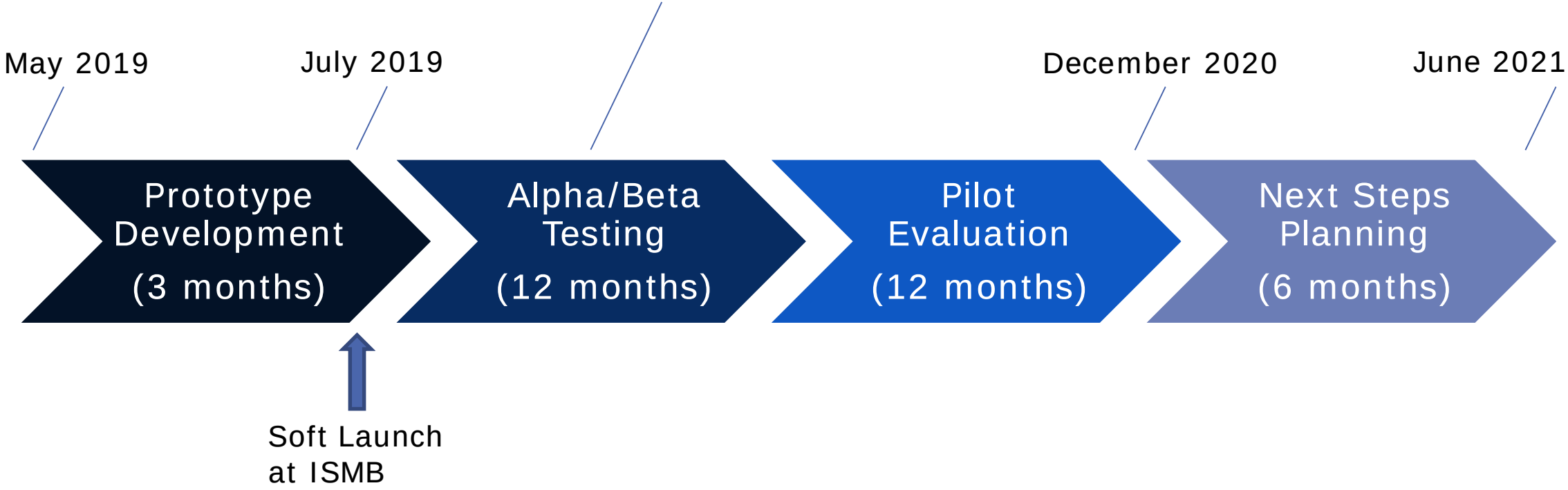
NIH figshare Pilot Timeline



After pilot, data will still be available in figshare.

NIH figshare Pilot Timeline

Workshop on “The Role of Generalist and Institutional Repositories to Enhance Data Discoverability and Reuse”



After pilot, data will still be available in figshare.

Sharing Publication-Related Data

NIH strongly encourages use of open domain-specific 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 GB**

Terabytes

PubMed Central (PMC)

Datasets associated with PMC articles

Separate unique identifiers assigned to datasets & articles

Generalist / Institutional Repositories

figshare pilot project 100 GB

Datasets assigned unique identifiers & linked to PubMed citations

Sharing Publication-Related Data

NIH strongly encourages use of open domain-specific 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 GB**	**Terabytes**	**Petabytes**
PubMed Central (PMC) Datasets associated with PMC articles Separate unique identifiers assigned to datasets & articles	**Generalist / Institutional Repositories** figshare pilot project 100 GB Datasets assigned unique identifiers & linked to PubMed citations	**STRIDES Cloud** Large-scale, high-priority datasets with unique identifiers Coherent authentication & authorization management

STRIDES Cloud

***S**cience & **T**ech **R**esearch **I**nfrastructure for **D**iscovery, **E**xperimentation & **S**ustainability*

STRIDES Cloud

***S**cience & **T**ech **R**esearch **I**nfrastructure for **D**iscovery, **E**xperimentation & **S**ustainability*

- Agreements with
  - Google Cloud
  - Amazon Web Services
  - Additional partnerships anticipated
- Other Transaction Authority used

STRIDES Cloud

***S**cience & **T**ech **R**esearch **I**nfrastructure for **D**iscovery, **E**xperimentation & **S**ustainability*

- Agreements with
  - Google Cloud
  - Amazon Web Services
  - Additional partnerships anticipated
- Other Transaction Authority used

- Benefits to NIH-supported Investigators
  - Discounted **rates** on cloud resources
  - Access to engineering, consulting, and other professional **services** from cloud service provider partners
  - Access to cloud **training** programs (standard & custom)

Examples of datasets moving / moved to STRIDES

- NLM Sequence Read Archive
- NHLBI Framingham Heart Study
- All of Us Research Program
- NCI Genomic Data Commons
- NHLBI Trans-Omics for Precision Medicine (TOPMed) Program
- NCI Proteomics Data Commons and Imaging Data Commons
- NIMH Data Archive
- Gabriella Miller Kids First Pediatric Research Program
- Transformative CryoEM Program Data

Repositories for NIH Data – An Embarrassment of Riches

Many & Many Types of Repositories

- Domain-specific
- PubMed Central-linked
- Generalist / Institutional Repositories
- STRIDES Clouds

Repositories for NIH Data – An Embarrassment of Riches

Many & Many Types of Repositories

- Domain-specific
- PubMed Central-linked
- Generalist / Institutional Repositories
- STRIDES Clouds
- NIH Funding Opportunity Announcements to fund Databases and Knowledgebases will be issued soon – **so, even more options!**

Repositories for NIH Data – An Embarrassment of Riches

Many & Many Types of Repositories

- Domain-specific
- PubMed Central-linked
- Generalist / Institutional Repositories
- STRIDES Clouds
- NIH Funding Opportunity Announcements to fund Databases and Knowledgebases will be issued soon – so, even more options!

Challenge: Which repositories are good enough to use?

BMIC identified **desirable characteristics** of repositories for NIH research data informed by:

- Clarivate Analytics
- CODATA
- CoreTrust Seal
- Data Seal of Approval
- Digital Curation Center
- Earth System Science Data
- FAIR Principles
- Interagency input
- Nature Scientific Data
- PLOS
- Research Data Alliance
- Science Europe
- Smithsonian
- US Geological Survey

Characteristics Mapped to

- Dryad
- figshare
- Mendeley Data

Desirable Characteristics of Data Repositories

- persistent unique identifiers
- long-term sustainability
- metadata
- curation & quality assurance
- maximally open access
- tracking data re-use
- free of charge
- secure
- common format
- provenance

Additional Considerations for Repositories With Human Data

- fidelity to consent
- restricted use compliant
- privacy
- plan for breach
- download audit & control
- clear use guidance
- retention guidance
- plan for use violations
- request review


NIH
National Institutes of Health
Turning Discovery Into Health


EXECUTIVE OFFICE OF THE PRESIDENT OF THE UNITED STATES
OFFICE OF SCIENCE AND TECHNOLOGY POLICY


NIH
National Institutes of Health
Turning Discovery Into Health


USGS
science for a changing world


EXECUTIVE OFFICE OF THE PRESIDENT OF THE UNITED STATES
OFFICE OF SCIENCE AND TECHNOLOGY POLICY


NSF


AHRQ
Agency for Healthcare
Research and Quality


DEPARTMENT OF DEFENSE
UNITED STATES OF AMERICA


NASA


NIH
National Institutes of Health
Turning Discovery Into Health


DEPARTMENT OF ENERGY
UNITED STATES OF AMERICA


DEPARTMENT OF TRANSPORTATION
UNITED STATES OF AMERICA


UNITED STATES AGENCY
USAID
INTERNATIONAL DEVELOPMENT


U.S. DEPARTMENT OF
HOMELAND SECURITY


ACL
Administration for Community Living


CDC
CENTERS FOR DISEASE
CONTROL AND PREVENTION


INSTITUTE of
Museum and Library
SERVICES

Government-wide **RFI on desirable repository characteristics** expected soon

Successful DS & OS Strategies Depend on Good RDM

Policy implementation for open science

Enhancing the repository ecosystem

Connecting repositories through metadata

The Miracle of Metadata

- Publications report on >200,000 datasets generated with NIH funding each year
- Datasets in many and many types of repositories

The Miracle of Metadata

- Publications report on >200,000 datasets generated with NIH funding each year
- Datasets in many and many types of repositories
  - How to find relevant repositories?
  - How to find relevant datasets across repositories?

The Miracle of Metadata

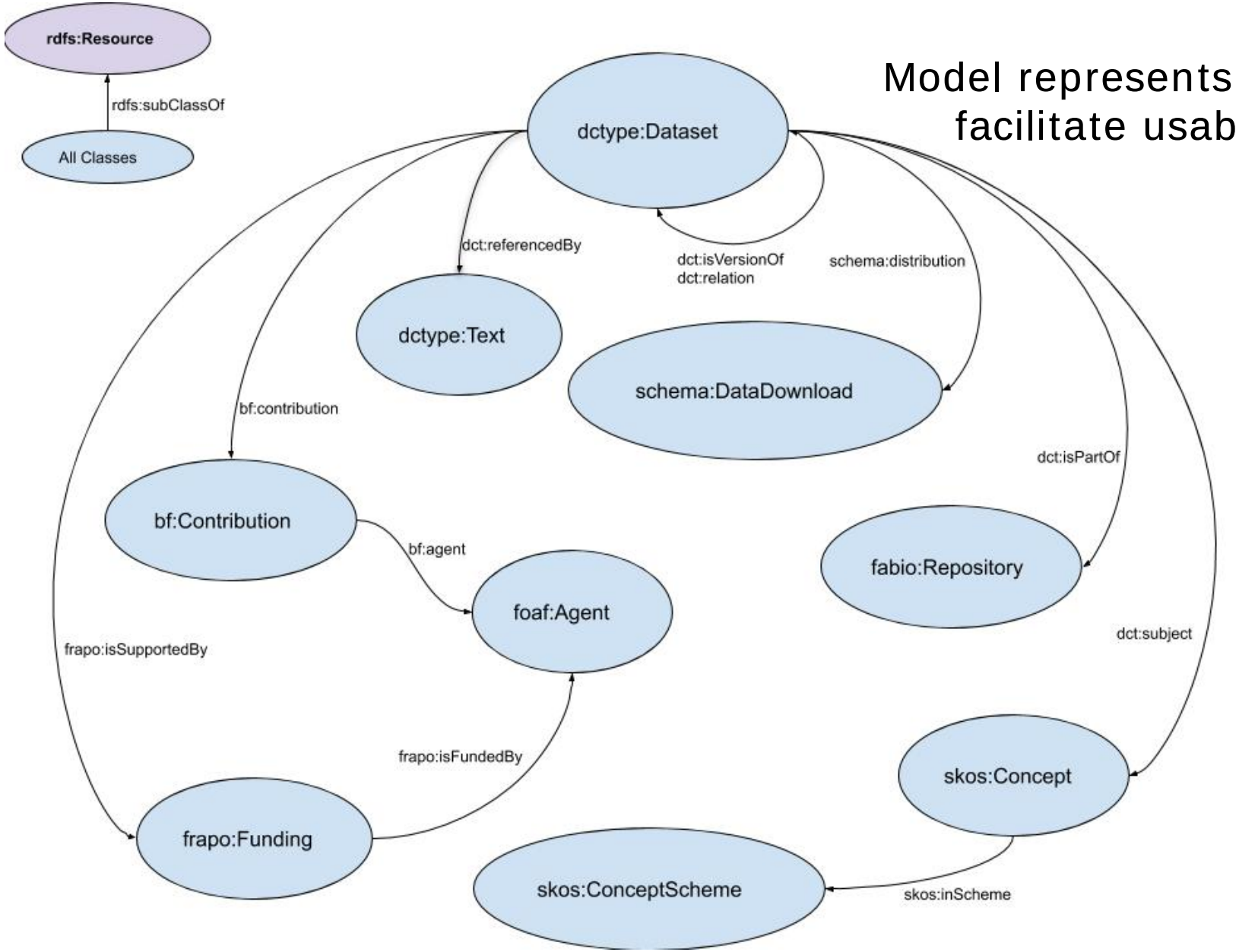
- Publications report on >200,000 datasets generated with NIH funding each year
- Datasets in many and many types of repositories
  - ➢ How to find relevant repositories?
  - ➢ How to find relevant datasets across repositories?

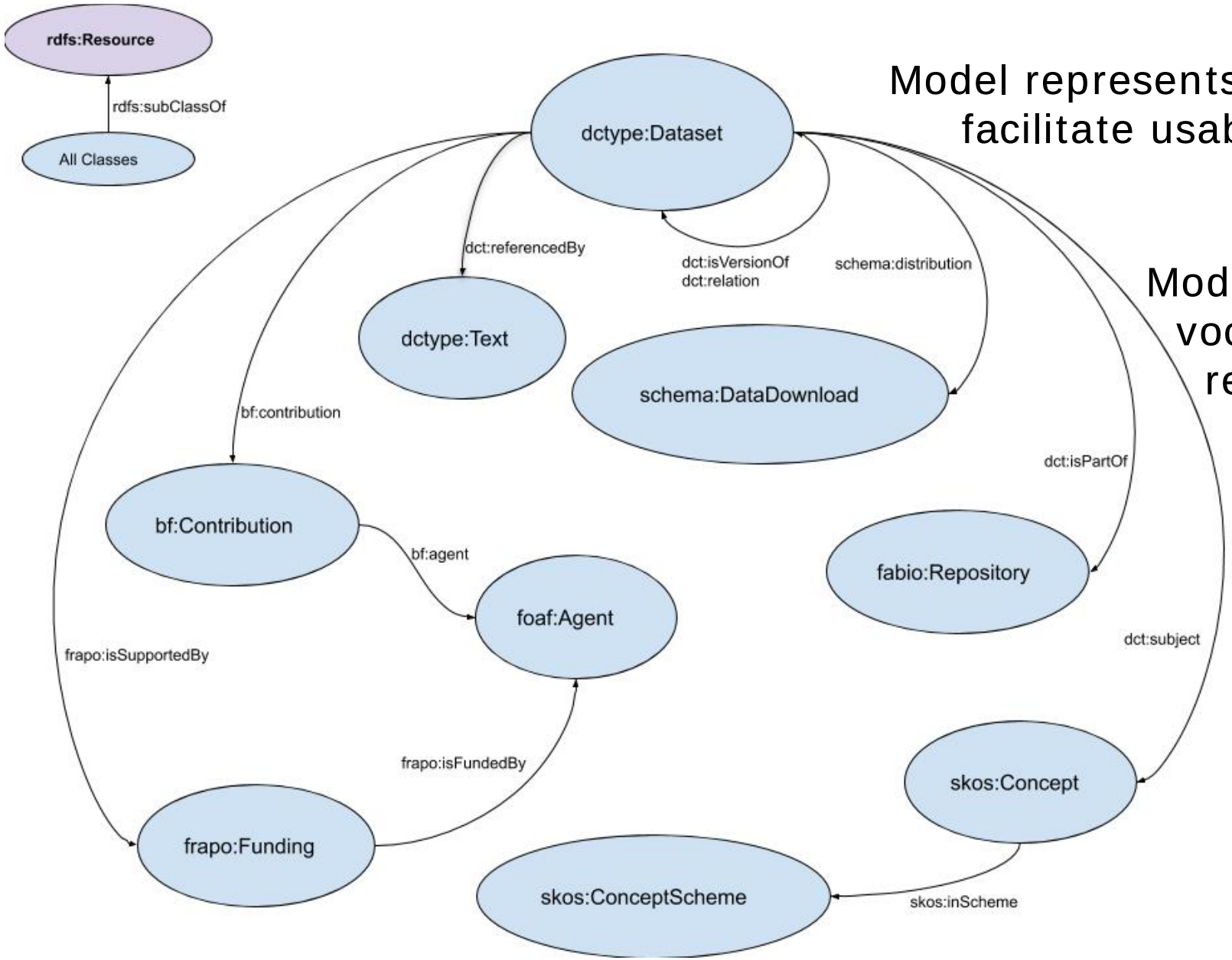
NLM Dataset Metadata Model (DATMM) as a common way to find and "connect" repositories

Trans-NLM DATMM Team

- **Pete Seibert - LO**
- Jeff Beck - NCBI
- Philip Chuang - OCCS
- Dan Davis - OCCS
- Jason Eshleman - NCBI
- Nancy Fallgren - LO
- Lisa Federer - OSI
- Rob Guzman - LO
- Karen Nimerick - LO
- Alvin Stockdale - LO
- Ying Sun - OCCS

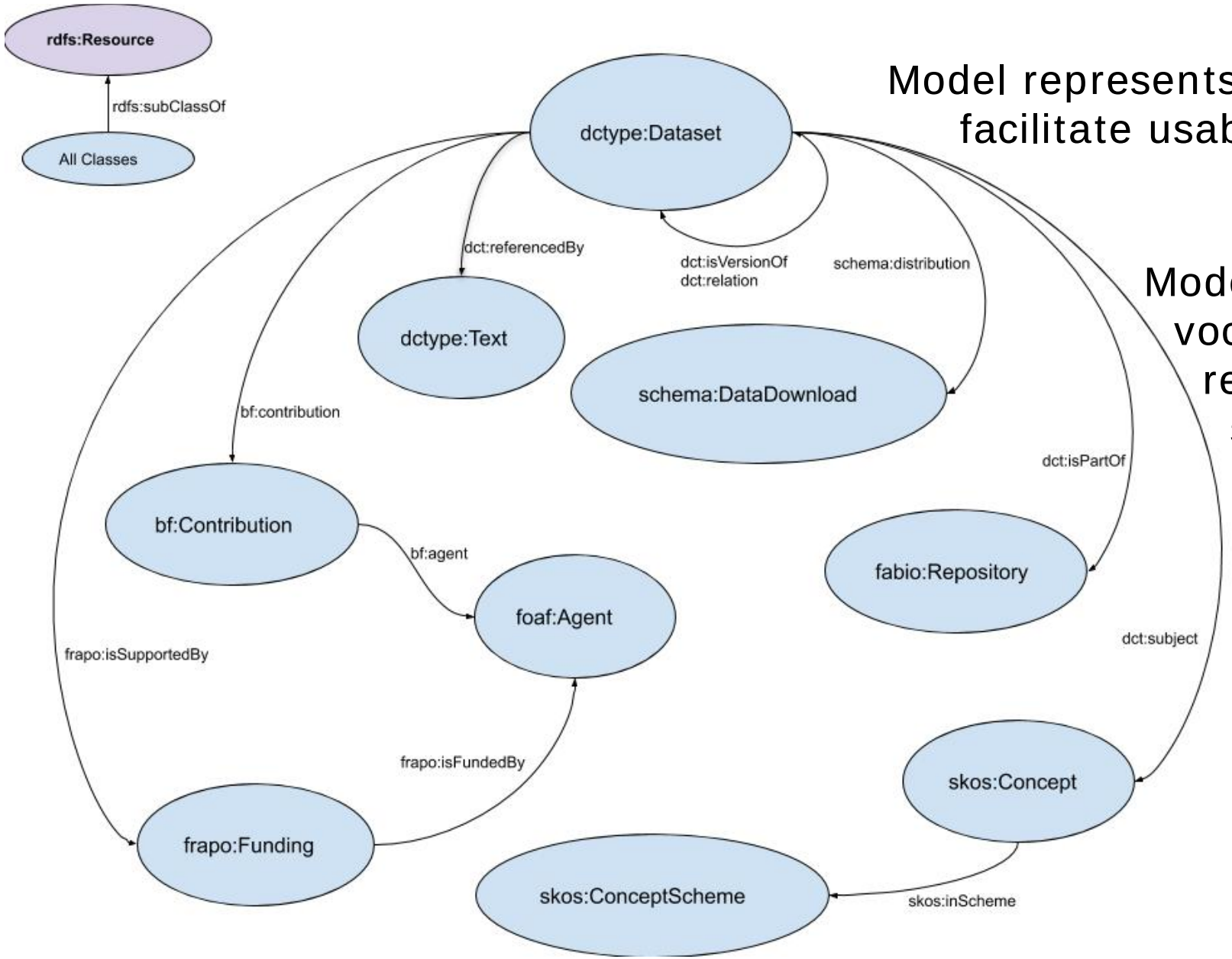
Model represents **simple** core metadata to facilitate usability and encourage adoption





Model represents **simple** core metadata to facilitate usability and encourage adoption

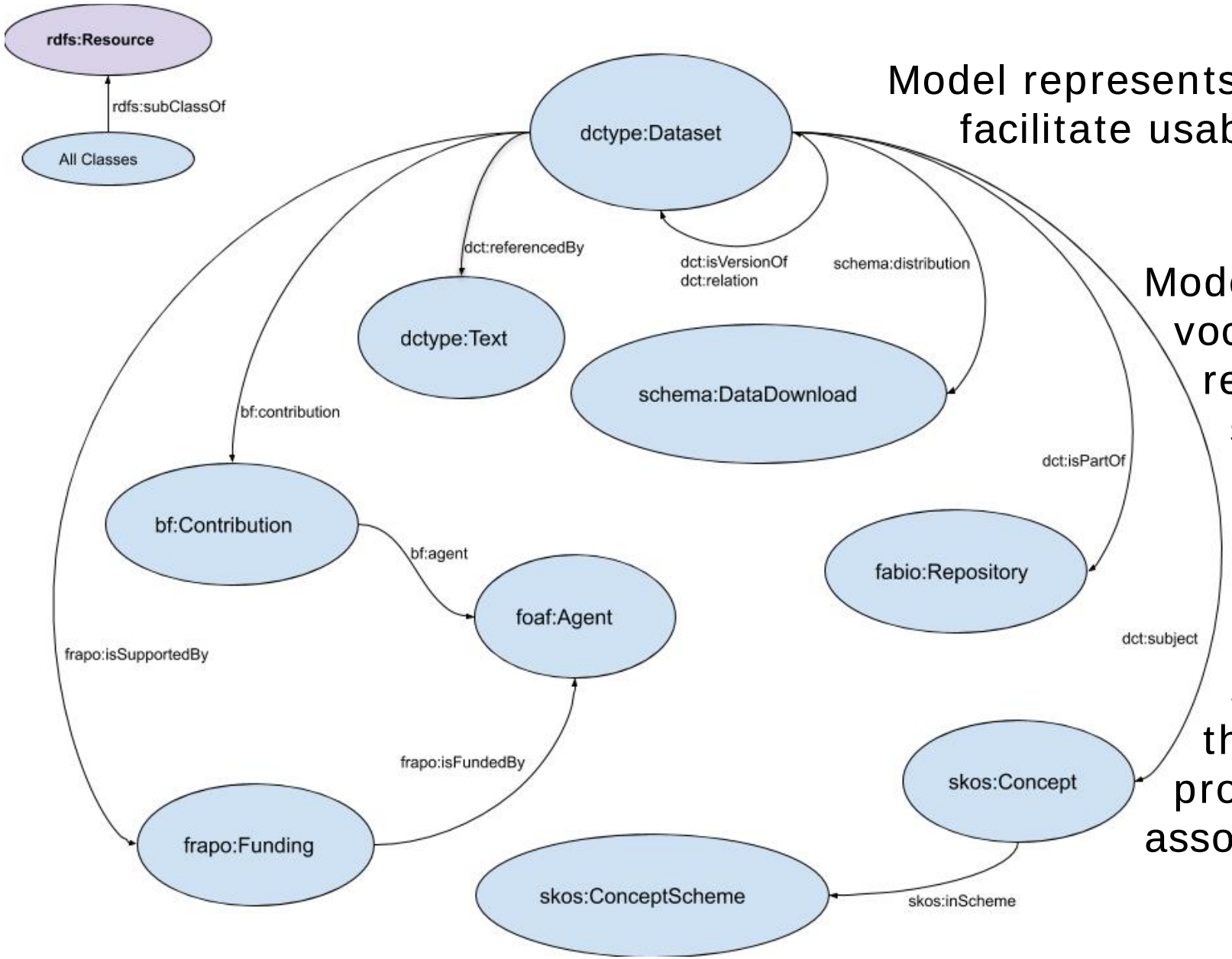
Model implemented via **existing RDF** vocabularies for interoperability, re-use, and linking specific data values



Model represents **simple** core metadata to facilitate usability and encourage adoption

Model implemented via **existing RDF** vocabularies for interoperability, re-use, and linking specific data values

Works with **schema.org**



Model represents **simple** core metadata to facilitate usability and encourage adoption

Model implemented via **existing RDF** vocabularies for interoperability, re-use, and linking specific data values

Works with **schema.org**

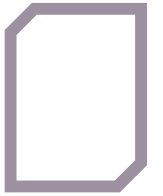
Discoverability is driven by **semantic connections** among the classes/entities and property elements that are associated with a dataset

DATMM – Next Steps

Map use cases to individual or multiple metadata elements

Author model into **working schema** using CEDAR

Use test metadata and data to **validate** model/schema



Documentation and post for external **review**

-**Implement** in central data catalog

-Socialize, promote, & encourage **adoption** by repositories (optional)

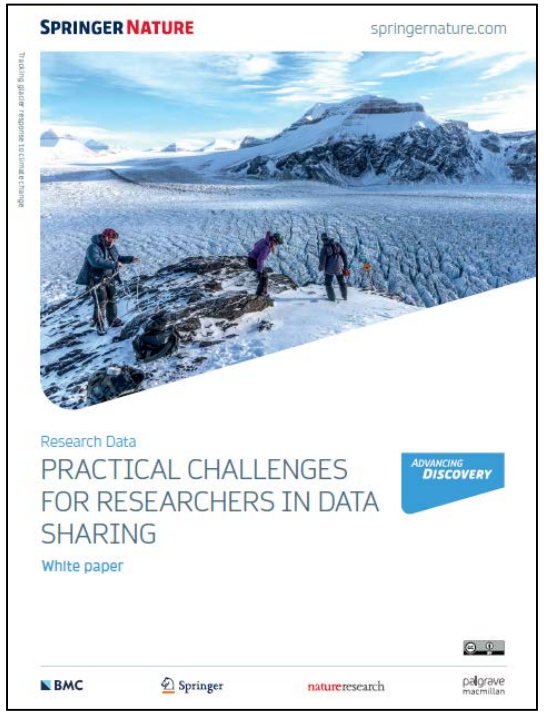
Successful DS & OS Strategies Depend on Good RDM

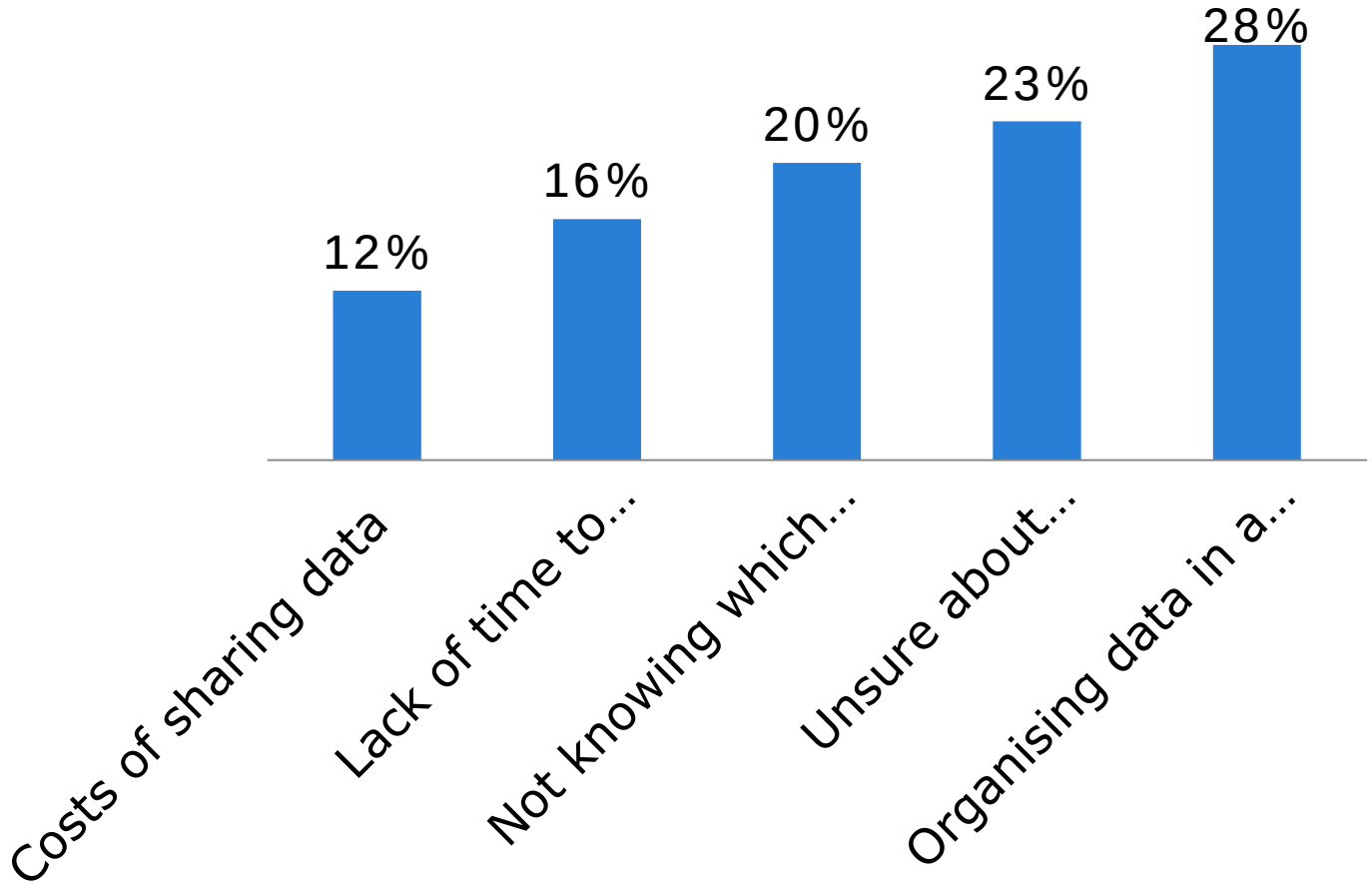
Policy implementation for open science

Enhancing the repository ecosystem

Connecting repositories through metadata

What problems do authors have in sharing datasets?





- From a Springer Nature researcher survey. Total respondents: 7719
- https://doi.org/10.6084/m9.figshare.5975011

Grace Baynes / 4/29/19

NLM Office of Strategic Initiatives

Data Science & Open Science Team

Lisa Federer, PhD, MLIS
NLM Data Science & Open Science Librarian

Maryam Zaringhalam, PhD
NLM Data Science & Open Science Officer

Teresa Zayas-Caban, PhD
Coordinator, NIH FHIR Acceleration
Chief Scientist, ONC, DHHS

Rebecca Goodwin, JD
Policy Analyst & Open Science Specialist

Tony Chu, PhD, MLIS
Information Scientist

NLM Office of Strategic Initiatives

Data Science & Open Science Team

Lisa Federer, PhD, MLIS
NLM Data Science & Open Science Librarian

Maryam Zaringhalam, PhD
NLM Data Science & Open Science Officer

Teresa Zayas-Caban, PhD
Coordinator, NIH FHIR Acceleration
Chief Scientist, ONC, DHHS

Rebecca Goodwin, JD
Policy Analyst & Open Science Specialist

Tony Chu, PhD, MLIS
Information Scientist

THANKS!