Evolution of the Biomedical Data Repository Ecosystem

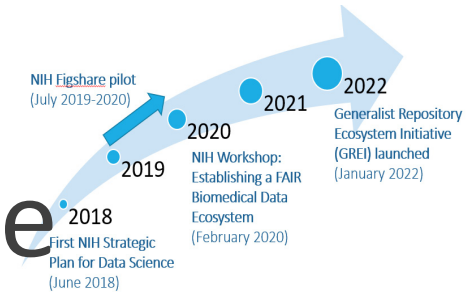
the role of generalist repositories

Jennie Larkin, PhD
Deputy Director,
Division of Neuroscience, NIA

GREI Workshop Panel on NIH Stakeholder Perspectives
January 25, 2023

Timeline of NIH Generalist Repository Activities

2018 NIH Strategic Plan for Data Science



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 observational and clinical data into biomed data science

Data Management, Analytics & 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

2018 Guidance: Repositories for Biomedical Data Sharing

NIH encourages researchers to share data using DOMAIN-SPECIFIC REPOSITORIES when available.

When domain-specific data repositories are not available, NIH is developing options to support data sharing.

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

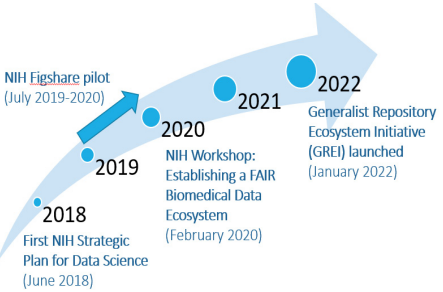
- Assign Unique Identifiers to datasets associated with publications and link to PubMed.
- Store and manage datasets associated with publication, up to 20* GB.

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.

Data Archipelago, not a Landscape



NIH needed to:

- Find additional data sharing solutions to build out the data ecosystem
- Understand the potential role of the rapidly growing Generalist Repositories

NIH encourages researchers to share data using DOMAIN-SPECIFIC REPOSITORIES when available.

When domain-specific data repositories are not available, NIH is developing options to support data sharing.

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.

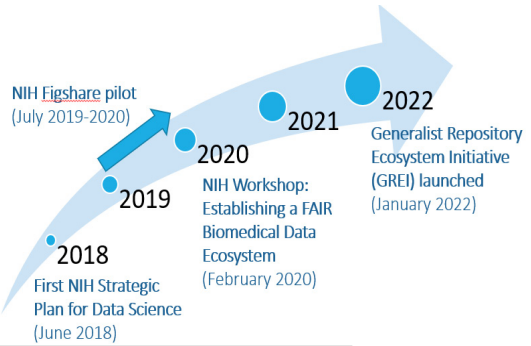
High Priority Datasets **petabytes**

STRIDES Cloud Partners

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

Established a one-year pilot to investigate the potential of Generalist Repositories

Opportunities: Generalist Repositories & NIH



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 observational and clinical data into biomed data science

Data Management, Analytics & Tools

- Support useful, generalizable, and accessible tools
- Broaden utility of, and access to, specialized tools
- Improve discovery and cataloging resources

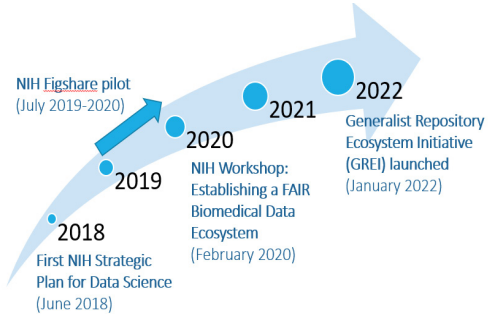
Workforce Development

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

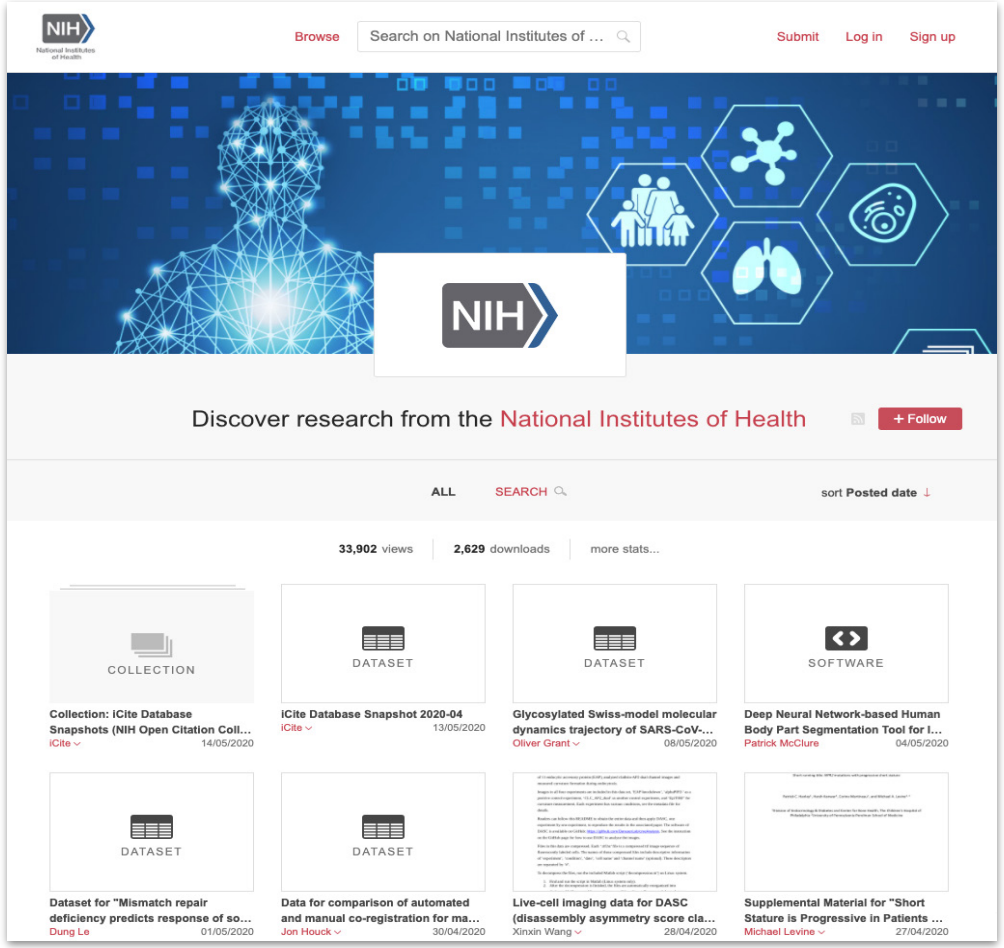
Stewardship and Sustainability

- Develop policies for a FAIR data ecosystem
- Enhance stewardship

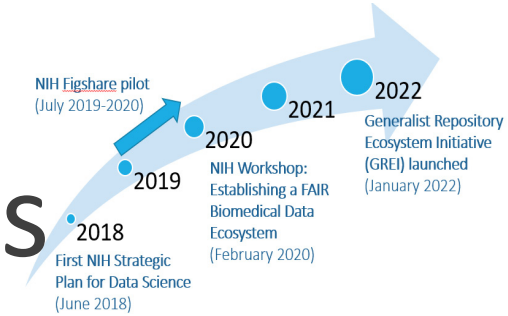
2019 Figshare Pilot



- **Pilot a data sharing platform** to enable NIH funded researchers to easily share data that was generated from NIH support.
- For research that **did not have an alternative venue** through which to share publication-related data.
- Enable the NIH to **better understand the patterns of data sharing by the researchers** it funds.
- Allowed storage of **datasets, spreadsheets, multimedia**, but NOT posters, slides or preprints.
- Increased funding **metadata and QC support**.

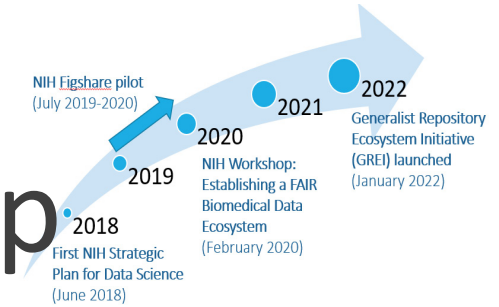


NIH Figshare Pilot – Key Takeaways



- **Generalist repositories are growing –** more researchers are depositing data and more publications are linking to generalist repositories.
- **Researchers need more education and guidance –** where to publish data and how to describe datasets in metadata fields effectively.
- **Metadata enhancement enables greater discoverability –** metrics indicate greater access but need longer time scale to observe data reuse.
- **There is a clear need for the services that repositories like Figshare provide –** researchers have data (and other materials) that they want to share but there are not suitable repositories.

2020 Generalist Repository Workshop



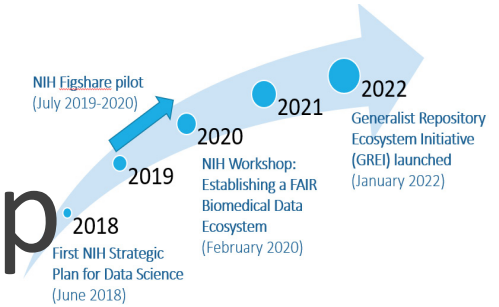
- **Establishing a FAIR Biomedical Data Ecosystem: Role of Generalist and Institutional Repositories to Enhance Data Discoverability and Reuse** (February 2020)
- Highlighted the breadth and depth of activities of generalist and institutional repositories in supporting biomedical researchers and biomedical data

Day 1: A Landscape of Generalist Repositories and Challenges in Data Discovery and Reuse

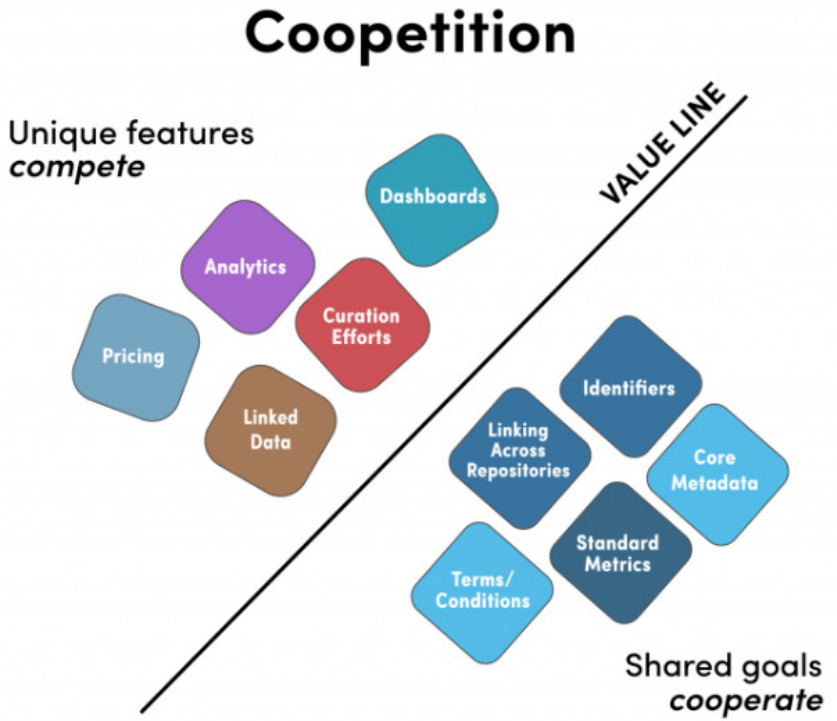
Day 2: Facilitating Reproducibility and Managing Technical and Cultural Change in Research

https://datascience.nih.gov/data-ecosystem/nih-data-repository-workshop-summary

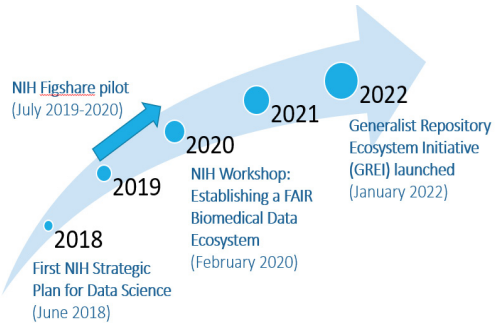
2020 Generalist Repository Workshop



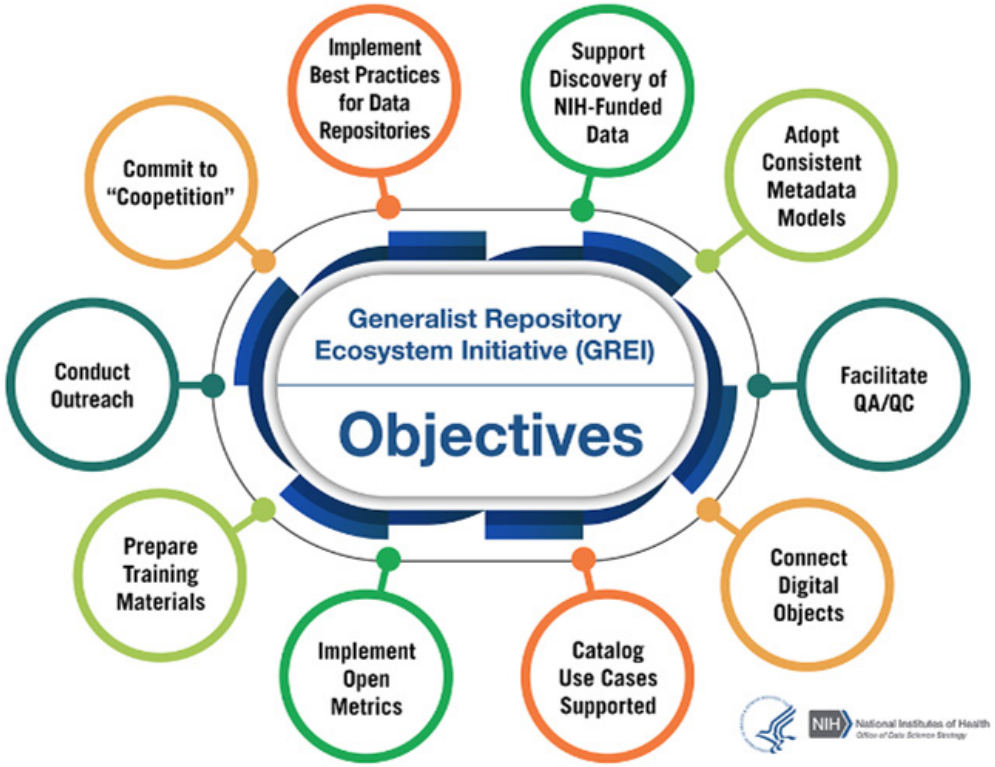
- **Coopetition**: A vibrant repository community includes participants in industry, academia, and government
- **Value line** to determines where to compete on services and when to collaborate.
- Following the workshop, generalist repositories worked together to develop a summary of repository capabilities to establish the **"coopetition" framework**.



2022: Generalist Repository Ecosystem Initiative (GREI)



- GREI is intended to **supplement the domain-specific data repositories** that are critical components of the NIH biomedical data ecosystem for data sharing.
- GREI repositories are identifying and implementing **collaborative activities**: shared metadata for improved discoverability, inclusion of funding and grant information, etc.
- GREI aims to make it easier to find and reuse NIH-funded data.



NIA & Generalist Repositories

Who is NIA?

- NIA is the third largest NIH Institute ($4.4B in FY23)
- Supports a **broad range of science**: the breadth of science from across NIH: basic (molecular cellular and genetic), clinical, behavioral, epidemiological, and therapeutics.
  - NIA leads a broad scientific effort to understand the nature of aging and to extend the healthy, active years of life.
  - Dedicated funding to support research on NIA Alzheimer's Disease (AD) and related dementias (ADRD).
- In addition to NIA's Intramural Research Program in Bethesda and Baltimore. NIA and NINDS support NIH Intramural **Center for Alzheimer's and Related Dementias** (CARD).
  - CARD seeks to advance AD/ADRD research through a data-driven and collaborative approach that emphasizes robust, replicable findings and cooperative progress over individual success.

Challenges to NIA Data Sharing

- Diversity of NIA research makes it challenging to also identify and support domain-specific repositories.
- NIA-supported data repositories and knowledgebases: https://www.nia.nih.gov/research/data-sharing-resources-researchers

AgingResearch Biobank

PAD

NIDUS Network for Investigation of Delirium: Unifying Scientists

AD Knowledge Portal

NACDA National Archive of Computerized Data on Aging

NIAGADS

Challenges to NIA Data Sharing

- NIA encourages the use of domain-specific repositories (supported by NIA, where possible, otherwise other repositories that align with the Key Characteristics).
- But Generalist Repositories will play an important role to:
  - Provide a FAIR repository when no specialized repositories exist
  - Facilitate in connecting datasets across multiple repositories, for projects that generate diverse data types.
- As DMS Policy gets implemented, we anticipate **increased use of all repositories** (including GREI) and **improved and more consistent metadata** at all repositories to support easier discovery and access … to realize the FAIR biomedical data ecosystem.
- Future Challenge: ensuring that NIA-supported data is not only Findable and Accessible but is also truly **Interoperable and Reusable**!

Thank you!

Jennie.Larkin@NIH.gov