NIH's Strategic Vision for Data Science: Enabling a FAIR-Data Ecosystem

Susan Gregurick, Ph.D. Senior Advisor Office of Data Science Strategy

June 13, 2019





VISION

a modernized, integrated, FAIR biomedical data ecosystem

premiegeum conce la prese, consecutorial adaptecing encladed even recentering reconserved incident et laorent deloren magnetialiquame alquamet volutati. Ut will encladed ad minister versione, quila costruid evensi tation utilianoosper autounte laborati utilitati galagia exita i consequati, soli sola autorn vel even intere dolor in heoderet in vulgutate velt evan molectile consequati, vel iven dolore evelinguti nulla facilitati at vero evise et accomune et iusto odio digretatem qui

Selection temporal default of events conservational commons, and the selection of the control of the control

A second second

IMAGINE...

the ability to link data in the Framingham Heart Study (NHLBI) with Alzheimer's health data (NIA) to understand correlative effects in cardiovascular health with aging and dementia.



Linking Data Platforms

IMAGINE... the ability to quickly obtain access to data, and related information, from published articles.



Negative stain EM reveals the principal architecture of the rhodopsin/GRK5 complex. (Image by Van Andel Research Institute)



Absorption spectra of purified CsR-WT (A) and CySeR (B) at pH 5 (green), pH 7.4 (red), and pH 9 (blue). R. Fudim, e al, Science Signaling, 2019



Energetics of Chromophore Binding in the Visual Photoreceptor of Rhodopsin, H. Tian et al, Biophysical Journal, 2017.

Dark Data

IMAGINE... the ability to link electronic health care records with personal data and with clinical and basic research data.



Data Security and Data Privacy

the new capabilities that artificial intelligence and advanced technologies offer medical research, treatment, and prevention.







IMAGINE...





This is the promise of the NIH Strategic Plan for Data Science

...and here's how we will get there.

Strategic Plan for Data Science: Goals and Objectives

Data Infrastructure	Modernized Data Ecosystem	Data Management, Analytics, and Tools	Workforce Development	Stewardship and Sustainability
Optimize data storage and security	Modernize data repository ecosystems	Support useful, generalizable, and accessible tools	Enhance the NIH data science workforce	Develop policies for a FAIR data ecosystem
	Support storage and sharing of individual datasets	Broaden utility of, and access to, specialized tools	Expand the national research workforce	
Connect NIH data systems	Better integrate clinical and observational data into biomedical data science	Improve discovery and cataloging resources	Engage a broader community	Enhance stewardship



Strategic Plan for Data Science: Goals and Objectives

FAIR Data and Data Infrastructure Connecting NIH Data Ecosystems Engaging with a Broader Community Enhancing Biomedical Workforce

Sustainable Data Policies



Implementation Progress: Oct. 2018 – Present

- FAIR Data and Data Infrastructure
- Sustainable Data Policies
- Connecting NIH Data Ecosystems
- Engaging with a Broader Community
- Enhancing Biomedical Workforce

Making Data FAIR

Findable	 must have unique identifiers, effectively labeling it within searchable resources. 		
Accessible	 must be easily retrievable via open systems and effective and secure authentication and authorization procedures. 		
nteroperable	 should "use and speak the same language" via use of standardized vocabularies. 		
Reusable	 must be adequately described to a new user, have clear information about data-usage licenses, and have a traceable "owner's manual," or provenance. 		

Overview of Sharing Publication and Related Data

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

and datasets.

The TRUST Principles for Data Repositories

Transparency	 is achieved by providing publicly accessible evidence of the services that a repository can and can not offer.
Responsibility	 is a commitment to provide high technical quality data services.
User community	 is the focus on the uses and potential uses of the data and services offered.
Sustainability	 is the capability to support long-term data preservation and use.
Technology	 is the infrastructure and capabilities to support the repository operations.

Develop Characteristics for Open Access Data Sharing Repositories



- Characteristics drafted, includes provisions for repositories with human data
- Developed and reviewed in trans-NIH process
- Planned Community Input: Request for Information (RFI)

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



Sharing Datasets as Supplementary Materials

<u>Autophagy</u> . 2017; 13(2): 386–403.	PMCID: PMC5324850
Published online 2016 Nov 22. doi: <u>10.1080/15548627.2016.1256934</u>	PMID: 27875093
Autolysosome biogenesis and developmental senesce both Spns1 and v-ATPase	ence are regulated by
<u>Tomoyuki Sasaki</u> , ^{a,†} <u>Shanshan Lian</u> , ^{a,†} <u>Alam Khan</u> , ^{a,b} <u>Jesse R. Llop</u> , ^c <u>Andrew V. S</u> <u>Daniel J. Klionsky</u> , ^e and <u>Shuji Kishi</u> ^a	amuelson, ^c Wenbiao Chen, ^d
 Author information Article notes Copyright and License information <u>Discl</u> 	aimer
This article has been <u>cited by</u> other articles in PMC.	
Associated Data	
 Supplementary Materials 	
1256934_Supplemental_Material.zip	
<u>kaup-13-02-1256934-s001.zip</u> (9.6M)	
GUID: AC7F9D11-8BEB-402D-9437-6E7942A3ACC6	

Piloting a Repository to Make Research Data Citable, Sharable, and Discoverable Using Figshare

Data is openly accessible	Documented with customizable, discipline-specific metadata	Authors can link grant information to data	All data is associated with a license	Self-publish any data type in any file format
Assign institutionally (NIH) branded DOI	Indexed in Google and discoverable across search engines	Ability to embargo data assets	Usage metrics tracked openly	FAIR implementation

Science & Tech Research Infrastructure for Discovery, Experimentation and Sustainability Initiative

- First **STRIDES** agreement: Google Cloud (July 2018)
- Second STRIDES agreement: Amazon Web Services (Oct. 2018)
- Other Transaction mechanism
- Additional partnerships anticipated https://datascience.nih.gov/strides

FAIR Data: Move/Access to high priority data sets in cloud service providers



Amazon And NIH To Link Biomedical Data And Researchers There is immense potential here to advance human health by driving new discoveries that enable more accurate disease risk prediction, tailored diag... forbes.com

Examples of Datasets Moving to the STRIDES Cloud

- NHLBI Framingham Heart Study
- All of Us Research Program
- NCI Genomic Data Commons
- NCBI data resources
- 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
- And many others!

Overview of Sharing Publication and Related Data

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

and datasets.

Implementation Progress: Oct. 2018 – Present

- FAIR Data and Data Infrastructure
- Sustainable Data Policies
- Connecting NIH Data Ecosystems
- Engaging with a Broader Community
- Enhancing Biomedical Workforce

NIH Data Management and Sharing Policy Development: Status

- Seek input from stakeholders
- Develop draft policy and any needed suggested guidance
- Seek more input from stakeholders
- Incorporate feedback and release final policy



Implementation Progress: Oct. 2018 – Present

- FAIR Data and Data Infrastructure
- Sustainable Data Policies
- Connecting NIH Data Ecosystems
- Engaging with a Broader Community
- Enhancing Biomedical Workforce

NIH's Data Environments are Rich, but Siloed



Single method for sign-on and data access across repositories and CSPs

Single 'Sign-on' Across NIH Data Resources

- Streamlined login for authorization of controlledaccess data
- Make use of industry standard technology (web tokens)
- Flexible for different NIH needs: 'do no harm to existing systems'

 End goal: NIH-wide system for a consistent method to access data across NIH data resources



Implementation Progress: Oct. 2018 – Present

- FAIR Data and Data Infrastructure
- Sustainable Data Policies
- Connecting NIH Data Ecosystems
- Engaging with a Broader Community
- Enhancing Biomedical Workforce

Leverage, Develop, and Extend Methods and Tools from Broader Communities

- Partner with other federal agencies (e.g. National Science Foundation) on data science activities
- Leverage SBIR/STTR to bring in industry expertise
- Engage a broader community through codeathons, citizen science, and challenges
- Improve software sustainability, efficiency and utility







Implementation Progress: Oct. 2018 – Present

- FAIR Data and Data Infrastructure
- Sustainable Data Policies
- Connecting NIH Data Ecosystems
- Engaging with a Broader Community
- Enhancing Biomedical Workforce

Enhance the Biomedical Workforce



Graduate Data Science Summer Program

- 10 undergraduate fellows for 2019 placed in admin or funding offices for 10-week summer program
- Student-led non-profit places tech-savvy students in federal agencies
- ODSS will coordinate on-campus networking opportunities for fellows

https://www.codingitforward.com/

- 13 master's-level interns for 2019
- Pilot driven by discussion with local universities consortium
 - UVA, George Mason, George Washington, UMD, University of Delaware/Georgetown, Johns Hopkins
- Open to students from any university

https://www.training.nih.gov/data_science_summer



NIH Data Science Senior Fellowships

- One- or two-year national service sabbatical in highimpact NIH programs
- Seeking data science and technology experts
- Work with large volumes of biomedical research data, impact public health, gain policy exposure

- Expecting 5+ fellows in first cohort, starting late 2019
- Program evaluation in 2024



Improve Data Science-Related Training through T Grants, F and K Awards

- Expand expectations for development of quantitative and computational skills for students and postdoctoral fellows supported by NIH training (T) grants
 - NIGMS T32/T34, Neuroscience T32, or NLM T15 FOAs
- Disseminate across all training mechanism and ICs

- Launch data science-focused training programs in specific biomedical research areas of high need
 - Biomedical behavioral and social science
 - Neuroscience
 - 🗸 RFA-OD-19-011

Improving R&R and RCR and Evaluating Efficacy of Interventions

- Support development of training modules to fill in gaps in rigor and reproducibility in data science
- Support training modules on responsible conduct of research in data science
- Improvement and expansion of K25 program (Mentored Quantitative Research Development – PA-18-396)



Building Diversity in Biomedical Data Science

- Boot camps/short training programs for diverse cohorts of biomedical research trainees
 - Includes STRIDES CSP and professional societies
- Increase emphasis on quantitative and computational skills development in existing diversity programs
 - E.g. new language in NIGMS FOAs (already in place)



VISION

a modernized, integrated, FAIR biomedical data ecosystem

chant ut laonent dokore magna aliquam erat volutput, Ut will enem all miner venlam, quit, trud exerci tation ullamourper inacipit kisostis mil ut aliquip ex ea comensió consequat, a utore vet eum intre dokor in hendrert in volguitate vette exat indepite consequat, set o dokore au broustit rulla focilita at vero anno et a comensant functio doko dokorium (set





Special Thanks

• **STRIDES:** Andrea Norris, Nick Weber and NMDS team

 \bigcirc

- Connecting NIH Data Resources: Vivien Bonazzi, Regina Bures, Ishwar Chandramouliswaran, Tanja Davidsen, Valentine Di Francesco, Jeff Erickson, Tram Huyen, Rebecca Rosen, Steve Sherry, Alastair Thomson, Nick Weber, and BioTeam
- Linking Publications to Datasets: Jim Ostell and NCBI Implementation Team
- Data Repository and Knowledgebase Resources: Valentina di Francesco, Ajay Pillai, Qi Duan, Dawei Lin, Christine Colvis, and James Coulombe
- Trustworthy Data Repositories: Dawei Lin, Kim Pruitt, Jennie Larkin, Elaine Collier, Christine Melchior, Minghong Ward, and Matthew McAuliffe
- Criteria for Open Access Data Sharing Repositories: Mike Huerta, Dawei Lin, Maryam Zaringhalam, Lisa Federer and BMIC Team
- Pilot for Scaled Implementation for Sharing Datasets: Ishwar Chandramouliswaran and Jennie Larkin
- Coding-it-Forward Fellows Summer Program: Jess Mazerik
- Graduate Data Science Summer Program: Sharon Milgram and Phil Ryan (OITE)
- Data Science Training: Valerie Florance, Jon Lorsch, Kay Lund, Kenny Gibbs, Shoshana Kahana, Erica Rosemond, Carol Shreffler
- Diversity in Biomedical Data Science: Valerie Florance, Jon Lorsch, Hanna Valantine, Roger Stanton, Charlene Le Fauve, Ravi Ravichandran, Zeynep Erim, Derrick Tabor, Rick Ikeda



Stay Connected





www.datascience.nih.gov

