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

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

September 4th, 2019
VISION

a modernized, integrated, FAIR biomedical data ecosystem
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.
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

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

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

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

[https://datascience.nih.gov](https://datascience.nih.gov)
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**

[https://datascience.nih.gov](https://datascience.nih.gov)
Enhancing Biomedical Workforce Engagement with a Broader Community
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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Findable</strong></td>
<td>• must have unique identifiers, effectively labeling it within searchable resources.</td>
</tr>
<tr>
<td><strong>Accessible</strong></td>
<td>• must be easily retrievable via open systems and effective and secure authentication and authorization procedures.</td>
</tr>
<tr>
<td><strong>Interoperable</strong></td>
<td>• should “use and speak the same language” via use of standardized vocabularies.</td>
</tr>
<tr>
<td><strong>Reusable</strong></td>
<td>• must be adequately described to a new user, have clear information about data-usage licenses, and have a traceable “owner’s manual,” or provenance.</td>
</tr>
</tbody>
</table>
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

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

Release draft for community input (targeting late summer 2019)

Release final policy (early 2020)

Sustainable data management and sharing policy
Overview of Sharing Publication and Related Data

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

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.
Published online 2016 Nov 22. doi: 10.1080/15548627.2016.1256934
PMCID: PMC5324850
PMID: 27875093

Autolysosome biogenesis and developmental senescence are regulated by both Spns1 and v-ATPase

Tomoyuki Sasaki,a,1 Shanshan Lian,a,1 Alam Khan,a,b Jesse R. Llop,c Andrew V. Samuelson,c Wenbiao Chen,d Daniel J. Klionsky,e and Shuji Kishi,a

► Author information ► Article notes ► Copyright and License information Disclaimer

This article has been cited by other articles in PMC.

Associated Data

► Supplementary Materials

1256934_Supplemental_Material.zip
kaup-13-02-1256934-s001.zip (9.6M)
GUID: AC7F9D11-8BE8-402D-9437-6E7942A3ACC6

FAIR Data: Linking datasets to publications (PubMed)
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

Providing FAIR-enabled, open-access options for datasets
Persistent Identifiers and Tracking
Attention, Use, and Reuse

• All submissions have a DOI
  • Supports data citation
  • Usage and citation statistics
  • Other alternative metrics

• Platform and dataset statistics and metrics
  • Openly available
  • Exported to other NIH systems using the API
Where will we be in 1-2 years?

- **Stronger Data Repository Ecosystem**
  - Knowledge of biomedical data repository landscape.
  - Where are the gaps?
  - How generalist repositories fit into the landscape.
  - Useful characteristics of generalist repositories.

- **Strengthen FAIRness of all data repositories**

  - *Why? To make it easier for researchers to more easily share, find, and reuse data*

  - To accelerate research and discovery!
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
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
Examples of Datasets in the STRIDES Cloud

- NHLBI Framingham Heart Study
- All of Us Research Program
- NCI Genomic Data Commons
- NCBI data resources (12 PB!)
- 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!
Opportunities for Data Analytics using STRIDES Cloud

• Large scale metadata search and retrial
• Artificial Intelligence data algorithms at scale; inference of data anomalies for example in gene sequences
• Challenges in large scale compression, data duplication, data quality issues
NIH’s Data Environments are Rich, but Siloed

Connecting NIH Data Systems:
Single method for sign-on and data access across repositories and CSPs
Single ‘Sign-on’ Across NIH Data Resources

• Streamlined login for authorization of controlled-access 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
A Simplified Model for a Distributed World

Before

Federated search exposes more data to more researchers

After

Federated authentication (and account linking), protocol translation services, federated authorization simplify controlled access to more data for more researchers

Federated search exposes more data to more researchers
NIH Access – Authorization Service: Conceptual Overview

1. Authenticated User

2. AuthZ Request

3. AuthZ Lookup

4. AuthZ Response

Identity Sources

- AD
- eRA
- dbGaP
- TBD

User Role 1
User Role 2
User Attr 1
User Attr 2
User Attr 3
User Attr 4
User Attr 5
User Attr 6
User Attr 7

Controlled Access 1
Controlled Access 2

Extended Identity Attributes (if desired)
NIH Is Committed to an Enterprise Auth MVP

We will be working to develop a Minimum Viable Product that addresses three key areas:

1. **Authentication**
   *Establishing or confirming who you are*

2. **Authorization**
   *Verifying what you have access to*

3. **Auditing and Logging**
   *Recording events that have security significance (e.g., logins)*
We Will Take a Standards-Based Approach

A robust, standardized approach to authentication, authorization, and auditing/logging will maximize efficiency and value now and in the future.

- **Industry Technologies & Standards**
  Utilizes industry best practices, technologies, and existing standards.

- **Minimal Re-Engineering**
  Requires a minimal need to reimagine or restructure existing processes and solutions.

- **Flexible to Support Future Standards**
  Looks towards future standards, technologies, and capabilities.

- **Policy Driven Approach**
  Decisions are informed, based, and driven by NIH Data Access 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
FHIR Standard and Application Program Interface

- Developed by Health Level Seven International (HL7), a non-profit organization
- Designed specifically for exchanging electronic health care record data
- For patients and providers, it can be applied to mobile devices, web-based applications, and cloud services
- FHIR is already widely used in hundreds of applications across the globe for the benefit of providers, patients and payers

Modernize the Data Ecosystem

NIH GUIDE NOTICE by July 30th
SBIR/STTR NOSI by August 8
We Generate Enormous Volumes of Data Daily

The intersection of Technology, Computing, and Artificial Intelligence Algorithms

https://people.rit.edu/sml2565/iimproject/wearables/index.html
AI in Biomedicine: Opportunities

- Extract medical information from text in EMRs/EHRs
- Interpret genomic sequence data to understand impact of mutations on protein function
- Monitor sleep and vitals to send information about health at home to doctors
- Determine which calls to child welfare systems warrant deployment of family support and prevention resources to protect at-risk children
- Read medical images and help diagnose diseases like pneumonia and cancer
• New Methods for Image Analysis

• Systems Pharmacology and Drug Efficacy

• Prediction Models, Early Detection and Screening

• Advanced Methods Development (includes deep learning)
AI in Biomedicine: Legal and Ethical Challenges

- No clear rules on consent for data use
- Threats to privacy that can affect generations
- How can people opt out? at the beginning or later on?
- Potential for bias and discrimination
- Use of incomplete or selective data
- Misuse of data
ACD Artificial Intelligence Working Group Members

Rediet Abebe, Cornell
Kate Crawford, PhD, AI Now Institute
Greg Corrado, PhD, Google
David Glazer, Verily (Co-Chair)
Daphne Koller, PhD, Insitro
Eric Lander, PhD, Broad Institute
Lawrence Tabak, DDS, PhD, NIH (Co-Chair)

Michael McManus, PhD, Intel
Barbara Engelhardt, PhD, Princeton
Dina Katabi, PhD, MIT Computer Science & AI Lab
Anshul Kundaje, PhD, Stanford University
Jennifer Listgarten, PhD, Berkeley
Serena Yeung, PhD, Harvard
Charge to the AI Working Group

• Are there opportunities for cross-NIH effort in AI? How could these efforts reach broadly across biomedical topics and have positive effects across many diverse fields?

• How can NIH help build a bridge between the computer science community and the biomedical community?

• What can NIH do to facilitate training that marries biomedical research with computer science?
  • Computational and biomedical expertise are both necessary, but careers may not look like traditional tenure track positions that follow the path from PhD to post-doc to faculty

• Identify the major ethical considerations as they relate to biomedical research and using AI/ML/DL for health-related research and care, and suggest ways that NIH can build these considerations into its AI-related programs and activities
Themes

- more AI-ready data
- more multilingual researchers
  - ELSI: ethical, legal, and social implications
- important areas to apply AI
- important areas to advance AI
ELSI: ethical, legal, and social implications

- Inappropriate use can present real harms, especially to under-represented and marginalized populations.
- Build the guardrails to ensure safety, ethical deployment, and non-discriminatory impacts.
- Set the quality standard, develop more rigorous frameworks around potential harms and challenges.
- Strong oversight and accountability mechanisms for the use of AI in biomedicine.

*These tools have sharp edges -- let’s “do no harm”.*
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

• 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
Enhance the Biomedical Workforce

Coding it Forward

• 9 undergraduate fellows for 2019 placed in NIH Institutes and Centers

• These fellows spend 10 weeks at NIH channeling their computational expertise toward hands-on experience with biomedical data-related challenges

https://www.codingitforward.com/
NIH Data Science Senior Fellowships

• One- or two-year national service sabbatical in high-impact 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 in 2020

• Program evaluation in 2024

COMING SOON
VISION

a modernized, integrated, FAIR biomedical data ecosystem
Enhancing Biomedical Workforce Engaging with a Broader Community
Special Thanks

- **STRIDES**: Andrea Norris, Nick Weber and NMDS team
- **Connecting NIH Data Resources**: Vivien Bonazzi, Regina Bures, Ishwar Chandramoulswaran, 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 Chandramoulswaran 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 Valentine, Roger Stanton, Charlene Le Fauve, Ravi Ravichandran, Zeynep Erim, Derrick Tabor, Rick Ikeda
Stay Connected

@NIHDataScience

/NIH.DataScience

www.datascience.nih.gov
"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov