Session 4: Considerations for Data Generated through the HEAL Initiative
NIH’s Strategic Vision for Data Science: Enabling a FAIR-Data Ecosystem

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

May 17, 2019
VISION

a modernized, integrated, FAIR biomedical data ecosystem
IMAGINE… the ability to link electronic health care records with personal data and with clinical and basic research data.
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 data in the HEALing Communities Study with data on opioid prescribing practices and measures of opioid use in other HEAL studies.
This is the promise of *Data Science at NIH*

...and here’s how we will get there.
Recent Progress Toward NIH’s Vision for Data Science

<table>
<thead>
<tr>
<th>datasets</th>
<th>Data repositories</th>
<th>Data ecosystems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link datasets to publications (PubMed)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provide FAIR-enabled, open-access options for datasets that underly a publication resulting from NIH funded research</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supporting data repositories and knowledgebase resources</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop criteria for open-access NIH data sharing repositories</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Connecting ecosystems</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-priority datasets moved to cloud service providers (CSPs)</td>
</tr>
<tr>
<td>Single method for sign-on and data access across repositories and CSPs</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tools</th>
<th>Communities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engaging with a broader community</td>
<td></td>
</tr>
<tr>
<td>National Science Foundation partnership</td>
<td></td>
</tr>
<tr>
<td>SBIR/STTR utilization</td>
<td></td>
</tr>
<tr>
<td>Hackathons, bug bounties, citizen science challenges</td>
<td></td>
</tr>
<tr>
<td>Software sustainability extension through hardening</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data policies</th>
<th>Workforce</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data management and sharing policy for NIH</td>
<td></td>
</tr>
<tr>
<td>Enhancing biomedical workforce through internships</td>
<td></td>
</tr>
<tr>
<td>Coding it Forward</td>
<td></td>
</tr>
<tr>
<td>Graduate Data Science Summer Program</td>
<td></td>
</tr>
<tr>
<td>NIH Data Science Senior Fellowships</td>
<td></td>
</tr>
</tbody>
</table>
Making Data *FAIR*

<table>
<thead>
<tr>
<th>Findable</th>
<th>must have unique identifiers, effectively labeling it within searchable resources.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessible</td>
<td>must be easily retrievable via open systems and effective and secure authentication and authorization procedures.</td>
</tr>
<tr>
<td>Interoperable</td>
<td>should “use and speak the same language” via use of standardized vocabularies.</td>
</tr>
<tr>
<td>Reusable</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>
Sharing Datasets as Supplementary Materials


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

Tomoyuki Sasaki, a,1 Shanshan Lian, a,1 Alam Khan, a,2,3 Jesse R. Llo, 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: AC7F9D11-8EB-402D-9437-6E7942A3AC06

Link 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

NIH recommends domain-specific repositories when available.

Provide FAIR-enabled, open-access options for datasets that underly a publication resulting from NIH funded research
## The TRUST Principles for Data Repositories

<table>
<thead>
<tr>
<th>Principle</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transparency</strong></td>
<td>is achieved by providing publicly accessible evidence of the services that a repository can and can not offer.</td>
</tr>
<tr>
<td><strong>Responsibility</strong></td>
<td>is a commitment to provide high technical quality data services.</td>
</tr>
<tr>
<td><strong>User community</strong></td>
<td>is the focus on the uses and potential uses of the data and services offered.</td>
</tr>
<tr>
<td><strong>Sustainability</strong></td>
<td>is the capability to support long-term data preservation and use.</td>
</tr>
<tr>
<td><strong>Technology</strong></td>
<td>is the infrastructure and capabilities to support the repository operations.</td>
</tr>
</tbody>
</table>
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)
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
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!

Move/Access to high priority data sets in cloud service providers
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 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
Principles for Data Sharing and Open Access in HEAL Research

Rebecca Baker, Ph.D.
Director, HEAL Initiative
Office of the Director, NIH

May 17, 2019
Considerations for HEAL Data

HHS has declared the national opioid crisis a public health emergency.

Many HEAL projects are funded through cooperative agreements.

Plans for a central data repository for HEAL.

HEAL should leverage ongoing data science innovations at NIH.
Maximizing the Utility of HEAL Research Data

• **Goal: Simple and FAIR data through HEAL**
  • Publications and underlying research data should be made available
    • Any file format
    • Assign an institutionally (NIH) branded DOI
    • Central HEAL or other data repository
  
• Documented with customizable, discipline-specific metadata
  • Enabling research across different HEAL projects

• Discoverable content across major search engines and frameworks
Data Sharing Policy Landscape at NIH

• Projects with budgets > 500K direct costs must submit a plan for data sharing in their applications

• Special considerations for certain types of data and projects, e.g. genomic data, Cancer Moonshot

• Publications resulting from NIH-funded research must be deposited into PubMed Central no later than one year after publication
### Plan for Open Access to HEAL-Funded Publications

<table>
<thead>
<tr>
<th>Incorporate into terms and conditions of certain awards:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rapid deposition of electronic copies of publications in PubMed Central with proper tagging of metadata.</td>
</tr>
<tr>
<td>Publications will be published under the Creative Commons Attribution 4.0 Generic License (CC BY 4.0) or an equivalent.</td>
</tr>
<tr>
<td>Publications will be made publicly available immediately with no embargo period.</td>
</tr>
<tr>
<td>Underlying primary data for the publications will be made broadly available through an appropriate data repository such as the HEAL central data repository.</td>
</tr>
<tr>
<td>To the extent feasible, underlying primary data will be shared simultaneously with the publication and made immediately accessible.</td>
</tr>
</tbody>
</table>
Implementation Plans for HEAL Data Sharing Policy

• Some awards will need to wait until FY2020

• Broad and responsible sharing of data that protects and maintains privacy and confidentiality

• Investigators required to plan for protecting and maintaining privacy rights of participants and confidentiality
Leveraging NIH Data Science Opportunities for HEAL

- **HEAL Central Data Repository**
- Characteristics for NIH-supported data repositories

- Storage of HEAL data
- STRIDES program

- HEAL data *not* in the HEAL Central Data Repository
- Figshare

- “Protected” HEAL data
- single sign-on system