Communicating Data Science Impacts & Accessing Data Science Resources

Kristianna Pettibone, Chris Duncan, Amanda Garton, & Nidhi Gera, NIEHS

National Institute of Environmental Health Sciences

Final NIH Policy for Data Management and Sharing

https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

Key Points:
Effective Jan 25, 2023

The policy requires submission of a Data Management and Sharing Plan outlining how scientific data and any accompanying metadata will be managed and shared, taking into account any potential restrictions or limitations.

Elements of an NIH Data Sharing Management and Data Sharing Plan

- **Data Type:** Briefly describe the scientific data to be managed, preserved, and shared.
- **Related Tools, Software and/or Code:** An indication of whether specialized tools are needed to access or manipulate shared scientific data to support replication or reuse, and name(s) of the needed tool(s) and software. If applicable, specify how needed tools can be accessed, (e.g., open source and freely available, generally available for a fee in the marketplace, available only from the research team) and, if known, whether such tools are likely to remain available for as long as the scientific data remain available.
- **Standards:** An indication of what standards will be applied to the scientific data and associated metadata (i.e., data formats, data dictionaries, data identifiers, definitions, unique identifiers, and other data documentation). While many scientific fields have developed and adopted common data standards, others have not. In such cases, the Plan may indicate that no consensus data standards exist for the scientific data and metadata to be generated, preserved, and shared.
- **Data Preservation, Access, and Associated Timelines:** Plans and timelines for data preservation and access
- **Access, Distribution, or Reuse Considerations:** NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data generated from NIH-funded or conducted research, consistent with privacy, security, informed consent, and proprietary issues. Describe any applicable factors affecting subsequent access, distribution, or reuse of scientific data.
- **Oversight of Data Management and Sharing:** Indicate how compliance with the Plan will be monitored and managed, frequency of oversight, and by whom (e.g., titles, roles).

Allowable Costs for Data Management and Data Sharing

Reasonable, allowable costs may be included in NIH budget requests when associated with:

- **Curating data and developing supporting documentation,** including formatting data according to accepted community standards; de-identifying data; preparing metadata to foster discoverability, interpretation, and reuse; and formatting data for transmission to and storage at a selected repository for long-term preservation and access.
- **Local data management considerations,** such as unique and specialized information infrastructure necessary to provide local management and preservation (e.g., before deposit into an established repository).
- **Preserving and sharing data through established repositories,** such as data deposit fees necessary for making data available and accessible. For example, if a Data Management and Sharing Plan proposes preserving and sharing scientific data for 10 years in an established repository with a deposition fee, the cost for the entire 10-year period must be paid prior to the end of the period of performance. If the Plan proposes deposition to multiple repositories, costs associated with each proposed repository may be included.

Selecting a Data Repository

- For some programs and types of data, NIH and/or Institute, Center, Office (ICO) policy(ies) and Funding Opportunity Announcements (FOAs) identify particular data repositories (or sets of repositories) to be used to preserve and share data. For data generated from research subject to such policies or funded under such FOAs, researchers should use the designated data repository(ies).
- For data generated from research for which no data repository is specified by NIH or the NIH ICO (as described above), researchers are encouraged to select a data repository that is appropriate for the data generated from the research project and is in accordance with the desired characteristics, taking into consideration the following guidance:
  - Primary consideration should be given to data repositories that are discipline or data-type specific to support effective data discovery and reuse. NIH makes a list of such data repositories available (see https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html).
  - If no appropriate discipline or data-type specific repository is available, researchers should consider a variety of other potentially suitable data sharing options:
    - Small datasets (up to 2 GB in size) may be included as supplementary material to accompany articles submitted to PubMed Central (see https://www.ncbi.nlm.nih.gov/pmc/about/guidelines/#suppm).
    - Data repositories, including generalist repositories (see https://www.nlm.nih.gov/NIHbmic/generalist_repositories.html) or institutional repositories, that make data available to the larger research community, institutions, or the broader public.
    - Large datasets may benefit from cloud-based data repositories for data access, preservation, and sharing.

FAIR Guiding Principles

Find out more at www.go-fair.org/fair-principles or review the seminal article The FAIR Guiding Principles for scientific data management and stewardship, Wilkinson (2016).

Findable
The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the **FAIRification process**.

Accessible
Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorization.

Interoperable
The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

Reusable
The ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

Communicating Data Science Accomplishments to NIEHS

NIEHS has made significant investments in data science, data infrastructure, and related activities. We would love to hear about your data science impacts, including activities and accomplishments.

In general, if you are working on or have completed any of the following types of activities, products or impacts, please be sure to describe them in section B2 of your RPPR/FRPPR or email your accomplishments to your program officer.

Scientific Data Management and Stewardship
- Created a formal Data Management and Sharing Plan
- Adopted community standards to be applied to scientific data and associated metadata
- Implemented an enhanced quality assurance and quality control plan across the data lifecycle
- Reconsented participants to broaden approved data sharing
- Created publicly-accessible protocols/guidance for accessing restricted datasets

Data Science Approaches and Tools
- Developed or improved data science tools, workflows, software, and/or code
- Leveraged best practices in software development and advances in cloud computing
- Developed methodologies and infrastructure to support interoperability of environmental health data with other relevant datasets
- Developed and applied standard, unambiguous terminologies, utilizing descriptive terms from existing biomedical ontologies

Data Sharing and Dissemination
- Deposited data to a domain-specific open-access data sharing repository, institutional repository, and/or generalist repository. Please provide persistent identifiers (e.g. DOI, accession number) if available.
- Created or improved infrastructure or platforms for data sharing
- Published articles describing new data science activities, products, or impacts
- Developed or implemented techniques for data visualization and/or dissemination of data to study participants

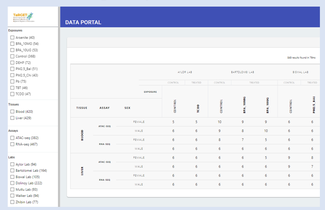
Workforce Development and Training
- Developed, tested, and/or conducted data science-focused courses, trainings, workshops
- Trained students on data science topics
- Formed new collaboration between EHS scientists and data scientists

Challenges/Barriers
You can also describe any challenges or barriers related to data science/data sharing so that program staff can identify possible solutions.

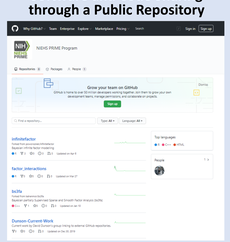
Are there other Data Science related impacts we should be tracking?
Email your ideas to Kristi Pettibone at pettibonekg@niehs.nih.gov

Examples of NIEHS Grantee Data Science Accomplishments

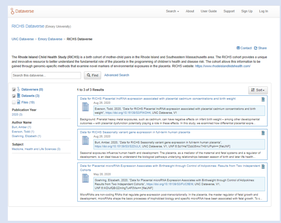
Data Submission to a Domain- or Program-Specific Database



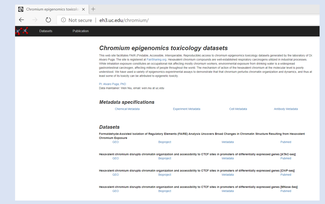
Software and Code Sharing through a Public Repository



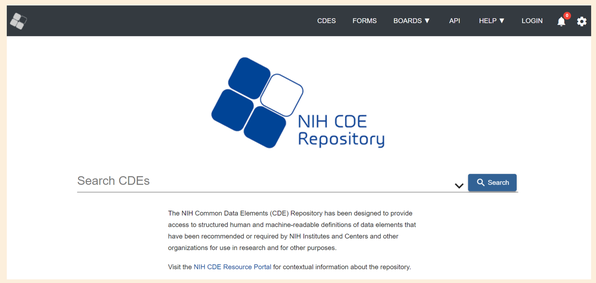
Data Submission to a Generalist Repository



Association of Publicly-Available Datasets with Rich Metadata

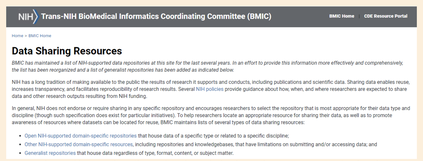


Other Resources



https://cde.nlm.nih.gov/home

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



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

https://www.phenxtoolkit.org/