Medical Library Association Comments to the National Institutes of Health
Re: Request for Information on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research
Notice Number: NOT-OD-19-014
Submitted December 10, 2018

Name of Submitter: Mary M. Langman

Name of Organization: Medical Library Association

Type of Organization: professional org/association

Research Area Most Important to You or Your Organization (e.g., clinical, genomics, neuroscience, infectious disease, epidemiology)*: biomedical research, clinical data

Section 1, Definition of Scientific Data

Metadata:

There are many categories of metadata (e.g., descriptive, preservation, technical, structural, administrative). For this reason, MLA recommends that the policy notes the different categories, and provides recommendations about which to leverage in the case of research documentation efforts.

If electronic lab notebooks are called out as not being an example of scientific data, they could be provided as an example of research documentation or metadata that are required to make the data usable.

Scientific Data:

MLA recommends clarifying at what level the data should be shared/deposited: raw data, processed data, and/or analyzed data. If the data is going to be usable, guidelines should also clarify what else must be made available (e.g., software to process the data).

In order to replicate research findings, researchers need the raw data and the processing code or a software citation. Having examples of what this scientific data includes, along with examples of what scientific data does not include, would be helpful for researchers.

Scientific data should NOT be defined just as those datasets that underlie publications, but also include data outputs from research not necessarily published to avoid bias.

Section 2, Requirements for Data Management and Sharing Policies

Sharing Scientific Data:

If there are perceived barriers to sharing scientific data; guidelines should outline what would be acceptable reasons for not sharing.

Plan Review and Evaluation:

NIH should provide guidelines that include details for how the plan review and evaluation will be accomplished. Guidelines could include training for reviewers and/or the development of a rubric for evaluation. Perhaps a more consistent way would be to have plans reviewed by a set group of NIH personnel, possibly from the National Library of Medicine (NLM) or the NIH Library, with the necessary skill sets for evaluating Data Management Plans (DMP) with a score that is integrated into the evaluation of the entire application.

Plan review and evaluation should affect the overall score of an application and affect the overall success of an application. Otherwise, these will not gain traction or importance and remain afterthoughts to the process for some researchers.

Plan Elements:

A two-page limit will restrict much of the required elements from being described in depth (i.e., data types, related tools and software, data standards, data preservation, access (including timelines) and discoverability, terms for re-use and redistribution, limitations on access, and oversight of data management). We recommend that this limit not be put into place.

Budget:

We recommend requiring researchers to provide reasonable costs associated with data management and sharing under the budget for the proposed project to ensure that researchers account for this cost at the outset. It would be helpful for NIH to provide examples of potential costs associated with each component of the Research Data Management (RDM) plan; examples should include links to the costs repositories might charge for storage.

Support Documentation and Guidance:

NIH should make available a portfolio of successful sample applications and include the DMP. The National Institute of Allergies and Infectious Diseases (NIAID) makes sample applications available (https://www.niaid.nih.gov/grants-contracts/sample-applications#r15) but their samples do not include data management or sharing plan information.

NIH should acknowledge that DMP are living documents that are likely to change throughout the research process, and provide guidelines about how to update and version plans as they change.

Data Types:

A requirement to list the types and estimated amount of scientific data resulting from NIH-funded or supported research should include an estimate of size as many repositories have size limits on both file size and total size of data uploaded. This should be spelled out in the policy as researchers may not equate “amount” with “size” and not include this information.

For scientific data derived from human participants or specimens, applicants should briefly describe the process of de-identification or aggregation they plan to use.

Related Tools/Software and/or Code:

The guidelines should require that all created code and scripts be shared alongside the data in order to make it replicable.

While open source software is preferred, all research software should be well documented, and version and packages used need to be included, especially for reproducibility.

Standards:

MLA recommends including guidance for metadata when identifying standards and ontologies that apply to the scientific data to be collected; e.g. suggested data formats, data identifiers, definitions, and other data documentation.

Data Preservation and Access:

In funding announcements, MLA recommends that institutional data repositories are recognized in funding announcements as potentially acceptable solutions for data deposit, and that appropriate discipline-related repositories are detailed for researchers consideration, such as directing them to use the Registry of Research Data Repositories (https://www.re3data.org/).

Repository selection should be guided by criteria (https://www.datasealofapproval.org/en/information/requirements/) that increase discoverability, guarantee the integrity and authenticity of the data, and have a continuity plan that ensures ongoing access to and preservation of its holdings, rather than the cost of the repository. For this reason, MLA suggests that the following sentence be removed from the provisions: "Investigators would be encouraged to consider using repositories that make scientific data available at no cost for extended periods of use".

We recommend including a statement in the DMP outlining why a certain repository will be selected. Alternatives/suggestions should be provided by reviewers when appropriate if those listed in the DMP are inadequate/not standard for a discipline.

As proposed in the National Library of Medicine's Strategic Plan (https://www.nlm.nih.gov/pubs/plan/lrp17/NLM_StrategicReport_Synopsis_FINAL.pdf) linkages between publications and datasets must occur.

If researchers are not able to deposit datasets into a repository for any reason, we recommend creating a metadata record which describes the data. Many health sciences libraries curate research datasets, regardless of where they are stored (varying from a personal server to repositories), and make them discoverable via a data catalog; e.g. several are funded through the National Network of Libraries of Medicine (https://www.datacatalogcollaborationproject.org/partners/). Research institutions and NIH

should consider this model for insuring the discovery of funded research data especially if the data was not deposited into a repository.

Oversight of Data Management:

We strongly recommend prompting researchers to seek assistance from specialist librarians and information professionals who have expertise in managing data. Librarians are pivotal in facilitating the storage, description, maintenance, preservation, and access to scientific data, and providing expertise, consultations, and outreach to the campus community and others around scholarly resource assessment and metrics (https://www.coar-repositories.org/files/Competencies-for-RDM_June-2016.pdf)

Section 3, Compliance and Enforcement

MLA supports the provision for developing data management and sharing guidelines that support compliance and enforcement of the policy.

Libraries and librarians should play a leading role in coordinating an institution-wide initiative to educate and support the efforts of researchers and others in compliance and enforcement of data management and sharing policies and procedures.

MLA recommends that NIH - in addition to those definitions already provided - direct researchers to existing resource(s) that define additional terms used in this document related to data management and sharing. An example of an existing resource is the National Network of Libraries of Medicine Data Thesaurus (https://nnlm.gov/data/thesaurus).