Office of Science Policy (OSP)
National Institutes of Health
6705 Rockledge Drive, Suite 750
Bethesda, MD 20892

Submitted electronically to:
https://osp.od.nih.gov/provisions-data-managment-sharing/

Re: Notice Number: NOT-OD-19-014 "Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research"

Dear Office of Science Policy, Office of the Director, and Acting Director Bonham,

The University of Massachusetts Amherst Libraries write in response to the Request for Information (RFI) on Proposed Provisions for a Data Management and Sharing Policy for NIH Funded or Supported Research.[1]

The University of Massachusetts Amherst Libraries maintain an institutional repository, "ScholarWorks", at https://scholarworks.umass.edu/data/ . We host at least 68 datasets since our repository began accepting data in 2016. The Libraries' Data Working Group was established in 2010 and has regularly contributed feedback on data management plans, provided data consultations, and instructed scholars in best practices for data management. To build on this work, the Libraries hired a dedicated Data Services Librarian (Thea Atwood, one of the authors of this comment) in 2017 to improve data management capacities and competencies across campus. In that capacity, she works with research groups across campus, as well as interfacing regionally and nationally.

As described in the RFI, the NIH seeks comments on key provisions for a future policy for the management and sharing of data, to replace the 2003 NIH Data Sharing Policy.[2] Specifically, NIH requests public comment on I. The definition of Scientific Data, II. The requirements for Data Management and Sharing Plans, III. The optimal timing for implementation, and any other relevant topic. We write on topics I. and II. and offer some additional commentary.

I. The definition of "Scientific Data".

The proposed definition is:

---

[1] Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research, October 10, 2018, available at https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-014.html.

[2] Final NIH Statement on Sharing Research Data (February 26, 2003), available at https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html .

> Scientific Data: The recorded factual material commonly accepted in the scientific community as necessary to validate and replicate research findings including, but not limited to, data used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens. For the purposes of a possible Policy, scientific data may include certain individual level and summary or aggregate data, as well as metadata.[7] NIH expects that reasonable efforts should be made to digitize all scientific data.
>
> [7] NIH Policy on the Dissemination of NIH-Funded Clinical Trial Information (September 16, 2016) https://grants.nih.gov/grants/guide/notice-files/NOT-OD-16-149.html

We appreciate the thoughtfulness of this definition, which carefully describes the material to be included in broad, functional terms, and equally carefully excludes specific types of communication in appropriately narrow and discrete terms. We have two comments regarding this proposed definition, as well as a suggestion for further study.

First, we recommend including specific reference to **negative results** as potential subjects of inclusion, possibly in the next to last sentence. This sentence could read:

> For the purposes of a possible Policy, scientific data may include certain individual level and summary or aggregate data, metadata, or relevant unpublished data demonstrating negative results.

As many scientists have observed[3], the lack of access to negative results and null data may result in duplicative or misdirected research, distorting assessment of scientific research and hindering the progress of research. The proposed definition properly applies to material "including, but not limited to" data supporting publications, which can be interpreted as including negative data. However, without express mention of data *not* included in publications, such as negative results / null data, researchers will not be prompted to consider preservation of this data. Indeed, due to current conventions of publishing only positive data, researchers may in many cases not feel it appropriate to include negative data without express "permission" to do so.

---

[3] See, e.g., Mlinaric et al, "Dealing with the positive publication bias: Why you should really publish your negative results," *Biochemia Medica (Zagreb)* 2017 Oct 15; 27(3):030201; Weintraub, "The Importance of publishing negative results," *Journal of Insect Science* 2016, 16(1): 109; Matosin et al, "Negativity towards negative results: A discussion of the disconnect between scientific worth and scientific culture," *Disease Models & Mechanisms* 2014 Feb., 7(2): 171-173; and Sandercock, "Negative results: Why do they need to be published?" *International Journal of Stroke* 2012 Jan; 7(1):32-33.

Secondly, we recommend additional detail about the potential **"metadata"** requirements. In particular, the final policy should specify that metadata dictionaries or codebooks, if used or developed, should be made available to the public on similar terms, either published separately or as materials and methods, or included within the data repository for a given proposal.

Third, we note that there is significant ambiguity in the expectation that reasonable efforts should be made to "digitize all scientific data." The modifier "all" in front of the phrase "scientific data", in the context of a definition of scientific data, opens this sentence up to further interpretation: For instance, does "all" mean the kinds of scientific data that were not included in the definition? Moreover, what standards are implied by "digitiz[ing]"? Scanning text without character recognition, for instance, provides an exceedingly minimal and not very useful amount of digitization. Images may be digitized at high quality or low quality, with significant differences in usability. Digital formats may be encrypted, non-standard, or sui generis, and in other ways "digital" but not necessarily useful. We recommend clarifying this language with functional qualifications; for instance, "digitize … in open formats that are usable by researchers" or "digitize … in formats that are broadly available."

Finally, we recommend further investigation of the question of preservation of **laboratory notebooks**. We agree that **laboratory notebooks** are properly excluded from the definition of "scientific data" in the "Data Management and Sharing Policy." Because laboratory notebooks are so idiosyncratic and may include material from many projects, as well as confidential or non-scientific content, it is inappropriate to treat laboratory notebooks as subject to NIH's open data policies.

However, laboratory notebooks often include data, procedures, fundamental research techniques, and observations vital to reproducing research findings. The preservation of laboratory notebooks is therefore of high concern to scientists, both as individual scientists and as managers and principal investigators of laboratories.

Unfortunately, the current proliferation of electronic laboratory notebook software, as well as the use of non-dedicated software including wikis, word processing, spreadsheet, and cloud-based storage, raise fresh concerns about long-term preservation and access to lab notebooks. The very multiplicity of options raises concerns, as the "lab notebook" environment in many laboratories is fracturing, leaving laboratories without consistent lab notebook data. We recommend, therefore, that the NIH study and develop standards for preservation and retention of lab notebooks. Scientists need assistance in understanding how to maintain and preserve lab notebooks into the future, and the progress of science will suffer without attention to this detail.

II. The requirements for Data Management and Sharing Plans.

PLAN REVIEW AND EVALUATION

The proposal to incorporate plan review and evaluation into **Contracts**, after technical evaluation performed by NIH staff, offers the best opportunity for consistent and high-quality assessment. NIH staff would be able to develop relevant expertise in preservation and access, and could connect with peers at other agencies in developing relevant standards and guidelines for repositories, and procedures for preservation, retention, data migration, and other data management policies.

Based on our work with faculty researchers, we are not persuaded that extramural grant reviewers would be consistently well-positioned to assess the acceptability of Data Management and Sharing Plans when reviewing proposals. This is not their area of expertise, generally, and assessment of these plans may well suffer by comparison with assessment of science. We believe, therefore, that assessment of the plans is more appropriately considered part of the technical review. However, extramural reviewers have a key role to play in review and assessment of repositories.

Incorporation of review of plans into reviews by the Scientific or Clinical Director may offer an additional level of review, but should not substitute for a consistent review by staff with relevant expertise.

Finally, we note that evaluation of individual plans in context of other funding / support agreement mechanisms may be appropriate in some cases but should not substitute for routine assessment by staff with appropriate expertise. Some plans, however, may warrant additional review. For instance, proposals to establish a new repository or new method of access may reasonably benefit from *additional review* over and beyond the technical evaluation performed by NIH staff.

PLAN ELEMENTS

Plan Elements Section 2 – Related Tools, Software and/or Code.

Computational methods for generating data must be preserved, and presently, software programs and scripts are made available at any number of private repositories (such as GitHub), which can close at any time. We note that by comparison, in Plan Elements Section 3, the NIH has established a “Common Data Element Resource Portal”. We recommend development of a “Common Software Tools Resource Portal” to connect to significant or recommended software repositories, and recommend assessment of the feasibility of developing an NIH-based software repository.

Plan Elements Section 4. Data Preservation and Access.

We recommend a working definition of “preservation” be included.

We also recommend development of standards for repositories, adoption of privately-developed standards or audits for repositories, or guidance to researchers on how to assess repositories. Researchers are currently not well-prepared to assess data security and management of repositories, including data migration, data security practices, compliance with applicable data laws and standards. While the University of Massachusetts Amherst offers consulting services to our researchers, many institutions are not so well-positioned. In cases where no such on campus resource exists, the NIH should provide guidance on how to assess repositories.

As previously mentioned, staff doing technical reviews for data plans would be well-positioned to collaborate with peers in developing or adopting standards for repositories, assessing repositories for compliance with standards, and certifying them to researchers.

Plan Elements Section 5 – Data Preservation and Access Timeline.

We strongly believe that data preservation and access timelines are an integral part of data management plans. However, researchers rarely know how to assess how long data should be kept; some researchers would prefer to destroy data after the project is over, for instance, while other researchers rely on IRB standards whether appropriate or not in other circumstances. Researchers also rarely receive guidance on a reasonable length of time to distribute data. NSF offers some guidance, suggesting that data should be available “at time of article publication,” but there is little guidance available on sunsetting data distributions.

We therefore agree that this section needs to be included, but recommend that it incorporate reference to additional standards and support from NIH or other federal agencies.

Plan Elements Section 6 – Data Sharing Agreements, Licensing, and Intellectual Property.

We recommend, first, that NIH include recommended sample licenses, such as Creative Commons, in Section 6.3.

Second, in section 6.2, we recommend that the NIH take this opportunity to return to the issue of Material Transfer Agreements (MTAs) that affect access to key reagents necessary to reproduction of scientific results. As an initial matter, the NIH should encourage use of the Uniform Biological Material Transfer Agreement (UBMTA).[4] However, this policy offers an opportunity to revisit the substance of the UBMTA, and consider appropriate limitations on attempts to own scientific data resulting from use of received material. The UBMTA was

[4] Uniform Biological Material Transfer Agreement (UBMTA), *Federal Register*, March 8, 1995.

published in 1995, and in the ensuing two-plus decades, open access and open data principles have become standard – as demonstrated by this RFI.

III. Additional comments

Finally, we recommend that the NIH develop a repository of successfully funded Data Management Plans. Researchers too often develop these plans "in the dark," particularly researchers from institutions that are under-resourced or do not have access to support staff within the libraries with an expertise in data management.

Conclusion

We thank you for the opportunity to comment on this important matter, and hope our comments prove helpful. Please feel free to contact us about any of our comments.

Sincerely,

Marilyn Billings, MLS
Director, Scholarly Communication Department

Laura Quilter, JD, MLS
Copyright and Information Policy Librarian

Thea Atwood, MLS
Data Services Librarian

Erin Jerome, PHD
Institutional Repository and Open Access Policy Librarian

Submitted electronically by:
Laura Quilter
LQuilter@umass.edu