December 10, 2018

Carrie D. Wolinetz, Ph.D.
Office of Science Policy
National Institutes of Health
6705 Rockledge Drive, Suite 750
Rockville, MD 20892

Subject: **Response to 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**

Dr. Wolinetz:

The University of Minnesota writes in response to the Request for Information listed above, published October 10, 2018. As a public land grant institution, we strongly support federal agency policies ensuring public access to scientific research data. Concurrently, we support a thoughtful approach to balancing the need for data access with the administrative burden and cost associated with data management planning, curating, and storing data.

Any policy implemented by the NIH would have direct implications to the researchers we support. In fiscal year 2017, the University of Minnesota received 244 million dollars in funding from the NIH, accounting for 55.6% of our federal research funding.

Specifically, with input gathered from key research committees on campus as well as individual faculty, we would like to respond to the following proposed policy across several major themes:

I. The Definition of Scientific Data

1. **We recommend that NIH adhere to the federal definition of data** in 2 CFR 200.315, which defines research data as:

   (3) Research data means the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This "recorded" material excludes physical objects. Research data also do not include (i) Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is projected under law; and (ii) Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study."

   By adhering to the national definition, NIH will promote commonality and consistency with other

federal agencies, and will facilitate implementation and understanding by researchers, data repositories, journals, and the public alike.

In addition, the federal definition offers a guidepost to what should not be included in the definition, including most importantly data that is preliminary, incidental to research findings, or involving certain objects impractical to share, such as physical specimens or laboratory notebooks. The definition also sends a clear message that HIPAA-controlled data and other data that is subject to privacy laws would be excluded or handled in a way that protects confidentiality in accordance with existing laws. Such protections are a critical element to any data sharing strategy.

2. While the reproducibility issue is of great importance, we note that the scope of that issue goes well beyond data. **Any conversations about expanding the scope of research data to include "necessary to validate *and replicate* (emphasis added) research findings" should be taken only after discussion in context of all of the components necessary for successful study replication**. If the definition of research data needs to be amended, it should be considered through the change management process for alterations to the Uniform Guidance.

3. **We also recommend removing the expectations about digitizing scientific data from the definitions section** (e.g., "*NIH expects that reasonable efforts should be made to digitize all scientific data*.") Instead, asking a researcher to specify what research outcome data is expected to be suitable for deposit into a digital data repository may be a better approach. The cost and effort involved with digitizing an overly broad definition of scientific data may inadvertently subvert the purpose of having public and scientific access to meaningful research data. In addition, the oversight cost for some types of data (including patient data) to ensure that privacy obligations are fully met and that inadvertent disclosure does not occur may be very labor-intensive. Some data (including information in hard copy lab notebooks) may not be worth the cost, curation, and storage efforts involved in being digitized.

4. **Conversely, other more rare information (such as 3d scans of rare physical specimens) that are excluded from the definition may well be worthy of being digitized and funding should be identified for this purpose when access would be deemed valuable**. Alternative approaches, such as NHLBI's BioINCC initiative that makes available "recorded" material about physical objections in a sharable database or inventory of laboratory sample holdings, should also be explored as a cost-effective option.

5. **Commonly adopted and understood standards for metadata and other coding schemes inherent in data collection, organization and management are paramount to the long-term success of this effort**. NIH should continue to invest in development, evaluation and adoption of broadly understood and accepted metadata schemas (such as the NIH common data elements: https://www.nlm.nih.gov/cde/) as a preface to requiring long-term data storage.

II. Requirements for Data Management and Sharing Plans

1. **The proposed requirements for the data management plan in the proposal are extensive. We are concerned about the level of detail requested at time of proposal submission.** We note that approximately 18% of NIH proposals are funded, and believe it is wasteful of researcher time to require detailed information to be embedded in the proposal when most proposals are not funded. As you know, based on the most recent national FDP Faculty Workload Survey, researchers are expending 44% of their available research time on administrative tasks – and that is before the addition of this new significant requirement. **We recommend that the data management plan be required at a later point in the process.** If possible we recommend that the DMP be furnished at Just-in-Time or, if NIH feels that the plan is essential in order for reviewers to have confidence in the investigator's commitment and ability to share data, then the Plan should be given additional importance in the review process (see item #2 below). If a Plan is required at time of original proposal, we recommend that requirements be bifurcated into only those those elements deemed critical for the proposal review (included in the proposal and the proposal budget) and then supplemented either at JIT or at time of RPPR to be expanded/refined. It is critical to mutually develop a mechanism that honors the importance of data sharing while minimizing its burden footprint. Faculty should be directly involved in the solutions developed.

2. Within the proposed plan review and evaluation criteria, the data management plan is currently proposed as an *Additional Review Consideration* for extramural grants. As an *Additional Review Consideration*, the DMP would not be individually scored nor would it influence the overall score, although there is an expectation that compliance with the plan "*would be integrated into terms and conditions as appropriate*" and that NIH staff would engage with potential awardees to modify the plan as appropriate prior to award. Given the extent of the proposed data management plan requirements and enforcement expections, it would appear that the effort and implications of the data management plan are not aligned in their value within the review process. **While we strongly prefer the option listed in Item 1 above, if NIH continues to believe that the DMP is essential at time of full proposal review, we recommend positioning the data management plan as an *Additional Review Criteria*. This would allow the Plan to not be scored individually but still be considered in the overall impact score.**

3. We feel that restricted access should not be the responsibility of the individual researcher. Not only is this administratively burdensome, but it also introduces a dependence on the researcher (and their current contact information) that undermines the goal of long-term data accessibility. NIH should recommend restricted access repositories that provide this level of control for sharing their data.

4. The "Compliance" section sends a strong message. However, it is our experience that data management and sharing is difficult to fully anticipate in detail**. Our researchers require**

flexibility to update and change their DMP as the project progresses. This is especially relevant as data management plans embedded in proposals may not be used (or fully used) until several years later, and repositories and data standards can reasonably be expected to evolve in the meantime. Complications may arise and a DMP should not simply be followed in order to meet requirements NIH mandates to measure compliance. **We recommend that the DMP could be revised annually or as needed (with explanation). Such revisions could be included in the project's annual report.** We note that this could be an excellent opportunity to increase dialogue between researchers and program officials, and to allow the natural sharing of advancements of data sharing mechanisms in a given field.

5. We appreciate the "Oversight of Data Management" section as an explicit component of the data management plan. This highlights the fact that the **management of research data is an active process which requires the long-term investment of resources. To the extent that these resources must be paid by grant budgets, NIH should help develop and widely publicize allowable, reasonable and allocable charging guidelines, and should incorporate the need for such costs into its own budgeting and planning.** Specifically, we note two challenges: (1) many researchers already feel that funding is very limited to support their science, and new "draws" on existing pots of available funding would have the inevitable consequence of reducing the amount available for the direct costs of science; and (2) some of the costs associated with data storage and sharing cannot reasonably be incurred with the period of the grant (or its closeout period). **A new charging mechanism is therefore needed to allow for costs to be either separately funded or, at minimum, set aside to cover these costs. This is a less significant issue if federally supported, non-profit repositories that do not require deposit fees can be made available.** We note that making available non-profit repositories will also facilitate implementation of national data deposit standards and record retention expectations, as well as public access to data. In addition, sustainable national data repositories can be expected to reduce confusion associated with data deposits from multi-site projects, and reduces issues associated with local storage failures (lost or corrupted computer, servers, or other storage facilities) or an inability of a university or other grantee to continue to support its data storage environment.

6. **Of particular importance is the degree of variation in the myriad types and sizes of data that require curation, storage and access**. We have attached a document from Jakub Tolar, Vice President for Clinical Affairs/Dean of the University of Minnesota Medical School that does an excellent job of articulating some of the specific issues associated with the different populations of data such as derived from clinical studies, basic sciences, informatics and "dry lab" data analysis, and large proprietary datasets. We ask that any data management strategy be significantly flexible to address the issues raised by Dean Tolar.

III. Timing for NIH to implement

1. **The opinions and needs of Researchers and Library professionals are critical to this issue, and**

must be sought and understood at a detailed level to properly plan for not only how data can best be accessed, but also appraised, selected, retained, and deaccessioned. We recommend that NIH proceed with sufficient lead time and engagement with the consumers of the data being produced in order to "get it right". Imposition of additional administrative and cost burden on NIH and on grantees should not be undertaken unless there is a high confidence that benefit can be demonstrated to exceed the cost.

IV. Additional Comments

1. In the "Purpose" section, the RFI states that "scientific data… should be managed, preserved, and made accessible in a *timely* manner for *appropriate* use by the research community and the broader public" [italics added]. **We recommend defining the words "timely" and "appropriate" more specifically and including examples of what is (and what is not) timely or appropriate.**
    a. In terms of timeliness, NIH could recommend sharing data X years after the close of the grant, x months after the publication of a paper that uses the data, or x weeks after the data sharing deadline stated in the DMP.
    b. If principal investigators or their delegates are responsible for determining what constitutes appropriate use, this should be clearly stated in the Licenses and Terms of Use section. For example, would PIs be able to place commercial-limiting licenses on their data? Also note that depending on the repository selected, the terms of use may not always be flexible (e.g., blanket repository-wide statement). The researcher should consult with the repository to help them consider what terms, licences are appropriate/available for their data.

2. **In the Preservation section, please link to clear preservation guidelines**. There do not appear to be well-established protocals listed in the NIH Strategic Plan for Data Science. Instead consider more explicit best practices captured in a dynamic way that will stay up to date with the evolving community of practice. One suggestion is the LOC Digital Preservation web site.

3. Preservation is not an inherent attribute of a repository. Specific actions must be taken to preserve digital files and not all repositories do so. Researchers should include a link to the preservation policy of a repository to provide evidence of such actions. A CoreTrustSeal, TRAC, or other trusted digital repository certification could also be evidence of this.

4. Finally, we noted a number of strong elements of the proposed plan and believe these should be highlighted for adoption by other funders (please see Appendix A) .

Finally, we thank you for giving us this opportunity to provide input on this critical topic to the NIH. We look forward to continuing the dialogue on this topic.

Sincerely,

Pamela A. Webb
Associate Vice President for Research

Lisa Johnston
Director, Data Repository for the University of Minnesota (DRUM), University Libraries

Enclosure: Memorandum dated November 30, 2018 from Dean Jakub Tolar, MD, Ph.D

cc with enclosure: Chris Cramer,
Wendy Lougee
Jakub Tolar

Appendix A: Elements in the proposed NIH Data Management and Sharing Policy We Liked

a. S4.4 – We like the suggestion that an alternative preservation plan be included should the original Plan not be achieved.
b. S1.2 – We agree that study protocols and data collection instruments should be considered metadata and shared along with the data.
c. S2.0 – We like that suggestion for justification when choosing a non open source software and highlighting where open source alternatives are available.
d. S4.2 - We like that a "persistent unique identifier" is used, rather than DOI specifically (brand name).
e. S7 - Oversight. We appreciate this section being included and other federal agency policies should follow your lead.

https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf
See section Objective 3-3 | Improve Discovery and Cataloging Resources (page 19 of the PDF)