

10 January 2020

Response to Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

The International Association of Scientific, Technical and Medical Publishers (STM) is the leading global trade association for academic and professional publishers. It has more than 150 members in 21 countries who each year collectively publish more than 66% of all journal articles and tens of thousands of monographs and reference works. STM members include non-profit scientific and scholarly societies, commercial publishers, and university presses who work collectively to ensure broad access to and use of the latest scientific and scholarly information. The majority of our members are small businesses and not-for-profit organizations, who represent tens of thousands of publishing employees, editors and authors, and other professionals across the United States and world who regularly contribute to the advancement of science, learning, culture and innovation throughout the nation. They comprise the bulk of a \$25 billion publishing industry that contributes significantly to the U.S. economy and enhances the U.S. balance of trade.

Publishers sit at the interface between researchers, their research and the rest of the world through our work to improve the quality and availability of information related to research. STM shares our members' commitment to supporting researchers in the sharing, discoverability, and reuse of research data. Individual publishers are developing tools and services to support researchers to make their data FAIR (Findable, Accessible, Interoperable, and Re-usable), and have actively responded to community demand for citation principles for data. STM itself has been involved in numerous projects looking at data access, citation, and preservation, the most recent examples of which have been our recently announced 2020 STM Research Data Year and our ongoing support for the development of [SCHOLIX](#), an easy and universal linking mechanism between scholarly publications and research data.

We therefore welcome the opportunity to comment on the "DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance," published on November 6, 2019, and offer the following as response to the "Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance" (NOT-OD-20-013 / 84 FR 60398). We wish to reiterate our interest – consistent with our commitment to promote sustainable Open Science – in ongoing dialogue with NIH on how to best to promote openness and sharing in research communication. We hope that we can engage further with NIH's Office of Science Policy on these issues over the coming year. Our submission builds on responses that STM has submitted to previous NIH RFIs

on research data and digital repositories, as well as responses that STM has submitted to previous government-wide RFCs on the Federal Data Strategy.

DRAFT NIH Policy for Data Management and Sharing

Section I. Purpose

STM supports the intent of NIH to promote data sharing in order to improve reproducibility and transparency, as well as to improve analysis and enable further discovery. We and our members share NIH's commitment to good data management practices, rooted in community practices and widely-accepted standards. We look forward to working with NIH and the research community to help enable data management and sharing, particularly through our [STM 2020 Research Data Year](#) initiative.

Section II. Definitions

As noted in [our response](#) to the "Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research," STM believes that the definition of data needs to distinguish between data itself and the various interpretations and presentations of data. We do not believe that the definition of "scientific data" in the draft policy is explicit enough. While we appreciate that the definition has been modified to focus on validation and the replication of research findings, the exclusions are not clear enough, or extensive enough to exclude analyses or creative presentations of such data. We therefore recommend that the exclusions be expanded to read: "Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts [or final versions](#) of scientific papers, plans for future research, peer reviews, communications with colleagues, [visualizations](#), or physical objects, such as laboratory specimens."

In addition, we recommend that several other terms be defined to increase the potential implementation and effectiveness of the policy and to help researchers understand and implement steps within their Data Management and Sharing Plans. Specifically, we offer the following terms and potential definitions for consideration:

- **Data Availability Statement:** a statement, often published within an article, that indicates what data are available and how to access them
- **Data Citation:** a reference to the available and relevant data in the reference list, including, where available, a link to the resource
- **Persistent Identifiers (PIDs):** Persistent identifiers assigned to digital objects, such as data sets, in order to make them Findable, Accessible, Interoperable, and Re-usable (FAIR).
- **Data linking:** using PIDs to data sets to create links between articles and datasets, or between datasets and datasets and other related research outputs and artifacts.
- **Data publishing:** making data publicly available and linked to curated information related to the research; especially when an article reporting on research is published, making related data sets available alongside the article, via deposits in trustworthy repositories, making them FAIR by means of linking via persistent identifiers, the inclusion of a Data Availability Statement in the article and proper citation in the reference list.

Section III: Scope

While STM supports better data management and sharing across the research ecosystem, we note that the implementation and extent of policy requirements may vary by the type of funding mechanism. Care must be taken to differentiate between requirements for contract work as opposed to researcher-led grant projects. This is in addition to – and distinct from – the already acknowledged differences between different research disciplines and communities.

Section IV: Effective Dates

STM appreciates that the implementation of the Policy will be dependent upon the willingness and ability of research communities to embrace FAIR principles and work towards greater data sharing. We are engaged in efforts to support both the infrastructure and cultural changes necessary to effect better data management and sharing and look forward to collaborating with NIH and others to accelerate the needed changes. We would welcome additional dialogue on how we can work together to achieve our shared goals.

In this section, and elsewhere in the document, there is a reference to “other funding agreements.” In the context of the scope of the Policy, it would appear that this means “other funding agreements with NIH,” and it would be helpful if that was so clarified here.

Section V: Requirements

STM agrees that researchers should have a plan for managing and sharing data, as appropriate. We also appreciate the flexibility intrinsic in the Policy requirements, including the recognition that there may be restrictions or limitations on sharing.

Section VI: Data Management and Sharing Plans

Good data management and sharing requires planning for long-term preservation and ensuring that data is FAIR, including through the creation of PIDs, proper data citation, and linking between articles and data sets. Although some of these issues are addressed in the “Supplemental DRAFT Guidance,” they are central enough that it would be valuable to include them in the first paragraph of this section, just as data security has been mentioned. For example, where the Policy says “NIH encourages shared scientific data to be made available as long as it is deemed useful to the research community or the public,” it could add “, and Plans should identify mechanisms for long-term preservation, where appropriate.” The first paragraph should also note “Plans should explain how researchers will maximize the discoverability of shared data, through the creation of PIDs, citation, linking, and the like.”

Section VII: Compliance and Enforcement

As noted earlier, STM publishers are working to assist researchers in making data sharing a conscious part of their efforts to communicate the results of their research. As appropriate to the diversity of research communities they serve, journals have a variety of approaches to support data management and sharing. These include creating an explicit data policy, encouraging or requiring data availability statements, ensuring PIDs for shared data sets, providing proper data citation guidelines, and creating

standard processes to link articles and data sets. All of these efforts would help to support implementation of and compliance with any commitments in a Plan.

Supplemental DRAFT Guidance: Allowable Costs for Data Management and Sharing

The costs, both direct and indirect, of good data management and sharing practices can be considerable. STM applauds NIH's recognition that these costs may not be captured in current research practices and its explicit recognition that researchers will need to consider these additional costs in their budgets. We especially appreciate that the document acknowledges that data management and sharing may have ongoing costs and that long-term preservation costs need to be considered.

In addition to the costs listed in this guidance document, STM encourages NIH to explicitly note the costs of the assignment of a persistent identifier, whether directly incurred or as part of the assessment of an appropriate repository. This could be achieved by adding to section 1: "ensuring that the data is deposited at a selected repository that will assign a Persistent Identifier (PID) to each data set, which is endorsed by the research community (DOI's or Accession numbers, etc)."

Supplemental DRAFT Guidance: Elements of a NIH Data Management and Sharing Plan (Plan)

STM's members publish in a wide variety of research areas, each of which has different practices with respect to data collection, use and sharing. The plan requirements must be flexible enough to support the diverse nature of the research that NIH funds, while also providing guidance to all researchers to encourage and enable sharing. With the diversity of data practices and differences in the intensity of data usage across different fields, it may not be appropriate to limit data management plans to two pages in all cases. NIH may want to consider providing the limit as a guideline, or adjusting it in the case of multi-institutional or more complex data plans.

We appreciate that the plan elements described in the guidance are generally flexible and open to interpretation by researchers to best suit their project, consistent with their research community standards. In particular, guidance on the data types and on related tools, software, and code provide appropriate openness to be adaptable to a variety of settings. At the same time, understanding the work involved in the development of good data management and sharing plans and practices, it might be helpful to provide additional guidance to researchers on repositories and practices, perhaps even with reference to a template for the creation of a plan or by providing examples.

We also appreciate the call for the use of standards that are community-endorsed, compatible and interoperable. In fact, this could even be strengthened to clarify that NIH is encouraging the use of standards that meet the two bulleted criteria.

To strengthen the guidance on data preservation and access we recommend several improvements:

- NIH may want to provide guidance to researchers on the criteria for an appropriate and trusted location for data, including plans for perpetual access and a commitment to the FAIR Data

principles. Several initiatives offer certification for or recommendations of trusted data repositories, including CoreTrustSeal (<https://www.coretrustseal.org/>) and Repository Finder (<https://repositoryfinder.datacite.org/about>; <https://www.re3data.org/>).

- NIH may want to strengthen the recommendation on PIDs, which are a widely accepted best practice in data sharing and critical to implementing the FAIR Data principles. Rather than considering “whether a persistent unique identifier or other standard indexing tools will be used,” researchers should consider “which persistent unique identifier or other standard indexing tools will be used.
- NIH may want to consider adding an additional consideration in support of data publishing. For example, “when and how the data will be made part of any article that reports on funded research; indicating whether journals will be considered that have data policies, entertain data availability statements and make data citations part of the reference lists, as well as linking from the published article to the Persistent Identifiers of the deposited data sets in trustworthy repositories.”

Finally, the guidance does not directly address researcher rights to the commercialization of data, which is a key incentive in the research enterprise. NIH may want to consider the use of language that currently appears in the Cancer Moonshot and HEAL Initiative Public Access and Data Sharing policies which explicitly allow researchers to use “licenses that retain intellectual property for commercialization.”

Other Considerations Relevant to this DRAFT Policy Proposal

One of the most significant challenges to better data management and sharing is the current lack of understanding in the research communities we serve of the benefits and motivations for data sharing. In an environment where researchers are under increasing pressure and have limited resources, any additional cost or effort needs to be motivated and aligned with incentives. Here, it can be helpful to note that it is increasingly clear that those that share data have a greater impact. Research has shown that publications with links to shared data receive more citations (see, e.g., Colavizza, G. et al. “The citation advantage of linking publications to research data.” *ArXiv* abs/1907.02565 (2019) <https://arxiv.org/pdf/1907.02565.pdf>).

Publishers stand ready to work with NIH and others to help researchers realize the increased impact for funded research and articles that report on that research. Further, publishers are in a unique position to help drive that change. Studies by a number of publishers that have introduced data policies and data availability statements have shown that such policies have significant impact. For example, when PLOS and BioMed Central introduced Data Availability Statement requirements, authors immediately responded with significant increases in sharing of datasets (see Fig. 2 of Colavizza, G. et al. “The citation advantage of linking publications to research data.” *ArXiv* abs/1907.02565 (2019) <https://arxiv.org/pdf/1907.02565.pdf>). Similarly, Elsevier has reported that the percentage of articles that carry links to deposited data sets increased from about 7% to more than 20 % in the 3 years since it

implemented data availability statements. Publishers are eager to support funders in our common mission to share more data.

STM is already contributing significantly to the development of the standards, resources, policies, and infrastructure needed to enable robust data sharing across the research community, through its involvement in the Research Data Alliance, our own STM [2020 Research Data Year](#), and other initiatives. We welcome further discussion on how NIH, STM, and our member publishers can work together to build greater trust in science and promote the use of research data for the benefit of research and the public. Please feel free to contact me or David Weinreich, Director of Public Affairs in the Americas, for further information.

Very truly yours,

A handwritten signature in black ink, appearing to read "Ian Moss", written over a horizontal line.

Ian Moss
CEO