NATIONAL SCIENCE FOUNDATION
2415 EISENHOWER AVENUE
ALEXANDRIA, VIRGINIA 22314

NSF 19-069

Dear Colleague Letter: Effective Practices for Data

May 20, 2019

Dear Colleague:

Open science principles are increasingly being adopted by industry, government, and academia. Open science gives rise to public benefits by offering broader access to publication, data, and other research materials; broader access enables broader circulation of scientific knowledge, greater return on investments in research data, and more opportunities for replicating and building upon scientific findings.

NSF's open science policy is articulated in the Foundation's Public Access Plan (NSF 15-052) and formally implemented in the NSF Proposal and Award Policies and Procedures Guide and in the Award Terms and Conditions that accompany each award that NSF makes. Implications of this policy are further clarified in an actively-maintained set of Frequently Asked Questions (NSF 18-041).

The purpose of this Dear Colleague Letter (DCL) is to describe — and encourage — effective practices for managing *research data*[1], including the use of persistent identifiers (IDs) for data and machine-readable data management plans (DMPs).

NSF's DMP requirement, as stated in NSF 15-052, expands on NSF's long-standing data-sharing policy. The DMP requirement specifies that every proposal submitted to NSF must include a supplementary document of no more than two pages, titled "Data Management Plan." This document should describe how activities described in the grant proposal will conform to NSF policy on the dissemination and sharing of research results.

As early as January 2013, NSF allowed principal investigators (PIs) to report data products in their biographical sketches. This extension put scientific data sets on a standing equal to traditional publications, such as peer-reviewed journal articles, juried conference papers, book chapters, and monographs.

Putting data in a form that others can use may require work that goes above and beyond a stated research activity. This additional work may be called "data curation" or "data cleaning."

PIs may budget for this expense – that is, they may budget for the work needed to prepare research data for distribution. See the NSF Proposal and Award Policies and Procedures Guide (PAPPG) Chapter II.C.2.g.(vi).b.

In some cases, PIs may have to pay a "data deposit fee" to place data in repositories that then make the data more accessible to others. A "data deposit fee" is a one-time charge paid at the time a dataset is deposited into a data repository. In exchange for this fee, repositories commit to making the data available into the future. NSF has clarified its policies on data deposit fees: these fees are allowable expenses in proposal and award budgets. Specific policies for deposit and length of agreement vary across repositories. Investigators are encouraged to identify such conditions during preparation of their DMPs and should understand that costs and period of retention might be considered during merit review of the DMPs. For more detail on these matters, see PAPPG Chapter II.C.2.g.(vi).b.

Looking forward, open science can be further advanced by two effective data practices: use of persistent IDs for research data, and use of DMP tools that create machine-readable DMPs.

> **Persistent IDs for Data:** Globally unique, resolvable, persistent IDs for research data make the data more findable and accessible, enable citation, and permit linking to data from within publications and other kinds of research presentations. Digital Object Identifiers (DOIs) offer a common example of a persistent ID. Global information trackers (such as Scholix[2]) use persistent IDs in publications or citations to facilitate greater information sharing about data and related materials.
>
> A benefit of a persistent ID for research data is that the dataset can be cited in a researcher's NSF biographical sketch, as previously noted, as well as in the "results of prior research" section[3] of future grant proposals. Use of persistent IDs confer other long-term benefits as well. For example, information about a dataset can be findable even though the dataset itself is no longer accessible.
>
> In a publication reference to a dataset, the citation to the dataset should appear in the body of the article with a corresponding reference in the reference list.[4] Researchers can obtain persistent IDs from their home institutions, repositories or data services in which the data or software are to reside, or other sources.[5]
>
> NSF notes that practices concerning citation of data vary among publishers. However, effective practice suggests that publications should include a statement of data availability that describes how the data underlying the findings of the article can be accessed and used.[6] In special cases where data access is restricted, these restrictions should be mentioned in the statement of data availability. Such restrictions might be determined by applicable laws, university and research

institution policies, funder terms, privacy, intellectual property and licensing agreements, and the ethical context of research.

Machine-readable DMPs: When written effectively, DMPs clarify how researchers will effectively disseminate and share research results, data, and associated materials. However, DMPs can also contain complex and/or ambiguous terms that produce uncertainty about the benefits of data management activities. Such ambiguity can produce situations where the DMP does not adequately explain what data will be created or where the data will be deposited.

For this reason, NSF encourages the use of DMP tools, such as EZDMP[7] or the DMPTool[8], to create machine-readable DMPs. The DMP specifies how data will be produced, prepared, curated, and stored. A machine-readable document allows a computer program to interpret the DMP, such as to prepare a data repository for an eventual deposit of a large or complicated dataset.

A machine-readable DMP, moreover, can be a living document that is modified as the project evolves with documentation of essential attributes of each modification. If there is a change to a DMP during the course of an award, this may need to be brought to the attention of an NSF program officer. See the Frequently Asked Questions for Public Access, question #48, at NSF 18-041.

A benefit of DMP tools for researchers is that they can generate both a PDF version of the DMP that is suitable for inclusion in a grant proposal and a machine-readable version suitable for sharing with an intended recipient data repository or the researcher's home institution.

Movement towards open science by the research community offers the potential to enhance public benefits of science and engineering research. Open science provides new opportunities for researchers to access research findings and data, which in turn has the potential to advance knowledge in many critical domains. The capacity of science and engineering research to achieve these advances depends on the extent to which researchers are aware of, and use, effective practices for data preparation, curation, and distribution. Through this DCL, NSF encourages researchers to learn about the practices described above, and to implement them in the proposals that they prepare for submission to NSF.

CONTACT

For any questions, please contact Beth Plale, bplale@nsf.gov, 703 292 7004.

Sincerely,

Joanne S. Tornow, Assistant Director, Biological Sciences

Jim Kurose, Assistant Director, Computer and Information Science and Engineering
Karen Marrongelle, Assistant Director, Education and Human Resources
Dawn M. Tilbury, Assistant Director, Engineering
William E. Easterling, Assistant Director, Geosciences
Anne L. Kinney, Assistant Director, Mathematical and Physical Sciences
Arthur W. Lupia, Assistant Director, Social, Behavioral, and Economic Sciences
Suzanne Iacono, Office Head, Office of Integrative Activities
Rebecca Keiser, Office Head, Office of International Science and Engineering

---

[1] Per 2 CFR 200.315, effective December 26, 2014, "research data" refers to "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 (e.g., laboratory samples). 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 protected 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."

[2] http://www.scholix.org

[3] See PAPPG Chapter II.C.2.d

[4] See, for example, Enabling FAIR Data, American Geophysical Union (AGU), http://www.copdess.org/enabling-fair-data-project/.

[5] NSF is exploring other services such as the U.S. Department of Energy's Data ID services (https://www.osti.gov/data-services).

[6] Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego, CA: FORCE11; 2014 (https://www.force11.org/group/joint-declaration-data-citation-principles-final).

[7] https://ezdmp.org/index

[8] https://dmptool.org