National Institutes of Health
National Institute of Diabetes and Digestive and Kidney Diseases
Data Management and Sharing Webinar Series
Virtual Meeting

Session 1: Writing a Data Management and Sharing Plan
April 25, 2023

SUMMARY

Welcome and Introductions

Michelle Engle, Research Triangle Institute (RTI) International

Dr. Michelle Engle, Bioinformatics Specialist, RTI International, welcomed the participants to the multipart Data Management and Sharing (DMS) Webinar Series, hosted by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) to educate and connect with the scientific community about the National Institutes of Health (NIH) DMS policy.

Data Management and Sharing (DMS) Overview

Jeran Stratford, RTI International

Dr. Jeran Stratford, Bioinformatician and Data Scientist, RTI International, emphasized the need for a culture shift around data sharing. Research findings typically have been reported in publications and meetings rather than by sharing data, but making the data available provides many benefits. NIH's new DMS policy encourages investigators to include data sharing within the research life cycle. Maximizing data availability increases the utility of research findings by showing how the data were generated and how results can be reproduced. Available data sets can be combined, and questions that previously were not able to be asked can be explored.

The NIH DMS policy applies to all studies that generate scientific data, which NIH defines as the recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings regardless of whether the data are used to support scholarly publications. Additional guidance and examples of what meets the definition of scientific data are available on sharing.nih.gov. The policy requires that researchers both submit a DMS plan with their application and comply with the approved plan, which may be updated over the course of the award.

The DMS submission consists of the DMS plan, which describes how the data will be managed and shared, and a description of the associated costs, for which funding can be requested. The costs are submitted as a line item for DMS in the budget, along with a narrative justification. These are included in the materials available during the peer-review process but are not part of the score. Peer reviewers will comment on whether the requested budget is appropriate to support the activities outlined.

NIH has identified six elements that should be addressed in each DMS plan. Investigators will provide information about the data types that will be generated and how they will be shared, with justification for any data that will not be shared. Links to any related tools, software, and code needed to view, access, or use the data should also be described, regardless of whether the data are published. Investigators also will describe the common data standards—such as ontologies, standardized vocabularies, file formats, and file structures—and what documentation will be provided, such as a study protocol, data analysis plan, or evaluation of quality control metrics, as well as include a schema for metadata to enable interoperability.

Investigators will also describe the data preservation and access plan, including the name of the proposed data repository, unique identifiers that will facilitate the findability and accessibility of the data, and

timelines of when data will be shared and how long they will be available. Investigators also must describe any factors that affect data access, distribution, or reuse. Researchers should maximize the appropriate sharing of their data; however, NIH understands that some ethical, legal, or technical factors may sometimes limit whether the data can be shared and the extent to which some data may be reused.

Research that involves scientific data generated from human research participants should indicate whether and how access to the data will be controlled, and researchers should outline how the privacy and confidentiality of research subjects will be protected. The final element is a description of the oversight of DMS activities, including how compliance with the plan will be monitored and managed within the institution and who is responsible for performing the oversight.

NIH has indicated that DMS plans should be one to two pages, and NIH provides an optional format page on sharing.nih.gov as a fillable, flexible Microsoft Word document that investigators can modify to prepare a plan that best fits their project. The DMS budget should include the estimated direct costs associated with DMS activities, and the justification should summarize the types of data that will be generated and shared, the repositories proposed, and the general cost categories. Unallowable costs include infrastructure costs and costs associated with routine conduct of research.

NIDDK recognizes that investigators have varying levels of DMS experience, and it is committed to making implementation of the DMS policy as smooth as possible by working with investigators as partners. An attitude of learning and adapting will help investigators in a new culture change of data sharing which will allow researchers to realize the maximum potential of their scientific data. NIDDK has developed guidance around DMS plans to supplement the NIH DMS policy. The guidance provides rich and detailed information on each of the six required elements and explains how a researcher can work to make their data as findable, accessible, interoperable, and reusable (FAIR) as possible. Additional topics covered in the guidance include best practices for making custom code and software publicly available, providing instructive documentation about appropriate use of code, using common data elements and machine-readable formats, providing raw and summary data, using persistent identifiers, and making the data available as soon as possible with as few restrictions as possible.

Flexibility will be key moving forward; Dr. Stratford encouraged investigators to reach out to their program officers with questions as they will inform NIDDK if additional resources could be developed to meet those needs. Dr. Stratford highlighted the tools and resources available on the NIH data-sharing website, including a roadmap to lead investigators through the process and help them identify other NIH policies that are relevant to their work, sample DMS plans that span multiple research domains, a repository selection aid, and tools to help investigators learn about common data elements and how to make their data interoperable.

Help From a Research Data Management Librarian
Lisa Federer, National Library of Medicine (NLM)

Dr. Lisa Federer, NLM Data Science and Open Science Librarian and Acting Director, Office of Strategic Initiatives, NLM, explained ways that research data management librarians can help investigators address DMS issues. Many of librarians' core competencies are applicable to DMS, and librarians now can enroll in data services training sessions to engage in services relevant to DMS activities.

The Data Curation Network (DCN), a network of libraries and institutional repositories working to improve data curation and make data more FAIR, provides the services of curators who have specialized knowledge and have worked with relevant types of data sets. A researcher bringing a data set to the DCN will be connected with a curator familiar with that type of data who can provide specialized suggestions and input on how to improve the data set before submission. Curators check the files, read the documentation to ensure its clarity, gather additional information from the submitter if necessary,

improve the submission, transform the format, evaluate the FAIRness of the data, and document the process. DCN also trains additional curators and creates curation primers for various types of data.

NLM's national library network includes seven regional medical libraries that engage with their communities and work with medical and public librarians in their region. The National Center for Data Services prepares the librarian community to support NIH's DMS policy and provides DMS guides and resources for researchers, as well as free training on data-related topics and a summer internship for graduate students to learn about data librarianship. Dr. Federer encouraged attendees to visit the libraries at their institutions to learn more about local data librarians and DMS resources.

The University of California, San Francisco Data Science Initiative—An Institutional Hub for DMS Support
Ariel Deardorff, University of California, San Francisco (UCSF)

Ms. Ariel Deardorff, UCSF Data Sciences Librarian, discussed work at UCSF to implement the NIH DMS policy and inform attendees of what might be available at their institutions. UCSF is a research-intensive, graduate-only health sciences university and the top public recipient of NIH funding. The UCSF library's data science team aims to build foundational skills in programming, statistics, and bioinformatics data engineering and data management. The library serves as a campus hub for education and a support for data science, with the goal of building connections to help UCSF researchers conduct innovative and reproducible data-driven science.

In preparing for the NIH DMS policy, UCSF anticipated questions about how to find and identify the right data repository, understand metadata standards, and determine relevant DMS concerns for research projects. Because some UCSF researchers already had been sharing data, a misalignment between the NIH DMS policy and local guidance was possible. To fill this gap, the library and the Office of Sponsored Research convened a stakeholder leadership group to ensure that the policy and its implications were understood and to prepare for compliance. Local policies were updated to align with the NIH requirements; new processes were developed; and tools, guidance, and outreach materials were designed. The UCSF grants management team also identified control points at which to monitor new DMS plans.

UCSF supports its researchers by connecting with them throughout the life cycle of a research project and answering any DMS questions. UCSF data science librarians conduct education and outreach, as well as one-on-one consultations, to translate complex and dense DMS policy language into easy step-by-step guidance. This requires significant behind-the-scenes coordination with other UCSF departments to ensure that the guidance is understood and that messages are aligned. UCSF's website includes sample plans, templates, and frequently asked questions. Further education and outreach efforts include ongoing training, monthly virtual workshops on the NIH DMS policy, a data management session for graduate students and postdoctoral researchers, and customized training and workshops by request. UCSF libraries also provide information on NIH institute–specific DMS policies and include DMS updates in the library newsletter and UCSF-wide resource emails.

UCSF librarians aim to help researchers navigate a field that often is complex. NIDDK's guidance can help with common knowledge gaps, and Ms. Deardorff recommended that attendees contact their local data librarians to identify any guidance specific to their campus and local experts who can help them navigate the DMS process.

Question and Answer Session

- Dr. Stratford explained that estimating DMS costs in clinical studies with human subjects is difficult; total costs will differ based on the study complexity, outcomes, and other factors. An example of a common element involved in a clinical study that would be appropriate for inclusion in a DMS budget is access to a standardized collection form or instrument to collect the data and increase its interoperability. De-identification of sensitive clinical information will be required, as well as deposition of data in a repository that can support controlled access.
- Dr. Federer encouraged attendees to consult their institutions to determine the librarians' capacity to help with DMS issues before including such assistance in the DMS plan, but she noted that librarians can be included in the funding for the research if investigators propose that a librarian be involved in the research. Ms. Deardorff reiterated that connecting with the library early in the project's development and identifying the level of support the library can provide is important.
- In response to a request for detailed concrete examples of data that should not be public, Dr. Stratford explained that specific types of data are particularly sensitive. Different policies often apply to Tribal, American Indian, and Alaska Native data, and special considerations may be needed. He encouraged attendees to review a recent two-part NIH webinar that discussed data privacy concerns. Ms. Deardorff pointed out that UCSF's website lists data types that are justifiable not to share. For example, data related to abortion would be sensitive in some locations, and even de-identified data could put people at risk. In such cases, a DMS plan still would be submitted, but the plan would describe why these data are not being shared. Dr. Stratford added that businesses generating data with intellectual property could also restrict sharing by law for a period of time. However, those research proposals are still required to submit a DMS plan and will outline in the plan when data will be shared. Finally, some contractual arrangements may prohibit sharing data. These must be described in the DMS plan so that NIH can evaluate and determine if the restrictions are acceptable.
- When asked how NIH is supporting institutional librarians so that they can engage more effectively with researchers, Dr. Federer explained that such support is primarily through training and providing opportunities for them to gain skills and expertise. NLM offers some small grant funding that librarians could pursue for specific activities.

Adjournment

Dr. Engle thanked the panelists and invited attendees to return for the second webinar in the series, which will focus on finding a repository for the data.