An Analysis of NSF Data Management Plan Guidelines

Yubing Tian[1], Anastasia Bennett[1], Will Sutherland[1], Amy Ferguson[1], Morgan Ford[1], Jessica Nga-Kei Li[1], Elizabeth Yarbrough[1], Megan Finn[1], and Amelia Acker[2]

[1] University of Washington
[2] University of Texas at Austin

Abstract. This poster reports on ongoing research into the National Science Foundation's Data Management Plan guidelines and its impact on science data lifecycles. We ask two research questions (RQs): 1) How does guidance about the formulation of DMPs vary across different research areas? And 2) How has guidance about the management of data changed since the first DMP policies were published in 2011? To this end, we collected, examined, and compared 37 DMP guidance policies from 15 different research areas. We identify the following three themes during document analysis: 1) Responsibility for the future of data; 2) Data maintenance changes over time; and 3) The use of data repositories. Based on these preliminary findings we believe that National Science Foundation guidance policies represent a unique view into changes in data management practices over the last decade.

Keywords: Data Management Policies · Scientific Data · Data Afterlives · Data Repositories

1 Introduction

Since 2011 the United States of America's National Science Foundation's (NSF) has required a Data Management Plan (DMP) of all project proposals, and the agency has provided specific DMP guidance policies to researchers about how to generate a management plan, share, preserve, and properly standardize data from research for reuse. These guidance documents are specific to the agency's seven research fields and subfields, of which many have been updated over the past decade. This poster reports on ongoing research into the impact of the DMP requirement on science data lifecycles.

In this study, we collected, examined, and then compared 37 DMP policies from 15 different research areas. We analyze the DMP requirements to address two RQs:

1. How does guidance about the formulation of DMPs vary across different research areas?
2. How has guidance about the management of data changed since the first DMP policies were published in 2011?

In our findings, we report on and address three major themes: First, we find that many of the DMP guidelines give research project leaders, or principal investigators (PIs), great flexibility in defining what counts as data and what appropriate metadata and standards should be used in data management, whilst also pushing that responsibility to individual PIs. Second, different layers of temporality appear in descriptions of data maintenance planning: there are substantial variations between the different research fields and the imagined afterlives of data beyond the end of the projects. Third, we note that starting in 2014, DMP guidance policies started to recommend working with general or field-specific repositories or data infrastructures for long-term data preservation.

2 Related Literature

The impetus for the NSF's 2011 requirement for DMPs [1] came from the National Science Board (NSB), an advisory board to the NSF. The NSB believed that sharing large datasets would be essential to the future of data-driven science increasingly underpinned by federally funded complex cyberinfrastructures and computing technologies [2]. These DMP data policies were also structured to help the NSF achieve compliance with President Obama's "Open Government Directive" announced in 2009 [3]. Following the NSF's example, many other federal agencies, and non-federal science organizations such as the National Academies of Science, now require DMPs [4]. The NSF DMP mandate addresses both data access, sharing, management and archiving, and argues that such a policy would enfranchise data sharing amongst communities of practice across federal funding agencies as well as scientific domains [5]. DMPs are required with all NSF grants; NSF grant proposal guidelines describe what should be addressed in DMPs [6], including types of data, standards to be used for data and metadata, plans for archiving data over the long term, and policies for access and reuse.

However, recognizing the vast number of fields funded by the NSF, it has not mandated that all DMPs follow the same guidelines. Different research divisions have published DMP guidance specific to researchers in that area. Like lab notebooks, cell cultures, and particle accelerators, we read DMP requirements as tools and evidence of different data cultures. We understand data cultures as epistemic communities that are brought together by the circulation of knowledge in the production, cultivation, collection, and use of data [7].

In her study of data sharing practices and rationales for sharing research data, Borgman analyzed the DMP mandate, arguing that "NSF sidesteps the definition of data" [8]. Borgman notes that the meaning of "data" across domains can be defined by its source, scope, or purpose in the study, handling, method of collection, or the way it is used in the research. The definition of data will impact a community of practice's perception of the data lifecycle and most importantly, what is meant by data sharing and norms for preservation and access: "Generally, the more handcrafted the data collection and the more labor-intensive the postprocessing for interpretation, the less likely that researchers will share their data" [8]. In Pasek's comprehensive 2017 analysis of the historical context of the NSF DMP policy and its implications for data curation efforts, she found that the NSF DMP policy "place(s) the burden of determining what specifically should be covered in a DMP onto the researchers who are seeking funding, rather than providing more detailed guidance" [9]. In this study, we build on previous work by conducting a close analysis of NSF's DMP policy documents, paying attention to longitudinal trends and variations between different directorates.

This research contributes to understandings of DMP requirements on scientific data lifecycles.

3 Research Design and Methodology

Our goal in analyzing DMP policies was to 1) analyze how what might be considered data, research practices, and data cultures vary across different fields as articulated in DMP policies and 2) examine how DMP policies might have changed over time. For our data collection, in order to review all iterations of the DMP policies across directorates and divisions, we divided their review amongst the paper co-authors. Using Table 2 in Pasek [9] as a starting point, the co-authors each took one directorate and found the current DMP guidance documents. Subsequently, the Internet Archive's Wayback Machine was used to find all past versions of DMP policy documents embedded or linked at the directorates' guidance policy URLs between 2011 and 2020. For some directorates, this process also involved collecting the DMP policies of each division within that directorate. For example, divisions within the Directorate for Mathematical and Physical Sciences (e.g. the Division of Astronomical Sciences and the Division of Physics) had multiple versions of their DMP policies since 2011. Prior versions of DMP guidelines found through the Wayback Machine did not consistently include publication dates. In their absence, we assigned the year in which they were captured by the Internet Archive.

In total, we reviewed 37 different DMP guidance policies (see Figure 1, Table 1). Each DMP policy was manually coded by one co-author for themes derived from our literature review including: discussions of data and "data product", standards, metadata, data retention, data infrastructure and repositories, and other data or non-data. We wrote memos summarizing observations about patterns or changes over time in each directorate and variations across directorates and discussed them. Based on our initial document analysis and discussions, we identified three themes (discussed in the next section) for further development. Teams of 2-3 researchers developed memos around themes based on the coded DMP policy data.

4 Findings

4.1 Responsibility for Data

We refer to responsibility as a shorthand for asking who bore the onus to decide what constitutes data and metadata. We found that different directorates'

Fig. 1. Data Management Plan Policies Reviewed Across National Science Foundation Directorates.

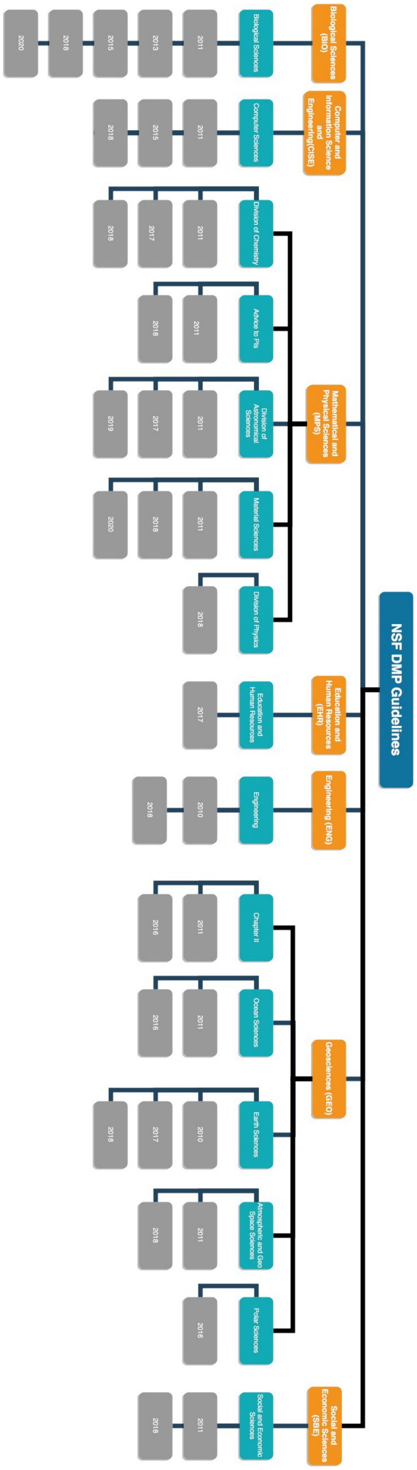

Table 1. Data Management Plan Policy Document Titles.

NSF DMP Policy	Document Title	NSF DMP Policy	Document Title
BIO 2011	Directorate for the Biological Sciences [BIO] Information about the Data Management Plan Required for all Proposals (6/15/11)	EAR 2018b	Division of Earth Sciences Data Policy Appendix
BIO 2013	UPDATED Information Information about the Data Management Plan Required for all Proposals (2/20/13)	AGS 2011	Division of Atmospheric and Geospace Sciences [AGS] Advice to PIs on Data Management Plans
BIO 2015	Directorate for Biological Sciences [BIO] UPDATED Information about the Data Management Plan Required for Full Proposals October 1, 2015	AGS 2018	Division of Atmospheric and Geospace Sciences Advice to PIs on Data Management Plans
BIO 2018	Directorate for Biological Sciences UPDATED Information about the Data Management Plan Required for all Proposals	OPP 2016	Dear Colleague Letter: Data Managment and Data Reporting Requirements for Research Awards Supported by the Office of Polar Programs
BIO 2020	Directorate for Biological Sciences UPDATED Information about the Data Management Plan Required for all Proposals September 2020	MPS 2011	Directorate of Mathematical and Physical Sciences: Advice to PIs on Data Management Plans
CISE 2011	Data Management Guidance for CISE [Computer and Information Science and Engineering] Proposals and Awards	AST 2011	Directorate of Mathematical and Physical Sciences Division of Astronomical Sciences [AST] Advice to Principal Investigators (PIs) on the Data Management Plans
CISE 2015	Data Management Guidance for CISE Proposals and Awards	AST 2018	Directorate of Mathematical and Physical Sciences Division of Astronomical Sciences (AST) Advice to PIs on Data Management Plans, January 17, 2018
CISE 2018	Data Management Guidance for CISE Proposals and Awards	AST 2019	Directorate of Mathematical and Physical Sciences (MPS) Division of Astronomical Sciences (AST) Advice to Principal Investigators (PIs) on the Data Management Plan (DMP)
EHR 2017	Data Management for NSF EHR [Education and Human Resources] Directorate Proposals and Awards	CHE 2011	Directorate of Mathematical and Physical Sciences Division of Chemistry (CHE) Advice to PIs on Data Management Plans
ENG 2010	Data Management for NSF Engineering Directorate [ENG] Proposals and Awards	CHE 2017	Division of Chemistry UPDATED Advice to Principal Investigators on Data Management Plans June 26, 2017
ENG 2018	Directorate for Engineering Data Management Plans Guidance for Principal Investigators	CHE 2018	Division of Chemistry Advice to Principal Investigators on Data Management Plans January 2, 2018
GEO 2011	Chapter II - Proposal Preparation Instructions	DMR 2011	Directorate of Mathematical and Physical Sciences Division of Materials Research (DMR) Advice to PIs on Data Management Plans
GEO 2016	Chapter II - Proposal Preparation Instructions	DMR 2018	Directorate of Mathematical and Physical Sciences Division of Materials Research (DMR) Updated Advice to PIs on Data Management Plans January 16, 2018
OCE 2011	Division of Ocean Sciences Sample and Data Policy	DMR 2020	Directorate of Mathematical and Physical Sciences Division of Materials Research (DMR) Updated Guidance to PIs on Data Management Plans March 18, 2018
OCE 2016	Division of Ocean Sciences Sample and Data Policy	DMS 2018	Directorate of Mathematical and Physical Sciences Division of Mathematical Sciences (DMS) Advice to PIs on Data Management Plans
EAR 2010	Division of Earth Sciences [EAR]	PHY 2018	Directorate of Mathematical and Physical Sciences Division of Physics (PHY) Advice to PIs on Data Management Plans January 2, 2018
EAR 2017a	Data and Sample Policy Division of Earth Sciences National Science Foundation	SBE 2011	Data Management for NSF SBE [Social and Economic Sciences] Directorate Proposals and Awards
EAR 2017b	Division of Earth Sciences Data Policy Appendix	SBE 2018	Data Management for NSF SBE Directorate Proposals and Awards
EAR 2018a	Data and Sample Policy Division of Earth Sciences National Science Foundation		

policies defined data differently and provided varying guidance for data management, sharing, and re-use. PIs were often given guidelines for how to determine if something is data, and ultimately PIs were responsible for deciding what was data in their projects. Deciding what counts as data and metadata have consequences for which data get preserved.

Although the responsibility for critical decisions regarding data management was relegated to PIs, key terms such as data and metadata were inconsistently explained in DMP policies. Guidance around definitions of data included the following:

1. "[T]he recorded factual materials commonly accepted in the scientific community as necessary to validate research findings" (BIO 2011, 2013; CISE 2011, 2015, 2018; ENG 2010; SBE 2011, 2018),
2. Providing examples of what could constitute data (AST 2011, 2018, 2019; CHE 2011, 2017, 2018; EHR 2017; ENG 2010; OCE 2011),
3. Defining what would not constitute data e.g. trade secrets (BIO 2011, 2013; SBE 2011, 2018),
4. No definition provided due to shifting norms (BIO 2018) [10].

While some directorates' definitions of data became more specific over time, others became more nebulous.

If data was defined unevenly, metadata remained largely undefined. Only three DMP policies provided examples for the kinds of data that required metadata (CISE 2011, 2015, 2018) [11]. The policies from Mathematical and Physical Sciences (MPS) required that metadata be "useful" (AST 2011, 2018, 2019; DMR 2018, 2020; MPS 2011). Some DMP policies noted that data and metadata were left undefined to facilitate cross-disciplinary collaboration, as less specific guidelines were envisioned to allow PIs more flexibility to work across different disciplines and collect data that may be considered atypical for their current discipline (BIO 2013; DMS 2018; EHR 2018).

4.2 Data Maintenance

We found themes of long-term planning for data maintenance in the DMP guidelines including how long data is to be kept, the cost of it being kept, and what the goals of the maintenance are. We found three long-term goals for data maintenance:

1. Retaining data to replicate published findings (DMR 2011; PHY 2011; 2018),
2. Making data publically accessible for reuse in future research (AST 2018, 2019; BIO 2015, 2018, 2020; CHE 2019; CISE 2011, 2016, 2018; DMR 2020; EAR 2010, 2017, 2018; ENG 2010, 2018; OCE 2011, 2016),
3. Managing future derivatives of data (AST 2019; BIO 2015, 2018, 2020; CHE 2018; CISE 2011, 2016, 2018; DMR 2020; ENG 2018; OCE 2011, 2016).

The goals for data management strongly shaped the discussion of future data management practices. Terms used throughout the DMP policies, such as "reuse", "access", "curating", "archiving", "maintenance", and "preservation" had different implicit data maintenance needs in regards to the timeframes PIs needed to plan for and the resources needed to manage data. For instance, some 2011 policies were concerned only with supporting published results and provided little discussion of different timeframes for data management (DMR 2011; PHY 2011). In contrast, policies which explicitly addressed the goal of reuse also delineated time frames such as how quickly PIs should provide public access (within a "reasonable" or "timely" timeframe), maintenance and access during the grant period, points at which the data became "final", after publication, and maintenance and access after the grant period ended (e.g., AST 2018, 2019; CHE 2018; DMR 2020; EAR 2010, 2017, 2018).

The temporalities of data management also had a close relationship with data infrastructures. As the DMP guidelines developed, responsibility for long term preservation and access moved from PIs to third-party repositories. As a result, data temporality concerns around long term and post-award maintenance were shifted to the more standardized practices of data infrastructures instead of the vernacular and ad-hoc practices of PIs.

4.3 The Turn to Data Repositories

All DMP policies emphasized data management for dissemination, but some directorates' policies specifically engaged with the afterlives of data by discussing tactics for re-use, re-distribution, or the status of derivatives from data collections (BIO 2018; ENG 2018). Data infrastructures refer to the systems and tools related to the retention, accessibility, and preservation of research data for the long term. After 2014, there was an increase in DMP guidelines that directed PIs to data infrastructures, commonly in the form of data repositories, data centers, and databases (AGS 2018; AST 2018, 2019; BIO 2015, 2018, 2020; CHE 2017, 2018; CISE 2015, 2018; DMR 2018, 2020; EAR 2017, 2018; ENG 2018; OCE 2016; OPP 2016; SBE 2018).

Although later versions of most DMP policies began to list data infrastructures for PIs to consult, they ultimately left it to the PI's discretion on where to store their data, as long as the retention plan for that data was clearly stated. Policies across directorates avoided taking a prescriptivist stance to allow PIs to find the most suitable institutional or community-approved repository for their data. Some policies offered resources, such as a repository registry (and search) service re3data.org (CHE 2018; ENG 2018), or toolkits from the Ecological Society of America (BIO 2015). Other DMP policies provided specific repositories through an appendix that listed suggested, recommended, and/or approved data repositories (CHE 2011, 2017, 2018; EAR 2010, 2017b, 2018b; ENG 2018; OCE 2011, 2016; OPP 2016). In turn, data standards and metadata became tied to the requirements set by third-party data infrastructures.

We also found that different NSF directorates ascribed diverse values to retaining and preserving datasets, samples, or other kinds of data within data infrastructures. The reasons provided included timely public access; reproducibility, interpretability, and transparency of scientific studies; education; and international treaty mandates (see Table 2).

Table 2. Values ascribed to retaining and preserving data within data infrastructures.

Values	DMP Policies
Timely public access	AGS 2011, 2018; AST 2011, 2018, 2020; BIO 2011, 2013, 2015, 2018; CHE 2018; ENG 2010, 2018; GEO 2011; OCE 2011, 2016; OPP 2016;
Reproducibility, interpretability, and transparency of scientific studies	DMR 2011; ENG 2018; PHY 2018
Education	BIO 2015, 2018, 2020
International treaty mandates	OPP 2016

5 Conclusion

Through document and thematic analysis, we analyzed 37 DMP policies in 15 research areas to examine how NSF DMP policies have changed throughout time, varied across research areas, and its impact on science data lifecycles. This research helps to uncover the ways that broad values around data sharing and preservation are articulated in specific policies and in turn how such policies might shape information management, scientific practice, and funding. Returning to RQ1, the results of this preliminary study indicate that pPolicies varied across directorates in terms of guidance provided surrounding DMPs. Notably, policies allowed PIs to determine what constitutes data and metadata, and how PIs should maintain and preserve their research data. Despite standard data policy goals for the NSF, different directorates' DMP data policies are varied in their contents or goals because different scientific domains will necessarily have different cultures of data sharing and modes of providing access. Returning to RQ2, we found that as the DMP guidelines changed from 2011 to 2020, the responsibility of long-term data preservation shifted from PIs to third-party repositories. Based on these findings we believe that guidance documents represent a unique view into how the NSF shapes data management expectations over time. Future research will examine DMPs to see how they relate to long-term data management, access, and re-use of archived scientific research data. While this research focused on the NSF, further research could examine if these patterns are consistent across funding agencies in the USA and in other countries.

References

1. National Science Foundation, Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans', News Release, May 10, 2010. `https://www.nsf.gov/news/news_summ.jsp?cntn_id=116928` (accessed Jan. 14, 2020).

2. National Science Board, Committee on Strategy and Budget, Task Force on Data Policies, "Statement of Principles." National Science Foundation, Feb. 16, 2011, [Online]. Available: https://www.nsf.gov/nsb/committees/archive/dp/principles.pdf.
3. M. Gascó-Hernández, E. G. Martin, L. Reggi, S. Pyo, and L. F. Luna-Reyes, "Promoting the use of open government data: Cases of training and engagement," Government Information Quarterly, vol. 35, no. 2, pp. 233–242, Apr. 2018, doi: https://doi.org/10.1016/j.giq.2018.01.003.
4. E. National Academies of Sciences, Data Management and Governance Practices. Washington, D.C.: The National Academies Press, 2017.
5. Interagency Working Group on Digital Data, "Harnessing the Power of Digital Data for Science and Society: Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council," National Science and Technology Council, Washington, D.C., Jan. 2009. Accessed: Dec. 26, 2019. [Online]. Available: `https://www.nitrd.gov/About/Harnessing_Power_Web.pdf`.
6. National Science Foundation, "Proposal & Award Policies & Procedures Guide," Jan. 25, 2016. `https://www.nsf.gov/pubs/policydocs/pappguide/nsf16001/` (accessed Dec. 26, 2019).
7. T. Clement and A. Acker, "Data Cultures, Culture as Data," CA, 2019, doi: https://doi.org/10.22148/16.035.
8. C. L. Borgman, "The conundrum of sharing research data," Journal of the American Society for Information Science and Technology, vol. 63, no. 6, pp. 1059–1078, 2012, doi: https://doi.org/10.1002/asi.22634.
9. J. E. Pasek, "Historical Development and Key Issues of Data Management Plan Requirements for National Science Foundation Grants: A Review," Issues in Science and Technology Librarianship, 2017, doi: https://doi.org/10.5062/F4QC01RP.
10. National Science Foundation, "Directorate for Biological Sciences UPDATED Information about the Data Management Plan Required for all Proposals." Feb. 2018.
11. National Science Foundation, "CISE Data Management Guidelines," Apr. 26, 2018. `https://www.nsf.gov/cise/cise_dmp.jsp` (accessed Jan. 26, 2020).