Evaluating Tools for Data Management Plans: A Comparative Study of the DART Rubric and the Belmont Scorecard

Sarika Sharma[1(✉)], Arlo Obregon[1], Zahir Shaikh[1], Yubing Tian[2], Megan Finn[2], and Amelia Acker[1]

[1] University of Texas-Austin, Austin, TX 78712, USA
sksharma2@utexas.edu
[2] University of Washington, Seattle, WA 98195, USA

Abstract. Data management plans (DMPs) are required from researchers seeking funding from federal agencies in the United States. Ideally, DMPs disclose how research outputs will be managed and shared. How well DMPs communicate those plans is less understood. Evaluation tools such as the DART rubric and the Belmont scorecard assess the completeness of DMPs and offer one view into what DMPs communicate. This paper compares the evaluation criteria of the two tools by applying them to the same corpus of 150 DMPs from five different NSF programs. Findings suggest that the DART rubric and the Belmont score overlap significantly, but the Belmont scorecard provides a better method to assess completeness. We find that most DMPs fail to address many of the best practices that are articulated by librarians and information professionals in the different evaluation tools. However, the evaluation methodology of both tools relies on a rating scale that does not account for the interaction of key areas of data management. This work contributes to the improvement of evaluation tools for data management planning.

Keywords: Scientific data management · Evaluation tools · Assessment

1 Introduction

The management of research data impacts the trust in and efficacy of science. The management of research data also creates possibilities and limitations for data futures. The perceived importance of research data management is embedded into science policy. Over the last decade, federal funding agencies in the United States, such as the National Science Foundation (NSF) and the National Institute of Health (NIH), have required research proposals to include data management plans (DMPs). A DMP states how data will be managed from federally funded grant projects. The DMP requirement by federal agencies aims to make outputs (whether data or publications) from taxpayer funded research openly available and accessible. But how effectively DMPs enact this vision of science is less understood by information scientists, scholarly communication researchers, and policymakers alike.

I. Sserwanga et al. (Eds.): iConference 2023, LNCS 13972, pp. 26–46, 2023.
https://doi.org/10.1007/978-3-031-28032-0_3

To address this gap, librarians and information professionals have developed evaluation tools to assess DMPs. The Data Management Plan as a Research Tool (hereafter the "DART rubric") and the Belmont Scorecard (hereafter "the Belmont score") are two tools that assess the "completeness" of data management. The DART rubric was developed in 2016 by professional librarians funded through a National Leadership Grant LG-07-13-0328 by the Institute of Museum and Library Science (IMLS) [8]. The Belmont score was formed two years later by the Belmont Forum with the goal to improve DMPs of funded projects around climate change [1].

Both tools provide an assessment of the technical aspects of data, management, access, and preservation in a DMP. Each tool assesses statements in DMPs by examining how well those statements reflect best practices for data management. Best practices from the LIS community address the key activities that determine the access to data and enables the interpretation of data. Thus, assessment tools examine DMPs for evidence of activities that enhance data access such as formatting, versioning, documentation such as metadata and data provenance, and technical systems for storage [2]. Assessing DMPs using existing, well-researched, best practices can provide insights into whether a DMP shows that a proposed science project can enable data to move from one evidentiary context to another – a shared interest across science funding agencies, scholars, and library professionals who want to promote open science.

Evaluation tools are vital for understanding whether DMPs can help funding agencies achieve their goals of promoting open science. This paper contributes to the small body of knowledge about DMP evaluation tools. Previous research has been published by researchers who developed the evaluation tools, providing vital insights into the formation of the metrics of analysis that constitute the rubric [1, 3, 10] The growing body of knowledge also informs the quality of DMPs. Studies about the DART rubric analyzes data management practices in a specific NSF program by researchers within the same universities [10] and compare DMPs across NSF programs and across different universities [1]. To date, no studies have drawn on a longitudinal corpus of DMPs; evaluated the Belmont Score; or conducted a comparative study of the tools and their underlying mechanisms of evaluation. Such an interest is motivated by practical matters of data management and the social studies of data. Evaluation tools are reflections of the organizational goals, best practices, and the current state of data management cultures. The purpose of this paper is to interrogate the tools to examine the possible outcomes of evaluation. This research is motivated by two big questions: What is the best tool to evaluate DMPs with? And what can evaluation tools tell us about DMPs? In order to speak to these broad, contextually specific questions, we have broken the question into two research questions: RQ1) What criteria are part of the DART rubric and the Belmont Score to assess DMP statements? RQ2) How do the DART rubric and the Belmont Score assess the same statements made in 150 DMPs from five programs at the NSF?

We specifically respond to the extant literature by reporting: the criteria of evaluation from the two tools; how they overlap and how each tool evaluated the same 150 DMPs; the methods we used to explore the comparisons between the DART Rubric and Belmont Score; and the results of the comparison. Results suggest that the content of the DART rubric and the Belmont score are similar. However, the evaluation of the 150 DMPs scored better in the DART rubric compared to the evaluation of DMPs by the Belmont

score. We end the paper with a discussion on the strengths and weaknesses of each tool; the pros and cons of specific evaluation techniques; a recommendation on which tool is ideal for assessment; and lastly data management practices across five NSF programs.

The contributions of this work are three-fold. First, this work provides a methodological contribution to assess evaluation tools to improve the criteria for evaluating DMPs. Second, it provides a guidance to policymakers and other evaluators of which tool (s) may be most efficacious to adopt for evaluating DMPs. Third, it provides insights into the data cultures across five NSF programs that represent different scientific fields. The contributions provide future follow-up work, specifically the question of how evaluation tools can better assess data cultures. In future work, we plan to qualitatively assess DMPs to examine data cultures.

2 Literature and Background

Increasingly federal agencies that fund academic science are focused on open science, or the processes to make science outputs available. The DMP is one way that the NSF has enacted open science. Beginning in 2011 the NSF required that all grant proposals include a DMP. This federal policy introduced data management planning to researchers. How effective this policy is for open science is still a question that looms today. Evaluation tools have provided one path to examine and analyze effectiveness of open science by focusing on the ways data are made mobile through data management planning in science.

DMPs were not always about open science. Since the 1960s, DMPs were used by researchers in technically complex projects to ensure the analysis of data generated from research projects [4]. Today DMPs are required by funding agencies (like the NSF and the National Institutes of Health (NIH)) to provide details of how data and other research outputs will be managed during the lifecycle of a project.

A typical DMP is a two-page document affixed to a grant proposal that outlines how an individual researcher or research team will collect, manage, and preserve research project data. The DMP ideally communicates a researcher's plans to manage scientific data, and other research outputs as part of the proposed research project. However, a DMP is not a blueprint for data management practices during the project. Such a document contains anticipated plans for research outputs. It signifies the aspirations for data management project goals. Plans communicated by a DMP are not always reflective of the actual practices that take place once a project is funded and underway [4].

Several attributes make a DMP an interesting object of study for information scholars concerned with access, institutions, and knowledge commons [5].

Similar to scientific documents such as laboratory notebooks and fieldnotes, DMPs are a genre for science communication that covers core research data management topics in information science including data documentation; data standards; metadata; preservation; cost; roles; intellectual property; and data access. A DMP can signify readiness for data management and insights into the resources scientists draw on to assist in their data management at their home institution.

Several issues have been identified in regard to the DMPs and their effectiveness for data management. First, data management planning varies across disciplines. DMPs are shaped by epistemic differences, the organizational aspects of research projects and collaboration structures, and data documentation standards across research domains. Disciplines show varying philosophies and cultures around the dissemination of research outputs [3, 8–10]. In some cases, despite its importance, researchers often perceive data management planning as a time-consuming administrative task, rather than a central aspect of current research practices [11]. Federal level DMP guideline are often vague leaving it up to researchers to decide what to include in their plans [5, 12]. A look at the NSF DMP guidelines show that policies and recommendations differ across NSF programs [12]. Further, DMP requirements do not consider how data management may change over the course of a project (specifically in the humanities) [13]. Together these findings suggest that data management planning is still not regularized within academic science research.

Regardless of disciplinary differences, data management planning is considered to be vital to the futures of data. Upstream practices around data determine the paths or futures of data. Planning is considered an antecedent step to data management practices because it takes into consideration the technical and social aspects of data management before data are even created. It invokes a time for researchers to think about the formats of data, the ways they will be stored, and the means to share those data across time and space.

2.1 Evaluation Tools

To augment the writing of DMPs and assess how well DMPs capture the key activities for DMPs, information professionals (librarians, research data managers and others) have developed tools to help researchers write DMPs that comply with DMP guidelines and tools to evaluate the completeness of DMPs. For example, the DMPTool was developed by librarians across eight institutions with NSF funding in 2011 and has been periodically updated. The tool guides researchers across data storage, formatting, sharing and long-term storage and provides NSF program and program specific templates for researchers to use during while drafting DMPs.

Given that DMPs provide important documentation to how data will be managed, a set of evaluation tools have emerged in the last decade to assess how well DMPs address management criteria or how well plans cover certain topics areas deemed to be essential for short-term and long-term stewardship of data. The DART rubric and the Belmont Score are two evaluation tools for the assessment of data management plans.

Following the DMP requirements, researchers started developing DMP evaluation tools. The tools provide a means for many stakeholders to evaluate data management planning by providing systematic ways to do so. Such tools contain different areas of measurement that guide the evaluation. Assessments from evaluation can provide information that guides librarians, funding agencies, and researchers. Assessments provide valuable information on whether DMPs meet best practice standards or require improvements. They can also be used by funding agencies to reveal the shortcomings of data management planning.

The DART rubric was the outcome of a two-year National Leadership Grant Libraries Demonstration Project led by research librarians across multiple universities. The DART rubric is available on the Open Science Forum (OSF) and includes several research instruments including the scorecard, and a 33-page guideline that explains how to use the tool to score DMPs [15]. The rubric aims to provide data librarians and other information managers with a standardized analysis tool to evaluate content in DMPs. The framework is based on the generic DMP guidance in the Proposal and Award Procedures and Policies (PAPPG) (specifically Chapter II.C.2.j at the NSF – updated every year). Five key areas of research data management are assessed in the DART rubric using a Likert scale (addressed, addressed but incomplete, not addressed). Those topics include the types of data produced; standards and metadata for data; security, data protection policies for access and sharing; and plans for archiving data. The DART rubric provides stakeholders a way to assess local research data management services including gaps in expertise and training programs [15]. This rubric was in response to the NSF program. It centers its evaluation criteria drawing on policies at NSF directorates. The DART rubric does not provide a score, but it acts as a guide to assess where DMPs can be improved.

The DART rubric has been used in an empirical study of DMPs within universities and across universities [8, 14]. Studies show that DMP evaluation tools can provide insights into the completeness of data management planning and data management practices in domains [6, 8]. What is learned from these studies is that there is a relationship between the completeness DMPs and domain-level efforts to build data infrastructure. For instance, DMPs from the NSF program biology specify the exact metadata standard, a reflection of ongoing efforts in the domain to build repositories [8]. Across different fields of science, DMPs did not have adequate information about data sharing and archiving [8].

The Belmont Score is another tool that was developed by librarians and information professionals to evaluate DMPs [1]. The tool was developed in 2019 by the Belmont Forum, a multi-institutional and international collaboration committed to transdisciplinary global climate change science. The Belmont score was the outcome of the Belmont Forum workshop on e-infrastructures and data management (e-IDM) to make the forum's Open Data Policy and Principles operational. The group took ideas from the DART rubric and incorporated them into the Belmont score. The Belmont score evaluates the Belmont forum's Data and Digital Objects Management Plan (DDOMP), a similar document to the DMP. In contrast to the DMP by the NSF, the DDOMP is a live document that is revisited during the lifecycle of the project. The tool quantitatively analyzes the DDOMPs associated with Belmont Forum proposals by scoring data management topics via a Likert scale. Likert scale responses are given a numerical score. Scores are added up and divided by the number of questions to provide an overall average score. An average closer to 2 signifies a DMP that has attended to all of the requirements of the rubric. A score close to 1 indicates that a DMP has met the minimum standards. A score closer to 0 implies a DMP is missing key areas of data management planning. The scoring framework is based on a combination of existing institutional policies and evaluation tools including the Belmont Forum Grant Operations (BFGO) process, the DART project rubric, the Open Data Policy and Principles, and FAIR Data Principles. The score is also intended to aid in the development of DMPs throughout the lifecycle

of a project. The Belmont Score is published in Zenodo. The publication includes a ten-page document with instructions that explains how to use the Belmont score [1]. To date, no studies have applied the Belmont score to the analysis of DMPs until our comparative analysis presented below.

Given the decade long implementation of the DMP requirement policy and the importance of the DMP in shaping futures of data, evaluation tools provide a means to assess DMPs. However, very little is known about the criteria for evaluation and the how the different criteria are similar or different when applied to the evaluation of DMPs. What are the key areas of data management evaluation? How to these key areas get evaluated to assess completeness?

This study examines the two evaluation tools and their criteria for evaluation. It focuses on the key areas of data management planning included in the evaluation tools; how those areas are measured; and the outcomes of evaluation given the criteria from each tool. To do that, we examine the tools and apply them to the same 150 corpus of DMPs. Doing so controls for variation and provides means to examine the two tools including their strengths and weaknesses. The next section provides a detailed discussion of the approach we took to compare the DART rubric and the Belmont score.

3 Method

To carry out our comparative analysis, we began by soliciting DMPs from scientists to create a corpus and then we evaluated the DMPs using both evaluation tools. The first part of this section describes the email study to collect DMPs and the second part discussed the evaluation of the 150 DMPs (see Fig. 1 for workflow that shows the process).

To create the corpus of DMPs, we gathered a comprehensive list of projects awarded since the policy was implemented in 2011 using the NSF's Awards Database. This database allows for program-level queries and the search results can be exported in CSV format. There were several pieces of relevant and administrative metadata in the awards database that were subsequently used in our email campaign for DMP collection, including PI names, email addresses, institutions, dates of awards, award numbers, project titles, and abstracts. From January 2011 to June 2021, awards from five NSF programs were gathered into a spreadsheet: Division of Biological Infrastructure (DBI); Civil, Mechanical, Manufacturing Innovation (CMMI); Secure and Trustworthy Cyberspace (SATC); Science and Technology Studies (STS); and Oceanography (OCE); and the Science, Engineering, and Education for Sustainability (SEES). The SEES program is an NSF wide program that incorporates the other directorates. Grants from SEES can be from the five other directorates.

Once relevant awards were identified, special kinds of awards for student education, early career researchers, or field-building work that were not likely to generate research data were removed. Specifically, the following types of awards were removed: Rapid Response Research (RAPID), Early Concept for Exploratory Research (EAGER), Faculty Early Career Development Grant (CAREER), Education (EDU), Research Coordination Network Grant (RCN), Workshop, Symposium, Research and Curriculum Unit (RCU), and all grants under $100,000. Though RAPID, EAGER, and CAREER Awards require DMPs and generate research data, they were removed as EAGER and RAPID

grants are typically short proposals for shorter projects that are more experimental in nature. Meanwhile, CAREER grants are awarded to early career researchers who are generally untenured and less experienced. These types of projects, we reasoned, were not necessary to our inquiry, and we didn't want to overburden PIs. If there were multiple awards for one PI, newer awarded projects were removed from our list, in order to avoid solicitation fatigue from multiple requests and try to obtain DMPs from older projects.

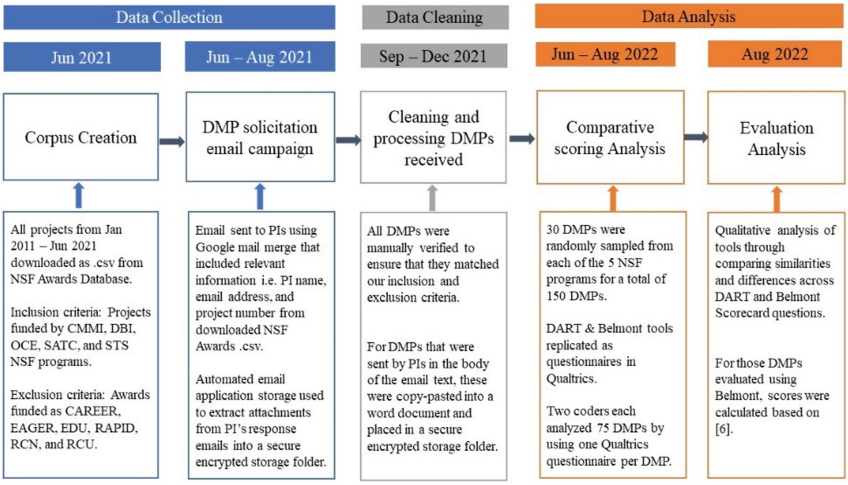

Fig. 1. Workflow

The email campaign was conducted from June through August 2021. Google's developer mail merge template was used to pull relevant data columns from our NSF awards spreadsheet including, PI name, email address, project name, and unique project number. A template email was created that requested PI's participation in our research project and included information about how their DMPs would be used. The mail merge template allowed each email request to be tailored to individual PIs, including their specific award titles and numbers in the email's subject line and message. Participants were asked to respond to the email by attaching their two-page DMP from the specific project we requested. An automated email storage application was used to collect email attachments sent by respondents that deposited them to a secure, encrypted storage folder. Some PIs responded by copying and pasting their DMP prose directly in the email's body without sending an attachment. In these instances, the relevant sections of the email were saved in a word document, .docx. Further document analysis was conducted to confirm that the DMPs received fit our inclusion criteria (e.g., occasionally PIs submitted another DMP or other proposal documents that were not relevant). An email archive was used to collect email responses from PIs for further analysis.

In total 1014 DMP submissions were received for a 18.38% average response rate across all programs (see Fig. 2). For this study, the 150 DMPs were retrieved from the corpus of 1,014.

NSF Program	# Of Emails Sent	# Of DMPs Received	Response Rate %
OCE	1689	395	23.39
DBI	1245	239	19.2
CMMI	1429	204	14.28
STS	221	48	21.72
SATC	915	122	13.33
SEES	40	6	15
Totals	**5339**	**1014**	

Fig. 2. DMPs per program

The comparative scoring analysis was conducted on a longitudinal sample of 150 DMPs from five NSF programs. 30 DMPs were selected from each of the five NSF programs from 2011–2021.

Prior to the comparative DART and Belmont analysis, research team members read all of the available documentation about DART and Belmont scoring processes available on OSF and Zenodo. DART and Belmont scorecards were subsequently recreated in Qualtrics. Scoring was conducted collaboratively by two team members (called "coders"). Each coder analyzed 15 DMPs for each NSF program for a total of 75 DMPs. Intercoder reliability ensured that coders scored DMPs in a consistent manner. For each scoring tool, each coder individually scored two DMPs and then discussed each answer with the team. Scoring discrepancies amongst the coders were settled through discussion and workflows. All DMPs were scored twice, once with the DART rubric and a second time with the Belmont score system. A detailed approach is discussed below.

3.1 Evaluation Tool 1: The DART Rubric

The corpus of 150 DMPs was evaluated using the DART score. The DART survey instrument was accessed from the OSF website and replicated verbatim in a Qualtrics survey. 26 questions had multiple choice responses; 13 questions consisted of both multiple choice and free responses. Following DART's instrument, each multiple-choice question included the following response options: (a) complete and addressed; (b) addressed but incomplete; and (c) did not address. Additional free responses and rotating dial questions were added to the beginning of the survey to keep track of the DMP. This included a free response to write in the PI's name associated with the DMP; a rotating dial to select the NSF program associated with the DMP, and a rotating dial to select the date associated with the DMP. Page breaks were incorporated to separate different sections. The team tested the Qualtrics survey twice for usability and accuracy.

Scoring DMPs required some familiarity with the DMPs' research domains, so the team spent one week learning about the rubric (including topics around outputs, concepts such as metadata and licensing), the rubric's evaluation criteria and reading current and old NSF program specific DMP guidelines as well as a few DMPs to acclimate to the genre.

Whilst scoring, coders kept a lab notebook open during the process to record insights or observations. Coders also kept the DART score rubric open to refer to examples. The coders evaluated the DMP by answering the Qualtrics survey questions. Each DMP was evaluated using this survey and coders answered each question one at a time. Over two weeks, this process was repeated for all 150 DMPs. Qualtrics survey results were retrieved via excel files.

3.2 Evaluation Tool 2: The Belmont Scorecard

After DART scoring was completed, the same corpus of 150 DMPs was evaluated using the Belmont Score. A similar approach to DART scoring was taken. The scorecard was downloaded from Zenodo. Sixteen questions were replicated verbatim to a Qualtrics survey. Each question was given a multiple-choice answer of (a) complete and addressed (b) incomplete response or a (c) no response. Following the Belmont Scorecard guidance, each response was given a weighted score. Responses that were scored complete and addressed received 2 points, responses scored incomplete received 1 point and a response that scored no response received 0 points. Additional free questions were added to connect the DMP including a free response to write the name associated with the DMP; a rotating dial to select the NSF program associated with the DMP, and a rotating dial to select the date associated with the DMP. The Qualtrics survey was tested for usability by the team. Once the instrument was ready, the team read the Belmont scorecard and its evaluation criteria. Due to the similarities across the tools, key concepts were familiar to the coders.

The same 150 DMPs were scored in two weeks. Coders accessed the DMPs in the google doc folders. Coders cross-checked the DMP with an excel file (that documents the DMPs context including the PI who submitted the DMP, the NSF program, and program, the project's proposal) to make sure the correct DMP was accessed. Whilst scoring, coders kept a lab notebook open to record any insights or observations. Coders also kept the Belmont score rubric open to refer to examples. A new survey was created for each DMP. The coders evaluated the DMPs by first reading each DMP. Then coders answered survey questions by reading line by line. Coders answered each question one at a time. Data from Qualtrics was retrieved via excel files.

3.3 Evaluation Analysis

Two methods were applied to conduct the evaluation. First, a qualitative approach was taken to assess the tools. Questions were extracted from both tools and were analyzed by comparing them for similarities and differences. Each question from both rubrics was placed into a category of data management planning. This approach was paired with the coding results in Qualtrics. Data from coding using the DART rubric and the Belmont score in Qualtrics was downloaded in an excel file. Response types were examined across the categories of data management planning. For Belmont, scores were calculated based on the method provided in [1]. The unit of analysis was the statement text at the sentence and paragraph level. The following reports on the comparison of the tools using the response types coded to the data management planning statements.

4 Findings

This section presents results from the comparative analysis of 150 DMPs using the DART rubric and the Belmont score. The findings are presented in two sections. The first section provides a comparison of the total responses that were collected by evaluating 150 DMPs by the DART rubric and the Belmont score. In the next section, comparisons are presented from five topic areas: scientific outputs produced from research grants; roles and responsibilities for data management planning; metadata planning; and planning for cost and volume of scientific data. For each specific topic area, the importance of the area is discussed in relation to data management planning, how each tool evaluates this area, the criteria for evaluation, and findings from using each tool to evaluate 150 DMPs. For each evaluation tool, the metric is presented in parenthesis. For instance, DART rubric (1.1) refers to question 1.1 in the rubric. Similarly, Belmont score (2.1) refers to questions 2.1 in the score.

4.1 Completeness of DMPs

Both the DART rubric and the Belmont score contain similar evaluation features and metrics. Both use the Likert scale model to assess DMPs and both evaluate key areas of data management planning evaluation including: defining outputs; roles and responsibilities; metadata; and cost and volume; security; and data protection.

To begin to understand the similarities and differences across the evaluation of the tools, response types at the statement level were collected from the analysis of 150 DMPs to see how many statements were complete, addressed but incomplete, or incomplete across both tools.

A larger proportion of DMPs evaluated from the DART received a higher complete response type for statements related to data management planning compared statements evaluated by the Belmont Score (Fig. 2 and Fig. 3). The finding is consistent regardless of the NSF program. DMPs in the five NSF programs evaluated by the DART rubric had a higher portion of "complete and addressed" response type. Further, both rubrics show that statements from DMPs across all five NSF programs did not meet the expectations of data management planning. Almost half of the responses by the DART rubric and Belmont score indicate that statements were incomplete (Fig. 3 and Fig. 4).

A Belmont score was calculated to see how each program compared in regard to completeness of data management. We found that DMPs from CMMI received a score of .85. DMPs from DBI received a score of .76. DMPs from OCE received a score of .70. DMPs from SATC received a score of .64. DMPs from STS received a score of .81. All five directorates received a score below one. This indicates that the DMPs across all five programs did not meet the minimum standards for data management planning according to the Belmont score.

To further investigate the differences in the evaluation tools particularly why statements in DMPs evaluated by the Belmont score scored lower, response types were collected by particular areas of evaluation across the two tools including four key criteria research outputs; roles and responsibilities; metadata; and cost and volume.

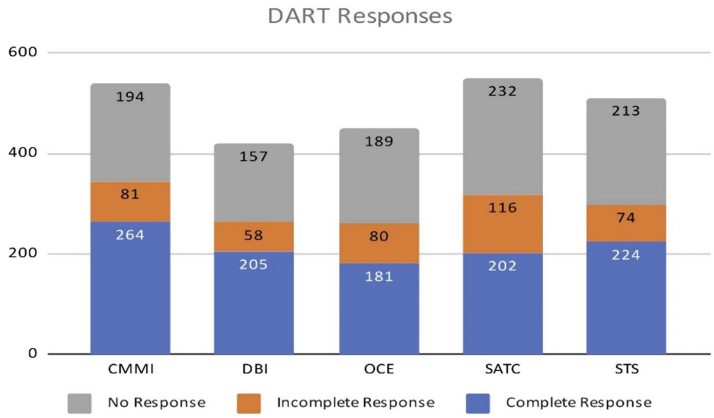

Fig. 3. The number of responses collected by the DART analysis

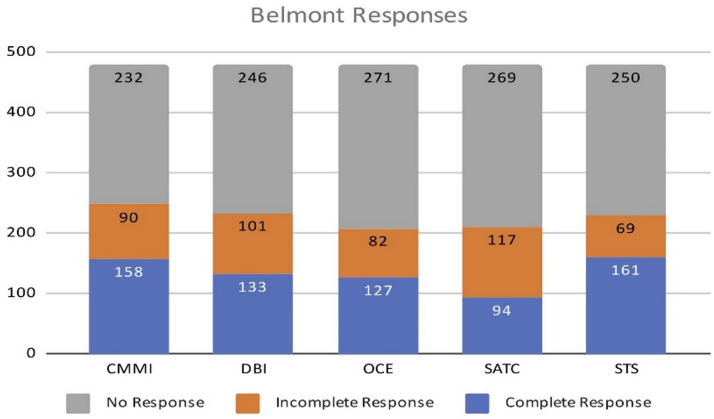

Fig. 4. The number of responses collected by the Belmont analysis

4.2 Research Outputs

Descriptions of research outputs signify what exactly is planned to be managed during the lifecycle of the project – an important statement to include in the DMP because it specifies the meaning of "data" to be managed by researchers. Data included many different objects. For instance, DMPs stated that numerical, media, text data and digital artifacts such as code, software, and databases were all described as the output from the project to be shared or archived.

Both tools evaluate statements of research outputs in DMPs. Evaluation questions include (a) what research output is defined in the DMPs (DART rubric 1.1 and Belmont score 1.1) and (b) how the research output will be generated (DART rubric 1.2 and Belmont score 1.2). However, the tools take different approaches to assess statements of outputs in DMPs. The DART score (1.1) emphasizes research data as a scientific output, and it leaves the definition of data up to a specific research agency or program.

In contrast, the Belmont score (1.1) defines a research output as any scientific object including software, code, or other outputs such as a database. To receive a complete response type, the statement also has to include the format of the data. The Belmont scores couples together the actual data and the format as important factor to data management. The DART rubric 1.2 specifies that the question only applies to specific NSF programs. Also, the Belmont score draws attention to the long-term output, not just any output for management during the project.

Statements about the research outputs and formation of research outputs were evaluated using both tools. The DART rubric 1.1 was applied to the 150 DMPs but not the DART rubric 1.2 (our sample did not have any DMPs from the suggested NSF programs). Both the Belmont 1.1 and 1.2 were applied to the corpus. A direct comparison was made of the DART rubrics 1.1 to the Belmont score 1.1 but a direct comparison of DART rubric 1.2 and Belmont score 1.2 was not conducted.

A comparison of DART rubric 1.1 and Belmont score 1.1 shows that a majority of DMPs are able to identify the outputs from research projects but the majority of DMPs scored higher when using DART as compared to the Belmont (Fig. 5). The higher scores in the DART may be due to the looser criteria for statements. The Belmont score evaluates a DMP to be complete if the output is described along with the format of the research output. The key differences here is that the Belmont score accounts the connection between the output and the format.

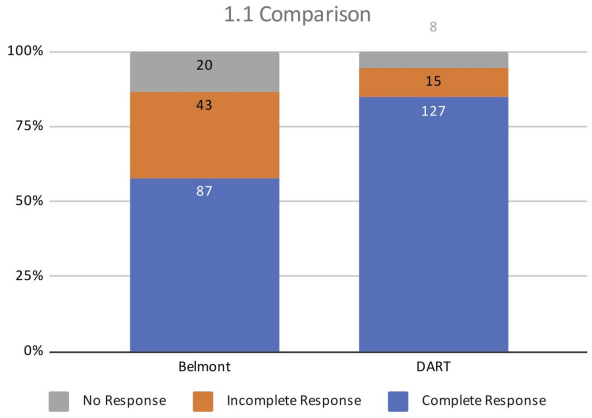

Fig. 5. The comparison of statements discussing research outputs

4.3 Roles and Responsibilities

The second area of overlap between the two tools is the evaluation of roles and responsibilities. The presence of statements that specify roles and responsibilities in data management planning outline the roles of individuals that will be responsible for the day-to-day data management activities. Anticipating the roles and responsibilities of data management is vital to planning because data management takes significant labor. Research outputs have to be cleaned, documented, stored, and licensed for sharing.

The DART rubric (6.10) evaluates roles and responsibilities in a free response question (this dimension is not evaluated from a Likert scale). The rubric provides a list of roles (PI; Co-PI; graduate student; post-doc; other/ N/A) for a coder to capture the roles in statements. The metric can capture specific roles and provide the coder more contextual information associated with the roles during the lifecycle of the project DMP.

In contrast, the Belmont contains a two-tier evaluation of roles and responsibilities compared to the DART rubric. The Belmont score (3.1) assesses whether the DMP defines the member of the team that will be responsible for developing, implementing, and overseeing the data management plan. In addition, the Belmont score (5.2) also assesses who will be responsible for managing the data after the project ends to ensure long-term accessibility. Criteria for a complete response entails statements that provide an exact description of the role during the project and after the project.

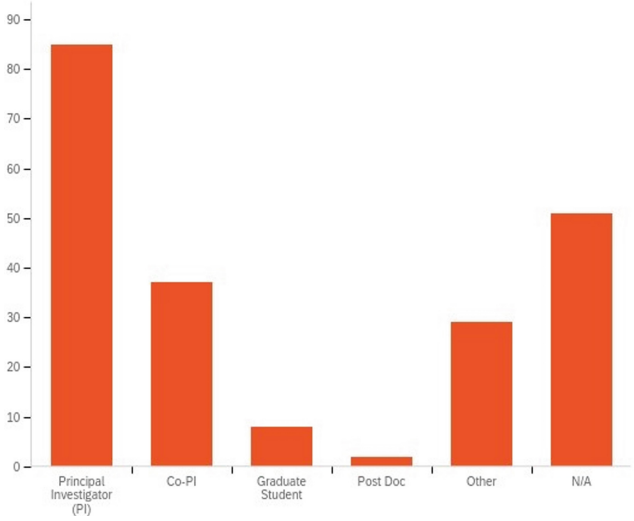

Fig. 6. Roles described in statements evaluated by the DART rubric

The DART rubric 6.10 and the Belmont score 3.1 and 5.2 were applied to the 150 DMPs. The DART rubric evaluation produced description of roles across the 150 DMPs (Fig. 6). The majority of DMPs stated that a PI would be in charge of data management planning. The Belmont score evaluation (3.1 and 5.2) provided an overview of the assessment of statements of roles and responsibilities but could not provide an in-depth look into the exact roles and responsibilities. The majority of DMPs scored low on both questions related to roles and responsibilities. Only one-third of the DMPs had text pertaining to the roles (Fig. 7).

We made qualitative observations about the differences between the NSF programs STS and OCE. When looking closely across NSF program, the DMPs from the STS had a very large number of DMPs that addressed the roles for long-term management of outputs. The coders attribute the high complete statements of roles and responsibilities in STS to the fact that the DMPs also state that sensitive data is a part of the project

including clear descriptions of how the data would be used in the project and outside the project. There was an overlap between sensitive data and descriptions for roles. Coders also noted that OCE had a very high number of DMPs that did not state roles and responsibilities but that this was related to the significant number of DMPs that stated a data repository for access. Oceanography researchers receiving grants from the OCE program could be signifying roles and responsibilities by stating the repository. Statements about roles and responsibilities overlap with statements made about data repositories and security of data.

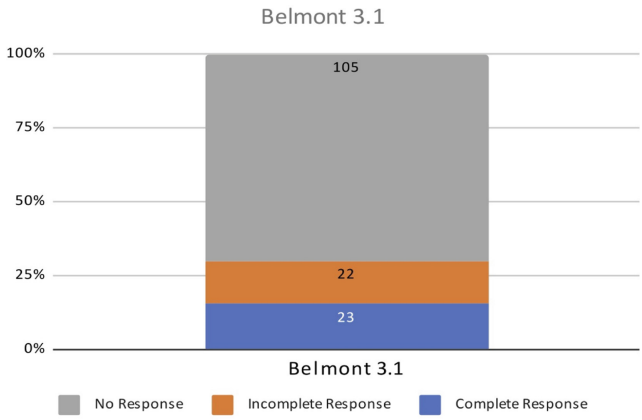

Fig. 7. Response types for roles in Belmont Score

4.4 Metadata

Metadata is vital to data management as it provides the necessary documentation for the data to be stored and reused. Metadata can provide the minimum descriptive documentation for data to be used in different settings.

Both tools evaluate metadata standards as part of data management planning. Two questions evaluate metadata in DART rubric (2.1 and 6.1). Question 2.1 evaluates the presence of a metadata standard and/or a format and question 6.1 qualitatively captures the exact standard that is described. In contrast, the Belmont score only includes one evaluation question for metadata (2.1). The Belmont score 2.1 evaluates metadata presence and kind of metadata. To be considered complete, both the presence and the specific kind of metadata need to be described (statements are not complete if they specify a workflow or an ad hoc standard). This is a much stricter criteria of evaluation compared to the DART rubric.

Comparison of the results show that statements in the DMPs scored by the Belmont score showed that statements scored lower compared to the DART rubric evaluation of statements pertaining to metadata. This is in part because the Belmont score asks for a specific standard as part of evaluation. The comparison of statements in DMPs across the five programs evaluated by the DART rubric (2.1) and the Belmont score (2.1) show

that the DART rubric 2.1 had more incomplete and no response types (shown in Fig. 8 and Fig. 9).

The high scores from the DBI program corroborates empirical work on data practices related to metadata standards in biology and the life sciences more broadly [16]. DMPs from this area described many different standards (such as Ecological Metadata Language and Darwin Core). In contrast, a majority of DMPs from all programs scored an incomplete in this area. For example, coders found that the DMPs from the SATC program did not list any metadata standards – an outlier from the other four NSF programs.

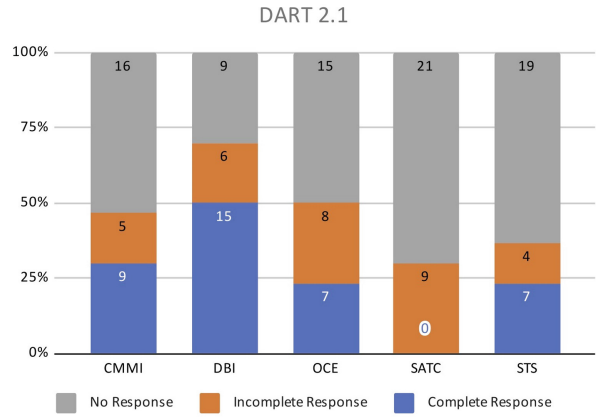

Fig. 8. Response types for metadata statements evaluated by the DART rubric

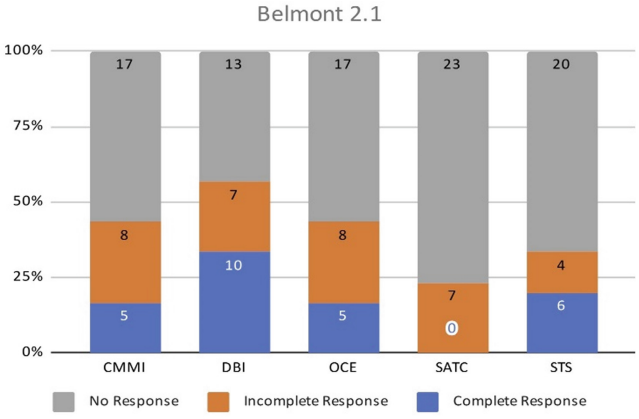

Fig. 9. Response Types for metadata statements evaluated by the Belmont score

4.5 Storage and Labor Costs

The next evaluation metric compared across the two tool was the cost of data management. This criterion involves the social and the technical costs of managing data. The costs associated with data management include infrastructure costs (university or external resources); preservation costs (archiving data for the longterm); time scales of management costs; and labor costs such as staff salaries and/or project-based fees. The DART and Belmont evaluation tools both address the cost of data management but in different ways.

The DART rubric (1.3) evaluates whether DMPs describe the amount of storage necessary, but the rubric does not evaluate cost. In contrast, the Belmont score evaluates both volume (1.3) and cost (9.1). Volume is assessed by the quantity of data that is stated in the DMP. Cost is analyzed in terms of the presence of a statement for the costs associated with long-term data management or costs associated with an assigned data manager. The metric does not specify whether to evaluate labor costs OR technical costs OR both.

The evaluation of volume and cost could not be compared across the tools. The differences between the two tools in regard to cost and volume create a gap in the comparison of the evaluation area.

According to the responses from the Belmont score (1.3), NSF programs across the board received low scores when it came to describing volume (Fig. 10). The majority of DMPs did not anticipate volume of data at all (Fig. 11). The DMPs from OCE had more complete statements about volume whereas the DMPs from SATC had the most DMPs with no response. As is shown though, across the programs a majority did not address the amount of data that the project would produce.

When analyzing how DMPs did on the question of cost, the numbers were striking. Of the 150 DMPs, only 18 addressed the cost of data management plans. Across all programs DMPs did not address cost (Fig. 11).

In summary, the comparison of evaluation tools drawing on the same corpus of 150 DMPs provided insights into differences, similarities and overlaps among criteria.

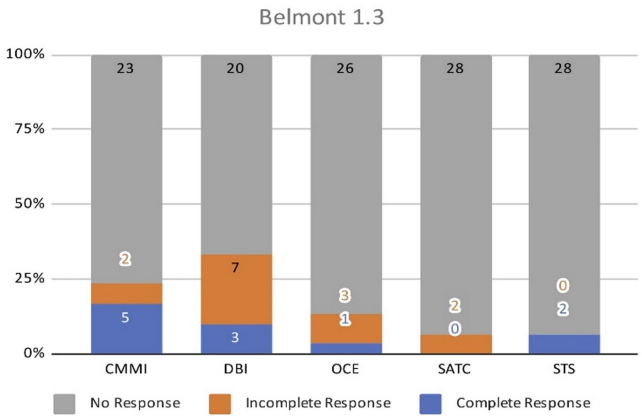

Fig. 10. Response types for storage evaluated by the Belmont score

However, it is unclear whether these scores are statistically different from each other. Qualitatively, the scores across the NSF programs are within a small range. This suggests that not one NSF program did better than the others. These findings though to do pose some interesting points about the state of evaluation.

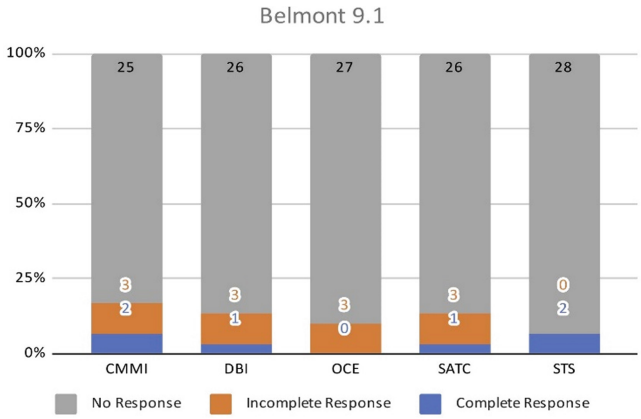

Fig. 11. Response types for cost evaluated by the Belmont score

5 Discussion

The findings above have implications to how we understand the state of evaluation of DMPs. Tools such as the DART rubric and Belmont score are the only tools currently available to assess DMPs. We compared the two to understand their similarities and compare their performance. We found that the 150 DMPs scored better with the DART rubric compared to the Belmont score. We found significant overlaps amongst the two tools but varying evaluative criteria. Overall, the Belmont score was stricter in evaluation criteria. Drawing on these results, we put forth three discussion points to the broader literature on the evaluation tools. First, we found that the Belmont score is ideal to assess DMPs. Second, the Likert scale while useful in some ways also constrains evaluation. Lastly, we found that DMP evaluative tools can provide general insights into data cultures.

First, we recommend the Belmont Score to evaluate DMPs. The Belmont score is a more precise evaluation tool for DMPs because of the specificity it demands. Next, we discuss the comparison of the tools and mechanisms for evaluation. We discuss the need for a tool that can evaluate the interactions among the criteria and the pros and cons of the Likert scale model for evaluation. Last, we also provide some insights into the data cultures of the five NSF programs based on the analysis of DMPs. The results indicate variability between programs, and thus scientific domains, in data management practices and planning but corroborate the need for more data management training. A program level perspective provides a broader look into the variables that constrain data management planning across domains of research.

5.1 Comparison of DMP Assessment Tools

The DART rubric and the Belmont score overlap in many ways. We presented several overlaps in our results: completeness; roles and responsibilities; metadata; storage and labor costs. In addition, both rubrics utilize a Likert scale model to assess evaluation with a set of criteria to evaluate complete, incomplete, or not present. This was expected given the fact that the Belmont score developed by drawing on criteria from the DART rubric.

The two rubrics also differ in unique ways. The DART rubric has many flexible criteria. For example, the DART rubric accepts any form of metadata as long as it documents data. The DART rubric also has a qualitative section that is not part of the Likert evaluation scale which does not cohere with the rest of the rubric but is useful to gain some qualitative text from the DMP. While it does provide a way to capture statements metadata standards, roles, and storage sites data it does not provide a standardized way of doing so for all DMPs. As noted by its developers [15] and confirmed here, it is an excellent tool for librarians and information professionals to identify areas that require attention and the resources that can be developed for data management planning. The downside of this approach though is that it is not a robust tool that can provide a full assessment of the whole document.

On the other hand, the Belmont Score's evaluation criteria are not flexible and requires a stricter interpretation of the criteria for DMPs to achieve a complete score. By stricter we refer to the guidelines for evaluation using a Likert scale. The comparison of the two tools across 150 DMPs shows that the Belmont score requires more detailed responses to data management planning areas compared to the DART in order for the text to be rated as complete. This is reflective of the fact that the Belmont score was developed to assess DMPs during and after the project. Most importantly the Belmont score takes into consideration interactions across data management topics. The Belmont score weights all of its criteria and provides a scoring mechanism to evaluate a DMP. This scoring mechanism provides evaluators with a quick way to score DMPs and compare them across grants.

Based on our comparative study, we recommend the Belmont Score for the assessment of DMPs. The purpose of the Belmont Score numeric score is to evaluate whether DMPs meet or exceed guidelines. This score can be calculated for each DMP. An average score can be calculated for a specific research domain. This score provides evaluators with a standardized method to examine many different kinds of DMPs within and across specific research domains.

5.2 Likert Scale: Pros and Cons

Both tools evaluate data management topics from a Likert scale. A Likert scale is an easy evaluative mechanism to use but it fails to capture the true quality of DMP content. For example, a DMP may specify a metadata standard and the tool would evaluate it as complete, but it does not tell us if this is a good choice for ensuring that future scientists can make use of the data. Future evaluation tools might need to be research field specific in order to evaluate quality of data management, access, and preservation.

We also recommend that evaluation tools need to be reexamined to assess not just key topics of data management but the relationships amongst those key areas. As the study has shown, data management planning provides a window into the intricacies of how data management topics interact. For instance, the technical qualities of data (including volume and format) are relational to its storage and access. Likewise, technical qualities of data are related to cost and labor associated with its management, curation, and preservation. Evaluation tools treat these topics are separate when in practice those topics must overlap. Further, the description of a data repository may cover metadata standards, access, and preservation all at once. Further work could be done to build evaluation tools that takes into consideration the maturation of the institution of data management.

5.3 DMPs Vary Based on Research Program

The comparative assessment of DMPs by tools also provided some insights into the DMPs by program. When statements of practices reoccur across individuals and groups, we say that researchers are drawing on similar norms for data management planning. The recurrence of statements across DMPs across five programs can be a proxy for how researchers across domains approach data management criteria. For instance, DMPs from DBI and OCE programs had higher rates specific statements that contained metadata standards. This is not surprising given the formalized guidelines and requirements of data repositories around metadata in the domain of biology and oceanography [16, 17]. We found that DMPs mostly lacked statements about volume of data and data management costs. This raises questions as to how storage and cost can be anticipated given the abstract nature of the volume of data. Finally, a majority of DMPs fail to define the roles and responsibilities associated with day-to-day management and long-term management. This is not surprising given the fact that data management planning is a form of invisible labor. How this particular area and other areas can be reassessed by exploring DMPs templates at federal funding agencies.

One takeaway from the study is that a majority of DMPs had no responses across the criteria described by librarian professionals as essential for data management planning. This points to a continued need to understand how and why DMPs fail to plan for topics around basic data management planning.

Future research could conduct a comparative analysis of the DART and Belmont rubrics using a larger sample of DMPs in order to investigate longitudinal trends in DMPs. Another takeaway is that assessing best practices provide great insights into how DMPs in NSF programs follow best practices, but we think there could more to examine here from a qualitative perspective. This kind of analysis would be useful to see how researchers theorize about planning and the ways in which they organize their futures for data mobility.

6 Conclusion

Over the last decade, funding agencies have required that researchers submit DMPs with their grant proposals to pursue federally funded science. However, it remains somewhat

unclear what sufficient DMPs should address, partly due to abstract guidelines provided by funding agencies [13]. As a result, data managers and librarians have developed rubrics for assessing the content within DMPs. Unlike research about data management policy and guidelines, which is at least a decade old, empirical research about evaluation tools of DMPs is a relatively new [7]. In part, this may be because DMPs are typically not published or public, and represent occluded documents often hidden but essential to the planning and practice of scientific research. Thus, our comparative study contributes directly to empirical research on DMPs by both reporting on the results of scoring a sample of 150 as well as assessing two types of evaluation tools.

In our study, we found that the DART rubric and the Belmont score use the same topics of evaluation but contain different criteria for completeness. Second, we found that the Belmont score takes into consideration overlap of criteria. Third, we found that evaluation tools provide insights into program specific trends around data management. Findings from the study fill several gaps. First, as of writing, this is the first study that applied the Belmont score to the assessment of DMPs. Second, this study uses the same corpus of DMPs to compare different DMP evaluation tools. Furthermore, our corpus of DMPs is unique. Previous studies that have assessed DMPs, collected their DMPs from a single institution. The DMPs used in this study range from multiple US institutions spanning over a decade from the beginning of the NSF DMP mandate in 2011 to 2021. Third, it provides a summary of the evaluation of the tools. It assesses the underlying mechanisms of each tool and how it shapes the analysis of DMPs. Further research should extend this analysis to DMPs for other funding agencies, particularly outside of the US.

References

1. Bishop, B.W., Ungvari, J., Gunderman, H., Moulaison-Sandy, H.: Data management plan scorecard. Proc. Assoc. Inf. Sci. Technol. **57**(1), e325 (2020)
2. Kowalczyk, S., Shankar, K.: Data sharing in the sciences. Ann. Rev. Inf. Sci. Technol. **45**(1), 247–294 (2011)
3. Bishoff, C., Johnston, L.: Approaches to data sharing: an analysis of NSF data management plans from a large research university. J. Librariansh. Sch. Commun. **3**(2), eP1231 (2015)
4. Smale, N., Unsworth, K., Denyer, G., Magatova, E., Barr, D.: A review of the history, advocacy and efficacy of data management plans. Int. J. Digital Curation **15**(1), 1–29 (2020)
5. Borgman, C.L.: The conundrum of sharing research data. J. Am. Soc. Inf. Sci. Technol. **63**(6), 1059–1078 (2012)
6. Mannheimer, S.: Toward a better data management plan: the impact of DMPs on grant funded research practices. J. eSci. Librariansh. **7**(3), e1155 (2018)
7. Hudson-Vitale, C., Moulaison-Sandy, H.: Data management plans: a review. DESIDOC J. Libr. Inf. Technol. **39**(6), 322–328 (2019)
8. Whitmire, A.L., Carlson, J., Westra, B., Hswe, P., Parham, S.W.: The DART project: using data management plans as a research tool (23 August 2021)
9. Tian, Y., et al.: An analysis of NSF data management plan guidelines. In: Proceedings of the 16th International Conference iConference (2021)
10. Carlson, J.: An Analysis of Data Management Plans from the University of Michigan. University of Michigan Library (2017)

11. Reichmann, S., Klebel, T., Hasani-Mavriqi, I., Ross-Hellauer, T.: Between administration and research: understanding data management practices in an institutional context. J. Am. Soc. Inf. Sci. **72**(11), 1415–1431 (2021)
12. Pasek, J.E.: Historical development and key issues of data management plan requirements for National Science Foundation grants: a review. Issues Sci. Technol. Librariansh. **87**, 1 (2017)
13. Poole, A.H., Garwood, D.A.: Digging into data management in public funded, international research in digital humanities. J. Am. Soc. Inf. Sci. **71**(1), 84–97 (2020)
14. Rolando, L., Carlson, J., Hswe, P., Parham, S.W., Westra, B., Whitmire, A.L.: Data management plans as a research tool. Bull. Assoc. Inf. Sci. Technol. **41**(5), 43–45 (2015)
15. Parham, S.W., Carlson, J., Hswe, P., Westra, B., Whitmire, A.: Using data management plans to explore variability in research data management practices across domains. Int. J. Digital Curation **11**(1) (2016)
16. Kim, Y., Burns, C.S.: Norms of data sharing in biological sciences: the roles of metadata, data repository, and journal and funding requirements. J. Inf. Sci. **42**(2), 230–245 (2016)
17. Chandler, C.L., Groman, R.C., Allison, M.D., Wiebe, P.H., Glover, D.M., Gegg, S.R.: Effective management of ocean biogeochemistry and ecological data: the BCO-DMO story. In: EGU General Assembly Conference Abstracts, p. 1258 (April 2012)