IJDC | *General Article*

Using Data Management Plans to Explore Variability in Research Data Management Practices Across Domains

Susan Wells Parham
Georgia Institute of Technology

Jake Carlson
University of Michigan

Patricia Hswe
Pennsylvania State University

Brian Westra
University of Oregon

Amanda Whitmire
Stanford University

Abstract

This paper describes an investigation into how researchers in different fields are interpreting and responding to the U.S. National Science Foundation's data management plan (DMP) requirement. As documents written by the researchers themselves, DMPs can provide insight into researchers' understanding of the potential value of their data to others; the environment in which their data are developed and prepared; and their willingness and ability to ensure the data are available to others now and in the long-term. With support from the Institute of Museum and Library Services, the authors conducted a content analysis of DMPs generated at their respective institutions using a shared rubric. By developing and testing a rubric designed to understand and evaluate the content of DMPs, the authors intend to develop a more complete understanding, at a larger scale, of how researchers plan for managing, sharing, and archiving their data.

Accepted 24 February 2016

Correspondence should be addressed to Susan Wells Parham, Georgia Institute of Technology Library, 266 4th St NW, Atlanta, GA 30332. Email: susan.parham@gatech.edu

An earlier version of this paper was presented at the 11th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v11i1.423

Introduction

In growing recognition of the importance of research datasets as standalone scholarly products of research, funding agencies have introduced requirements for the inclusion of a data management plan (DMP) with proposals. The primary purpose of a DMP is to describe the data resulting from a project, and how they will be made publicly accessible for reuse. In response to the growing need among researchers for support in addressing research data management (RDM) mandates, many research and academic libraries are allocating significant thought, effort, and capital toward developing RDM services.

The Data management plan As Research Tool (DART) project has as its premise that data management plans can be a rich source of information about researchers' data management knowledge, capabilities, practices, and needs. By using these plans as a window into research practices, we can discern variability in RDM habits across broad research domains, as well as the extent to which university and library resources are being consigned in the plans. Such investigations will help inform efforts to develop or improve RDM services and infrastructure. In this paper we show how librarians and other data support professionals can use DMPs as a tool for exploring local RDM behavior and identifying data management services needs. We also discuss our analysis of 500 data management plans written by researchers at five U.S. research institutions.

Background

The DART team conducted an analysis of 500 DMPs from awarded proposals submitted to the U.S. National Science Foundation (NSF), using an analytic rubric that we developed and tested for this task (Whitmire, Rolando and Westra, 2015). Our work builds upon previous research conducted on the analysis of NSF DMPs. Articles from librarians at Cornell University (Steinhart, Chen, Arguillas, Dietrich and Kramer, 2012), the University of Illinois at Urbana-Champaign (Mischo, Schlembach and O'Donnell, 2014), the University of Minnesota (Bishoff and Johnston, 2015), and Georgia Institute of Technology (Parham and Doty, 2012) have all noted that researchers have difficulty understanding and responding to the requirements. Many research libraries have set up a DMP review service as a means to support researchers (Dietrich, Adamus, Miner and Steinhart, 2012), and as a means of training librarians about the data needs of researchers (Davis and Cross, 2015).

In contrast to previous work, our research utilized a dataset of plans from multiple U.S. institutions, affording us a rich source of content for comparisons across the seven directorates that comprise the NSF. By investigating and comparing DMPs written for different directorates, we begin to see how researchers in different fields understand and interpret the NSF data management requirements. We also get an idea of how well equipped they are to meet these requirements, as well as a glimpse into how they are currently managing their research data.

In our findings, we focus on results from six of the seven NSF directorates: Biology (BIO); Computer and Information Science and Engineering (CISE); Engineering (ENG); Geosciences (GEO); Math and Physical Sciences (MPS); and Social, Behavioral, and Economic Sciences (SBE). We discuss variability in researcher data

management practices across these directorates in the areas of data sharing, data discovery and reuse, and use of data curation infrastructure. We close by discussing the implications of our findings for practitioners in data services.

Approach

To facilitate consistent review across the project team, we developed an analytic rubric for the assessment of NSF data management plans (Rolando, Carlson, Hswe, Parham, Westra and Whitmire, 2015; Whitmire et al., 2015). The rubric contains assessment criteria across three performance levels for both NSF-wide and directorate-specific DMP content requirements (Whitmire, Carlson, Hswe, Parham and Westra, 2016a). The rubric was tested and improved through two rounds of individual reviews of the same set of DMPs and subsequent assessment of inter-rater reliability (IRR). We used intra-class correlation (ICC) to assess IRR (McGraw and Wong, 1996; Shrout and Fleiss, 1979; R package 'irr'[1]), and, through improvements in the rubric, were able to achieve a median ICC score of 0.76, which is within the range of having excellent agreement between raters. We anticipate that the analytic rubric we developed to facilitate this work can be used by others to conduct their own RDM assessments.

Each team member assessed a random sample of 100 DMPs from their respective institution to create a dataset with 500 total DMP reviews (Whitmire, Carlson, Westra, Hswe and Parham, 2016b). This approach avoided potential rater bias for a given directorate, and distributed the work of reviewing plans evenly across the team. The resulting set of DMPs reflected the research strengths of each institution, but in aggregate also provided a sample distribution among the directorates that is similar to the national NSF awards. In addition to recording performance level ratings for the assessment criteria, we also gathered supplementary information, such as how researchers said they would share and archive their data, whether or not they mentioned the institutional repository or other university resources, if they mentioned a specific metadata standard, and so on. We translated the rubric into a Qualtrics survey to facilitate data collection and co-location, and to standardize scoring and collection of supplementary information. Some of the randomly selected plans stated that the research would not produce data and therefore, no DMP was needed. This analysis is based on the proposals that included a DMP (465 of the 500 selected). We did not drill down to the division level to ascertain differences that might be found there.

There are some inherent limitations in conducting a content analysis of DMPs. The information presented in a DMP regarding the data can be fairly complex, written for experts in the field. Without disciplinary knowledge, it may be difficult to fully understand the plan. The DMP is only one component of a grant proposal, and may make reference to other parts of the application that are unavailable. In addition, researchers write proposals to win funding, and therefore may be more motivated to write a DMP that appeals to the stated goals of the agency, rather than to provide an accurate description of their practices and intentions.

1 Various Coefficients of Interrater Reliability and Agreement: https://cran.r-project.org/web/packages/irr/index.html

 doi:10.2218/ijdc.v11i1.423

Results and Discussion

The distribution of DMPs selected for this study across the NSF directorates closely follows the overall funded proposal distribution (See Table 1). This indicates that our selection of DMPs was suitably random, and that findings may be generalized. Of the 500 DMPs in our sample, 465 (93%) stated that the proposed project would produce data (Table 2), and therefore described a plan. The numbers and percentages in the rest of the paper refer to this subset of 465 plans.

Table 1. Number and percentage of proposals funded for the National Science Foundation (NSF) as a whole (FY 2014) and for proposals reviewed for this paper.

	Number		Percentage	
	NSF [n]	DART [n]	NSF [%]	DART [%]
BIO	1272	53	12.0	10.6
CISE	1680	72	15.8	14.4
EHR	701	18	6.6	3.6
ENG	2145	116	20.2	23.2
GEO	1487	89	14.0	17.8
MPS	2343	95	22.1	19.0
SBE	994	51	9.4	10.2
Unk	n/a	6	n/a	1.2
Total	10622	500	100.0	100.0

Table 2. "Yes" responses to the question: "Will the project produce data?", by number and percent of DMPs from NSF-wide or within each directorate.

	Number	Percentage
All	465	93.0
BIO	52	98.1
CISE	66	91.7
ENG	106	91.4
GEO	83	93.3
MPS	85	89.5
SBE	50	98.0

NSF guidelines across all directorates stipulate that the DMP must describe the data to be captured, created or collected. We evaluated how well researchers described their data, and found variability between the directorates (Table 3). Proposals submitted to BIO and SBE had data management plans that better defined the types of data to be produced during research, while those submitted to CISE were significantly less complete. Among all directorates, 5.8% to 15.3% of DMPs (or 9.5% overall) failed to describe the data that would be produced in any way. Throughout our review, we note that the DMPs submitted to BIO do a consistently better job of meeting rubric criteria.

Table 3. DMP performance level ratings for the criterion: "Describes what types of data will be captured, created or collected."

	Complete/ detailed (%)	Addressed issue, but incomplete (%)	Did not address (%)
All	68.4	22.2	9.5
BIO	89.7	11.5	5.8
CISE	53.0	31.8	15.2
ENG	73.6	19.8	6.6
GEO	63.9	28.9	7.2
MPS	61.2	23.5	15.3
SBE	82.0	12.0	6.0

Data Sharing

Of the DMPs reviewed, only 2.6% included statements that the data would not be shared, while 7.5% failed to specify how the data would be shared (Figure 1). Options for data sharing were not mutually exclusive, as many of the DMPs noted several different avenues for sharing. The most popular means, observed in 36.1% of DMPs across all directorates, was through journals (tables, supplements, etc.). However, as Figure 1 and the following directorate summaries show, the relative percentages for data sharing methods varied considerably across domains. We discuss selected findings below.

	All	BIO	CISE	ENG	GEO	MPS	SBE	Scale
Journal / supplement	36	27	23	45	35	54	18	80
Data center or repository	34	75	14	8	66	25	42	70
On request	30	23	30	38	24	34	26	60
Personal website	25	13	44	31	25	20	12	50
Other method	22	27	30	15	23	18	22	40
Institutional repository	17	6	12	20	6	28	20	35
Conference / proceedings	13	8	11	23	8	13	8	30
Did not specify	8	0	18	9	2	8	4	25
Thesis / Dissertation	3	0	0	5	2	6	2	20
Not planning to share	3	0	5	3	0	1	10	10
Book	2	2	2	3	2	2	2	0

Figure 1. Methods of sharing research data as described in NSF data management plans. Numbers are percentages (shaded by color according to the scale).

An overwhelming proportion of BIO DMPs (75%) indicated that data centers or repositories would serve as the key platforms for sharing. This percentage is much higher than what was observed in DMPs overall, suggesting that researchers in biology fields are not only more likely to deposit data into repositories and data centers, but are also more familiar with this dissemination approach. The BIO DMP preference for data centers and repositories may also help explain why types of data in these DMPs are

 doi:10.2218/ijdc.v11i1.423

more thoroughly described than DMPs overall. In addition, the named data centers – GenBank, Dryad, and the Sequence Read Archive, most frequently – are fairly established centers and repositories, suggesting that the propensity for sharing, or at least the intention to share, is common in the fields associated with biology. No BIO DMPs stated that researchers were not planning to share their data, nor did any fail to specify how data would be shared. Given how thoroughly BIO DMPs described the data for their proposed projects, it's probably not surprising that these DMPs also stood out on the question of how data would be shared.

For CISE DMPs, personal websites maintained by project personnel were the top venue proposed for sharing data (43.9%). This is a much higher percentage than what was found in DMPs overall. Second to this option were sharing them on request and sharing them through "other method" (both 30.3%), such as through Github, SVN, or bitbucket repositories – systems that typically track versions of code as part of developing and maintaining software applications. The "other method" response was also higher in CISE DMPs than in DMPs overall (21.7%). A smaller proportion of CISE DMPs (13.6%) indicated that they would be using a subject-based data repository.

DMPs in the ENG directorate followed the trend of DMPs overall: publication of results in a journal was the leading means for sharing data (45.3%), followed by sharing on request (37.7%). ENG DMPs also displayed a preference for data sharing via conference presentations and proceedings – a venue similar to journal publications. Only 8.5% of ENG proposals indicated a data center or repository compared to 34.4% overall, although roughly 20% specified sharing via an institutional repository. For GEO, 66.3% of DMPs favored sharing data via specifically named data centers, repositories, or data-sharing platforms, almost double the percentage reflected across all of the DMPs (34%), and second only to the BIO directorate for this form of dissemination. The centers and repositories mentioned ranged from those associated with supporting journal articles (e.g., Dryad), to national data centers (e.g., the National Geophysical Data Center, the NSF-sponsored Biological and Chemical Oceanography Data Management Office, or the National Institute of Health's GenBank). Many small, boutique databases were indicated, as well as nationally federated systems like DataONE. Like the DMPs for BIO, GEO DMPs showed little preference for institutional repositories, which were only mentioned in 6% of plans.

By far the most popular means of sharing MPS data was via supplemental information for an article (54.1%). This number is much higher than that of DMPs overall (36.1%). Sharing data via supplemental information is an approach common to chemists, and although it is often in the form of a PDF rather than actual data files, this method is supported by NSF Chemistry Division DMP guidance. Another popular choice was institutional repositories (28.2%), higher than the 16.6% of DMPs overall.

As with DMPs to the BIO and GEO directorates, SBE DMPs also showed a preference for data repositories (42%), such as the Inter-university Consortium for Political and Social Research (ICPSR) and the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI). Institutional repositories marked another popular method for sharing SBE data (20%); SBE is second only to MPS in this regard. SBE DMPs also expressed a relative disinclination for data sharing: 10% stated they would not be sharing data, a proportion much higher than all the other directorates, perhaps because of the preponderance in SBE of projects collecting or utilizing human subjects or restricted-access data.

The preference in BIO, GEO, and SBE to share data through domain-specific repositories suggests that these researchers have more familiarity with such a dissemination practice than scientists in other domains and that infrastructure

(repositories) exists to host data in these fields. For CISE, researchers in computer science fields may be more inclined to develop a solution locally, rather than use an institutional or national solution. The reviews also revealed that ENG and MPS plans showed a preference for sharing data via journal publications and supplemental information, and when requested by other researchers – two of the most conventional ways of disseminating research results. Table 4 shows a comparison of how well researchers described their plans for sharing data, by directorate.

Table 4. DMP performance level ratings for the criterion: "Describes how the data will be made publicly available." Numbers are percentages, and are shown across all DMPs and by directorate.

	Complete/ detailed	Addressed issue, but incomplete	Did not address
All	50.3	40.9	8.8
BIO	67.3	30.8	1.9
CISE	40.9	43.9	15.2
ENG	36.8	52.8	10.4
GEO	61.4	34.9	3.6
MPS	48.2	43.5	8.2
SBE	58.0	34.0	8.0

Data Discovery and Reuse Metadata

In order for data to be reused, it must be discoverable, accessible, well documented, and in a format that facilitates reuse (Van Tuyl and Whitmire, 2016). Metadata and other types of documentation (e.g., readme files or data dictionaries) facilitate discovery and reuse. Most DMPs that we reviewed did not specify a metadata standard (85.1%; Figure 2), but again, inter-directorate differences reveal key behavioral variability between research domains. Unsurprisingly, BIO DMPs have the highest percentage of plans (38.5%) that mention a metadata standard (Ecological Metadata Language and Darwin Core most often), while only 6.1% of CISE DMPs specify a standard. Also note that the greater percentage of BIO DMPs (38.5%) provided complete, detailed information about metadata, compared with the CISE low of 9.1%.

	Identifies metadata standards/formats					
	Complete / detailed	Addr., but incomplete	Did not address	Most common metadata standards (N)	Yes (%)	Scale
All	19.4	23.7	57.0	Dublin Core (23), CSDGM*/ISO 19115 (12), EML^ (11)	14.9	60
BIO	38.5	21.2	40.4	EML (8), Darwin Core (5), CSDGM/ ISO 19115 (4), Dublin Core (3)	38.5	50
CISE	9.1	28.8	62.1	DIF# (2), Dublin Core (1), EML (1)	6.1	40
ENG	15.1	30.2	54.7	Dublin Core (7)	9.4	30
GEO	18.1	18.1	63.9	CSDGM/ISO 19115 (3), EML (2), ACADIS~ (2)	12.2	20
MPS	18.8	22.4	58.8	Dublin Core (9), FITS% (3)	15.3	10
SBE	26.0	16.0	58.0	CSDGM (5), ODM& (2) Dublin Core (2)	16.3	0

*Content Standard for Digital Geospatial Metadata; ^Ecological Metadata Language; #Directory Interchange Format; ~Advanced Cooperative Arctic Data and Information Service; %File Information Tool Set; &Observations Data Model

Figure 2. Aspects of how well metadata is addressed in DMPs. In the first three columns, the DMP performance level ratings for 465 DMPs are shown (in %). The most commonly named metadata standards and the percent of DMPs that name a specific metadata standard are also shown. Percentages are shaded by color according to the scale at right.

In many cases, DMPs that did not name a specific metadata standard were still assessed as having fully addressed the topic (Figure 2). Slightly more than fifteen percent of MPS plans mentioned a metadata standard, while 18.8% received a "fully addressed" rating. While initially seeming erroneous, this result makes sense in light of the fact that metadata standards do not exist for all data types or domains. In some cases, a DMP did not have to mention a specific metadata standard in order to receive a "fully addressed" rating. For example, in lieu of listing standards, several CISE DMPs mentioned creating locally relevant metadata fields and/or readme files. GEO DMPs that did not mention a particular standard (because one doesn't exist for that data type) often mentioned the creation of a readme file. In addition, some plans described important characteristics of the data that they would capture, such as equipment calibration settings or corrections, without defining these fields as metadata. A number of MPS plans discussed documentation for experiments, such as would be recorded in lab notebooks, which reflects common practices in chemistry, for example.

Many DMPs simply stated that they would create "metadata" or "documentation" without providing detail or explanation, a phenomenon noted across all of the directorates. This may indicate that researchers have a limited understanding of what metadata is, and its role in making their data discoverable and useable by external audiences. While the proportion of plans that did not address metadata at all is a discouraging 57%, we saw many researchers making an honest effort at addressing what can be a difficult topic.

Polices for Reuse, Redistribution, and the Creation of Derivatives

NSF guidelines state that DMPs should include statements on policies for data reuse, redistribution and derivative creation. The collection of DMPs we reviewed lacked detail about these policies: 56.3% did not mention reuse policies (Table 5); 63.4% did not address redistribution (Table 6); and 69% did not provide policies for

derivative creation (Table 7). In many cases, DMPs were rated as having partially addressed reuse if a policy could be inferred from the author's selection of a data center, repository, or other publisher through which the data were to be shared, even though the policies were not explicitly included or mentioned in the DMP itself. However, in some of these cases, clear reuse policies could not be located for the respective data centers or repositories.

Table 5. DMP performance level ratings for the criterion: "Describes the policies or provisions in place governing the use and reuse of the data." Numbers are percentages, and are shown across all DMPs and by directorate.

	Complete/ detailed	Addressed issue, but incomplete	Did not address
All	15.9	27.7	56.3
BIO	19.2	23.1	57.7
CISE	9.1	40.9	50.0
ENG	17.9	31.1	50.9
GEO	9.6	27.7	62.7
MPS	22.4	24.7	52.9
SBE	10.0	22.0	68.0

Table 6. DMP performance level ratings for the criterion: "Describes the policies or provisions for redistribution of the data." Numbers are percentages, and are shown across all DMPs and by directorate.

	Complete/ detailed	Addressed issue, but incomplete	Did not address
All	13.8	22.8	63.4
BIO	19.2	13.5	67.3
CISE	6.1	36.4	57.6
ENG	14.2	26.4	59.4
GEO	3.6	26.5	69.9
MPS	24.7	18.8	56.5
SBE	14.0	14.0	72.0

 doi:10.2218/ijdc.v11i1.423

Table 7. DMP performance level ratings for the criterion: "Describes policies or provisions for building off of the data, such as through the creation of derivatives." Numbers are percentages, and are shown across all DMPs and by directorate.

	Complete/ detailed	Addressed issue, but incomplete	Did not address
All	08.8	22.2	69.0
BIO	13.5	15.4	71.2
CISE	09.1	27.3	63.6
ENG	08.5	22.6	68.9
GEO	01.2	26.5	72.3
MPS	15.3	23.5	61.2
SBE	04.0	16.0	80.0

Across all directorates, the policy statements made on reuse, redistribution, or the creation of derivatives tended to be very permissive, but also vague, implying that these issues had not been given much consideration. For example, one DMP stated, "there are no limitations on any data or samples generated during the scope of this research", and another claimed, "no issues regarding… intellectual property are foreseen for this work." Other statements indicate that the researchers did not understand what was being asked of them: "We are constructing an original dataset, so there are no re-use or re-distribution issues to be addressed." A desire on the part of the researcher for others to provide some form of attribution or to cite the data appropriately was also observed. These statements were also rather vague, and typically did not define the method of attribution or citation standard.

Several DMPs referred to their institution's policies or technology transfer office; however, few if any details were provided as to what these policies actually permit for using, redistributing, or creating derivatives of the data. As such, these statements conveyed a sense that researchers felt the need to protect themselves against possible contradictions between what the funding agency required, and what their institutions, as presumed owners of the data, would permit. In addition, in some cases the results of the research were anticipated to have commercial applications, requiring the researchers to work with their institution's technology transfer or similar office before considering the release and reuse of their data.

Finally, a noticeable minority of DMPs specified a particular license for governing the reuse, redistribution and creation of derivatives from their data. The most common license to be assigned to data sets was some form of Creative Commons license (not always specified), a GNU General Public License, or a BSD license. In contrast, a few researchers referred to having or developing data use agreements of their own to address these issues.

Much like the results for metadata, these numbers as a whole suggest a need for improved understanding among researchers of what the concepts *reuse* and *redistribution* mean, and how to address them in a DMP through a stated policy or guideline. There is likely an assumption that data reuse and redistribution are natural by-products of data sharing, and are addressed by repository policies. It may also be the case that without well-publicized instances of data reuse, researchers are less aware of this prospect, and thus provide less detail in DMPs.

Data Curation Infrastructure

As part of our analysis, we also documented how often researchers mentioned campus infrastructure and library services. As shown in Table 8, nearly 21% of the reviewed DMPs mentioned using library services, from 3.7% in plans for GEO proposals, to 32.9% for MPS funding proposals. Most of these references were to library-run institutional repositories, either as a means of sharing or archiving data, or as a place to deposit articles and other products of research. Some plans did mention library consultation services for data management plans and metadata standards.

References to campus infrastructure were a bit more varied, and included the use of campus storage and backup services, and department or campus web servers. We found that many researchers seem to conflate archiving research data with simply using campus storage services. We also found that 29% of the DMPs did not specify how they were planning to archive their data, in contrast to those that did not specify plans for sharing data (8%) (Figure 3). Twenty-eight percent of DMPs specified a data center or repository – a number heavily skewed by DMPs submitted to the BIO and GEO directorates.

Table 8. The percentages of DMPs that mentioned the use of library services or campus-wide resources or services, across all DMPs and by directorate.

	Library services	Campus services
All	20.5	33.8
BIO	17.3	44.2
CISE	16.7	30.3
ENG	21.9	40.0
GEO	3.7	20.7
MPS	32.9	37.6
SBE	22.0	30.0

Centralized storage servers (23%) and PC/external storage (15%) were also notably present, and were particularly popular in ENG and MPS DMPs. This may indicate a lack of awareness of repository options, or it may indicate a reluctance to surrender local control over the data set to a third party for curation purposes. Furthermore, as archiving is often interpreted simply as long-term storage it may not be clear as to why a data center is needed or what value its preservation services would have for the researcher. We also noted that in some DMPs a repository was mentioned for sharing, but not specifically for archiving data. As noted above, there may be an assumption that all repositories provide both access and preservation services.

 doi:10.2218/ijdc.v11i1.423

	All	BIO	CISE	ENG	GEO	MPS	SBE	Scale
Did not specify	29	23	39	33	14	29	30	60
Data center or repository	28	58	9	9	59	14	36	55
Centralized storage servers	23	12	23	34	8	39	12	50
Institutional repository	15	8	14	15	6	26	20	45
PC / external storage media	15	12	9	20	10	25	10	40
Other method	12	17	15	11	8	15	10	35
Personal / project website	6	2	6	8	8	2	4	30
Journal / supplement	3	0	6	3	5	5	2	25
Not planning to archive	1	0	2	0	0	4	2	20
Conference / proceedings	1	0	3	2	0	0	0	15
Thesis / Dissertation	1	0	0	2	0	1	0	10
On request	0	0	2	0	1	0	0	5
Book	0	0	0	0	0	0	0	0

Figure 3. Methods of archiving research data as described in NSF data management plans. Numbers are percentages (shaded by color according to the scale).

Conclusion

The findings of this project indicate that the application of an analytic rubric to DMPs can yield valuable information. Though we have several caveats, we conclude that reviewing DMPs as an authentic artifact of researchers' intentions can present a useful snapshot of current data practices, uncover institutional challenges for compliance, and inform the development or augmentation of useful data services. As noted throughout the paper, we found a number of data management concepts that appear to be unclear to researchers across disciplines. This shortfall demonstrates the importance of building strong support systems to ensure that researchers respond adequately to funding agency requirements, and to ensure that they receive the full benefits that good data management, sharing, and preservation afford.

The process of generating a rubric that is aligned with both general and directorate-level NSF guidance highlighted the variability in guidance from directorate to directorate. The NSF expressly relies on research disciplines ("communities of practice") to promulgate and apply their own data management practices and infrastructure in the review of DMPs for funding decisions (National Science Foundation, 2015). This is a logical course of action, and one that should be supported. However, directorate guidelines would benefit from a shared understanding of concepts and terminology, expressed as clearly and unambiguously as possible. Common definitions and consistent approaches to accountability could help improve the quality of the DMPs and post-award compliance.

In addition to the data management concepts that researchers across domains did not understand or address (such as policies regarding data access and reuse), we also found that researchers in certain domains addressed some concepts more fully. Areas of divergence were particularly noticeable in terms of data description and sharing. We found that DMPs submitted to the BIO directorate provided more detailed descriptions of their data, how they would share data, description of metadata – including naming a

metadata standard, and specific domain repositories for sharing data. This trend leads us to speculate about the relationship between the existence of national, domain-specific repositories and the proclivity of researchers in that domain to use them effectively and in large number. The proportion of researchers in BIO who "fully addressed" the topic of metadata was nearly twice that of any other directorate. In many ways, the process of sharing data obligates the researcher to think about data formats, and about creating useful documentation. Can other research disciplines benefit from the creation of strong disciplinary repositories, with their attendant policies and standards?

As the NSF and other agencies rely on communities of practice to develop appropriate responses to the challenges in managing, sharing, and archiving data, mechanisms for communication across communities are also needed. Researchers in some fields, such as ecology, have developed support structures and processes – including data centers, publications on best practices, metadata standards and common tools – support which other fields might consider in shaping their own efforts. There is an opportunity to bring stakeholders together from across mature and emerging domains to move toward shared best practices or infrastructure. Cross-disciplinary and open membership organizations, such as the Research Data Alliance (RDA), are increasingly important conduits in leveraging efforts from one field to inform thinking and possible approaches for others.

Observations made in our study and others regarding the current quality of DMPs do not appear especially promising – an observation supported by work that shows that the presence of a data management plan does not, in most cases, lead to effective sharing of research data (Van Tuyl and Whitmire, 2016). If research institutions are committed to supporting researchers in meeting this requirement, we must acknowledge that crafting an authentic DMP will require researchers to re-conceptualize how they conceive and carry out their research on a fundamental level. The potential impact to the cultures of practice for many fields is likely to need time to fully take root and play out, and will require the support not only of disciplinary groups and funding agencies, but also of research institutions.

At the institutional level, librarians, IT personnel, grant administrators, and others have stepped up to provide assistance to researchers in responding to the DMP requirement, but clearly more collaboration is required. In addition to increased training on data management topics such as metadata and its applications, formats suited for sharing data, and documentation for data reuse, researchers clearly need guidance on data licensing options and intellectual property policies. Expertise in these areas resides in a variety of groups within one institution, so successful training programs and other support require partnerships that value and prioritize these efforts. Forging alliances and partnerships between libraries, IT centers, grant administrators, and others should become a priority to build data management capacity and address local needs.

Acknowledgements

We thank Lizzy Rolando for her critical contributions to this project. This project was made possible in part by the Institute of Museum and Library Services, National Leadership Grant LG-07-13-0328.

References

Bishoff, C., & Johnston, L. (2015). Approaches to data sharing: An analysis of NSF data management plans from a large research university. *Journal of Librarianship and Scholarly Communication*, *3*(2). http://doi.org/10.7710/2162-3309.1231

Davis, H.M., & Cross, W.M. (2015). Using a data management plan review service as a training ground for librarians. *Journal of Librarianship and Scholarly Communication*, *3*(2). http://doi.org/10.7710/2162-3309.1243

Dietrich, D., Adamus, T., Miner, A., & Steinhart, G. (2012). De-mystifying the data management requirements of research funders. *Issues in Science and Technology Librarianship*, *70*. http://doi.org/DOI:10.5062/F44M92G2

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. http://doi.org/10.1037/1082-989X.1.1.30

Mischo, W., Schlembach, M., & O'Donnell, M. (2014). An analysis of data management plans in University of Illinois National Science Foundation grant proposals. *Journal of eScience Librarianship*, *3*(1). http://doi.org/10.7191/jeslib.2014.1060

National Science Foundation. (2015). *Public access plan: Today's data, tomorrow's discoveries: Increasing access to the results of research funded by the National Science Foundation* (No. nsf15052) (p. 31). Arlington, Virginia, USA: National Science Foundation. Retrieved from https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf15052

Parham, S.W., & Doty, C. (2012). NSF DMP content analysis: What are researchers saying? *Bulletin of the American Society for Information Science and Technology*, *39*(1), 37–38. http://doi.org/10.1002/bult.2012.1720390113

Rolando, L., Carlson, J., Hswe, P., Parham, S.W., Westra, B., & Whitmire, A.L. (2015). Data management plans as a research tool. *Bulletin of the American Society for Information Science and Technology*, *41*(5), 43–45. http://doi.org/10.1002/bult.2015.1720410510

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Steinhart, G., Chen, E., Arguillas, F., Dietrich, D., & Kramer, S. (2012). Prepared to plan? A snapshot of researcher readiness to address data management planning requirements. *Journal of eScience Librarianship*, *1*(2). http://doi.org/10.7191/jeslib.2012.1008

Van Tuyl, S., & Whitmire, A.L. (2016). Water, water, everywhere: Defining and assessing data sharing in academia. *PLOS ONE*, *11*(1). http://doi.org/10.1371/journal.pone.0147942

Whitmire, A.L., Carlson, J., Westra, B., Hswe, P., & Parham, S. (2016a). Rubric and related files. *Open Science Framework*. Retrieved from http://osf.io/qh6ad

Whitmire, A.L., Carlson, J., Westra, B., Hswe, P., & Parham, S. (2016b). Data from: Using data management plans to explore variability in research data management practices across domains. *Open Science Framework*. Retrieved from http://osf.io/ewmsy

Whitmire, A.L., Rolando, L., & Westra, B. (2015). *Using data management plans as a research tool for improving data services in academic libraries*. Presented at the International Association for Social Science Information Services and Technology 41st Annual Conference, Minneapolis, MN. Retrieved from http://ir.library.oregonstate.edu/xmlui/handle/1957/56232

Whitmire, A.L., Westra, B., Carlson, J., Hswe, P., Wells Parham, S., & Rolando, L. (2015). Data management plans as a research tool. Retrieved from http://ir.library.oregonstate.edu/xmlui/handle/1957/55482