

JLSC

ISSN 2162-3309 | JLSC is published by the Pacific University Libraries | <http://jls-public.org>

Volume 3, Issue 2 (2015)

Approaches to Data Sharing: An Analysis of NSF Data Management Plans from a Large Research University

Carolyn Bishoff, Lisa Johnston

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), eP1231. <http://dx.doi.org/10.7710/2162-3309.1231>

External Data or Supplements:

Bishoff, Carolyn; Johnston, Lisa, 2015, "Instrument used to code DMPs", <http://dx.doi.org/10.7910/DVN/5JGNMM>, Harvard Dataverse. [Instrument]

Johnston, Lisa R; Bishoff, Carolyn; McGrory, John; Storino, Chris; Swendsrud, Anders. (2015). Analyzed Data Management Plans (DMPs) from Successful University of Minnesota Grants from the National Science Foundation, 2011-2014 [dataset]. Retrieved from the Data Repository for the University of Minnesota, <http://dx.doi.org/10.13020/D6TG6Z> [Data]



© 2015 Bishoff & Johnston. This open access article is distributed under a Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>)

Approaches to Data Sharing: An Analysis of NSF Data Management Plans from a Large Research University

Carolyn Bishoff

Physics, Astronomy, and Earth Sciences Librarian, University of Minnesota-Twin Cities

Lisa Johnston

*Research Data Management/Curation Lead and Co-Director of the University Digital Conservancy,
University of Minnesota-Twin Cities*

INTRODUCTION Sharing digital research data is increasingly common, propelled by funding requirements, journal publishers, local campus policies, or community-driven expectations of more collaborative and interdisciplinary research environments. However, it is not well understood how researchers are addressing these expectations and whether they are transitioning from individualized practices to more thoughtful and potentially public approaches to data sharing that will enable reuse of their data. **METHODS** The University of Minnesota Libraries conducted a local opt-in study of data management plans (DMPs) included in funded National Science Foundation (NSF) grant proposals from January 2011 through June 2014. In order to understand the current data management and sharing practices of campus researchers, we solicited, coded, and analyzed 182 DMPs, accounting for 41% of the total number of plans available. **RESULTS** DMPs from seven colleges and academic units were included. The College of Science of Engineering accounted for 70% of the plans in our review. While 96% of DMPs mentioned data sharing, we found a variety of approaches for how PIs shared their data, where data was shared, the intended audiences for sharing, and practices for ensuring long-term reuse. **CONCLUSION** DMPs are useful tools to investigate researchers' current plans and philosophies for how research outputs might be shared. Plans and strategies for data sharing are inconsistent across this sample, and researchers need to better understand what kind of sharing constitutes public access. More intervention is needed to ensure that researchers implement the sharing provisions in their plans to the fullest extent possible. These findings will help academic libraries develop practical, targeted data services for researchers that aim to increase the impact of institutional research.

Received: 02/27/2015 Accepted: 05/06/2015

Correspondence: Carolyn Bishoff, 117 Pleasant Street S.E., 108 Walter Library, Minneapolis MN 55455,
cbishoff@umn.edu



© 2015 Bishoff & Johnston. This open access article is distributed under a Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>)

IMPLICATIONS FOR PRACTICE

1. Research libraries and data specialists can use the method and coding instrument presented in this paper to extract and analyze the data sharing practices contained within data management plans at their institutions.
2. Data sharing practices of researchers are inconsistent overall, and more education and intervention is needed to ensure that researchers implement the sharing provisions in their plans to the fullest extent possible.
3. The findings represent the results from a local study, yet may be generalized to help libraries and data specialists of other institutions develop data service strategies that leverage existing practices based on our results.

INTRODUCTION

Many national and international funding agencies, journal publishers, and research institutions now require researchers to provide reliable access to their research data, and a growing number of federal funding agencies in the United States are asking principal investigators (PIs) to document their plans for describing, storing, securing, sharing, and preserving their research data. In 2015, data management plans (DMPs) are becoming the norm in grant proposals to federal agencies such as the Agency for Healthcare Research and Quality, Department of Energy, NASA, and others (Adler, 2015) in response to the Office of Science and Technology Policy's public access memo (Holdren, 2013). These requirements are largely modeled on those issued by the National Science Foundation (NSF) in 2010, which mandated that all NSF grants submitted after January 18th, 2011 include a 1-2 page DMP (NSF, 2010).

These developments align well with the goals and values of academic libraries. Our mission is to inspire learning and discovery, in part through preserving and supporting the research legacy of the university (University of Minnesota Libraries, 2014). This mission is in part fulfilled by developing a world-class print collection that reflects research objectives of the faculty, but the Libraries are increasingly called on to promote the responsible management and sharing of digital scholarship and research data. When the NSF began requiring PIs to include a DMP explaining how their research data would be shared, the University of Minnesota Libraries responded immediately with training, face-to-face consultation, DMP templates, and best practices for the long-term stewardship and access of data, all which strive to help researchers across campus manage their data responsibly. The Libraries recently launched an institutional data repository (Data Repository for the U of M

(DRUM), 2015) with a new suite of data sharing, archiving, and preservation services that help make it easier for researchers comply with data sharing requirements. In order to identify opportunities to influence the norms of researchers' sharing practices on campus, it is essential to understand their current practices and needs. DMPs allow libraries to look at data sharing intentions that, when supplemented with other user-needs assessments such as surveys and focus groups, can provide a complete picture of data management needs of our campus researchers.

To understand how PIs on campus plan to share their data, the Libraries conducted a review of DMPs from NSF grant proposals accepted from January 2011 through June 2014. During this time period the NSF was the only agency requiring written data management plans in a volume that would achieve the goals of our project. Our review analyzed the content of each DMP at a granular level and quantitatively summarized the researcher's planned strategies for sharing data. The results reveal intended practices across a variety of disciplines and are a significant step toward understanding the sharing practices of researchers at an individual level.

LITERATURE REVIEW

The benefits of data sharing have been widely studied and documented; however, the rationales for sharing research data, which include reproducibility, public access, reusability, and the advancement of science, are deeply intertwined with the complex questions of the nature of the research, the policies that support data sharing, and agreement over the basic definition of data itself (Borgman, 2012). Researchers might choose to share their data for any number of reasons, but only a few factors have been shown to be effective for increasing access to data. Piwowar (2011) identified several factors that increased the likelihood that a researcher would share data, and these include: publication in an open access journal, having a personal history of reusing data, and publishing in a journal with data sharing requirements. However, funding agency mandates do not seem to be one of those factors (Diekema, Wesolek, & Walters, 2014). Vines et al. (2013) found additional evidence that stringent journal data archiving policies dramatically increased the likelihood that a certain data type would be accessible online compared with journals with no policies. Other case studies have shown that researchers value data sharing for personal reasons as well, such as career advancement, soliciting feedback, and facilitating collaboration among peers (Van den Eynden & Bishop, 2014).

The motivations for sharing data, however, are distinct from data sharing practices, and these are significantly less understood outside a small selection of subdisciplines (Borgman,

2012). Part of the problem is the wide variability of data sharing practices, which are highly dependent on the environment in which the data are generated (Akers & Doty 2013). Arzberger et al. (2004) proposed five areas that contribute to an effective climate of data accessibility, including the cultural, institutional, political, technological, and financial domains. Cultural factors, including discipline, age, and geographic region have significant impacts of perceptions of data sharing and reuse (Tenopir et al., 2011). Results from the Data Curation Profiles (DCP) project found a wealth of specific information from a small subset of researchers, such as common data embargo periods, retention timelines, and appropriate audiences for sharing; to date, the project provides the most detailed information available on data sharing practices (Cragin, Palmer, Carlson & Witt, 2010).

Analyses of data management plans have the potential to provide information about data management practices on a granular level, comparable to interview findings such as the DCPs. In the same way that DCP interviews give detailed information about data collection practices, data sharing habits, and tools used in the research process, DMPs provide this same type of information, yet, can be more easily scaled to cover a large pool of researchers. Yet, no studies have utilized DMPs for this level of analysis aimed at data sharing practices. It is widely recognized that NSF grant proposals have generated a critical mass of DMPs and that DMPs are a valuable source of data that can address a variety of research questions. Parham and Doty (2012) measured the adoption of data services, including the local institutional repository (IR) and a DMP template, to assess the impact of data services outreach from their academic library. A multi-institutional project called Data Management Plans as a Research Tool (DART) is developing an analytical rubric from existing NSF DMPs that will support DMP consultation and other research data services at academic libraries (Parham et al., 2015). The forthcoming rubric and analysis focuses on assessing the quality of the DMPs to identify knowledge gaps, barriers to good data management practices, and to inform areas for further education and training (Whitmire et al., 2014).

Two DMP analyses have focused on specific data management practices. A national study combined 167 survey responses and 69 DMPs to identify gaps in sharing and archiving practices (Curty, Kim, & Qin, 2013). Mischo et al. (2014) conducted a more in-depth institutional study of 1,260 DMPs. The study confirmed a number of known data sharing practices, including the practice of sharing data through traditional scholarly publication (Swan & Brown, 2008), and reported on successful outreach over time with regards to references to the template and IR. Additional findings provided evidence that the storage and sharing mechanisms provided in DMPs had no correlation to whether the NSF grant proposal was ultimately funded.

It is clear that DMPs offer a unique opportunity to capture ground-level data from the researchers themselves. Therefore, this study looks to build upon the existing DMP analyses and develop a more comprehensive understanding of data sharing practices, comparable to the information gathered in the Data Curation Profiles project. Based on our local review of DMPs, this paper will focus on the results related to data sharing. Future DMP analyses from the NSF and other federal agencies have the potential to broadly classify data management norms and lead to the development additional outreach strategies within academic institutions that take advantage of existing practices.

METHODS

The University Libraries conducted an institutional review of DMPs that were funded by the NSF from January 2011–June 2014. The plans were presumed to include specific information about the research data file types and the PI's plans and practices for organization, sharing, and archiving the data. This review captured a snapshot of the data sharing practices of a key institutional audience, namely, STEM researchers from a variety of disciplines.

Soliciting DMPs

Initially the Libraries planned to gather DMPs directly from the local grants database under the supervision of the institution's Sponsored Projects Administration (SPA), but privacy concerns and lack of available staffing in SPA prevented this approach. Instead SPA and the Libraries agreed on a method to request the DMPs directly from the PIs via email. SPA provided a list of all NSF grants awarded to the University of Minnesota, a total of 450 since the NSF requirement went into effect. The list included the name of the grant's PI, the department affiliation of the PI, and the date the grant was awarded.

Due to the high volume of STEM plans in the NSF sample, the project team first approached the Associate Dean for Research and Planning in the College of Science and Engineering (CSE) to request support for a college-wide solicitation (see Figure 1, following page). The associate dean sent the email on our behalf to all department chairs in mid June 2014, who forwarded the request to PIs in their departments. This proved to be an effective way to obtain DMPs, and a significant number of the sample were collected soon after this initial request. Liaison librarians were then asked to directly contact researchers in the College of Food, Agricultural, and Natural Resource Sciences (CFANS); the College of Liberal Arts (CLA); the College of Biological Sciences (CBS); and associated professional programs. Three weeks after the initial call, liaisons to all colleges and departments sent follow-up emails to PIs that had not yet responded.

Subject: Sharing NSF Data Management Plans with the U Libraries

TO: Principal Investigator on a National Science Foundation Grant

The University Libraries are conducting an internal review of all data management plans (DMPs) submitted to the National Science Foundation since this requirement went into effect (Jan 2011). We would like obtain a copy of your DMP in order to help us better understand your data management challenges. The review of your DMP will inform the development of robust and targeted data services, both from the libraries and our campus partners, that aim to help you and all campus researchers to comply with increasing federal mandates on sharing and preserving digital research data.

Your Action Needed

Please reply to this email and attach a copy of your two-page Data Management Plan (DMP) submitted to NSF. If you have more than one NSF grant, you are encouraged to attach multiple plans.

Purpose of DMP Review Project

In order to analyze the needs of campus researchers and to help the university develop services to support data storage, access, sharing, and preservation, the University Libraries propose a review of all data management plans (DMPs) submitted to the National Science Foundation since this requirement went into effect (Jan 2011). This review will be for internal university purposes only, they will not be shared beyond the project team, and all identities of grant authors will be protected. The intent is not to critique the plans, but rather to gauge the current data challenges faced by campus researchers.

Further Resources for Your DMP

Data management tools and services exist across campus to help you implement your DMP. Please visit <https://www.lib.umn.edu/datamanagement> to learn about potential services that might be right for your data. The libraries also offer training for staff and students through our online and in-person courses, available at <http://z.umn.edu/datamgmt14>.

Figure 1. Email to Principal Investigators Soliciting a Copy of Their DMP

Response Rate

The response to the call for DMPs from faculty was strong considering the opt-in nature of the study. The Libraries received a total of 182 data management plans emailed directly from PIs to the project team or liaison librarians. This accounted for 41% of the total number of plans available from funded NSF grants in this time period. A considerable amount of credit must be given to the strong ties that the liaison librarians hold with their departmental faculty who responded to the call.

When the DMPs were received via email, they were downloaded and stripped of individually identifying information relating to the grant recipient, including names, grant award IDs, or titles. The file was renamed using a standard file name schema in the form of University_CollegeAbrv_Department_000.ext. Most DMPs arrived as Microsoft Word files or PDFs. However, some DMPs were sent as part of the entire grant application; in those cases, the two-page DMP was extracted, saved as a new file, and the rest of the application was deleted.

Analyzing the DMPs

NSF guidelines for creating a DMP formed the foundation of our instrument used in this analysis. According to the *NSF Grant Proposal Guide* the DMPs should include:

- (1) the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;
- (2) the standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);
- (3) policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;
- (4) policies and provisions for re-use, re-distribution, and the production of derivatives; and,
- (5) plans for archiving data, samples, and other research products, and for preservation of access to them. (National Science Foundation, 2014)

To standardize the DMP review process, an instrument was developed to collect and code DMPs methods using a mix of controlled vocabulary and free-text quotes for five broad categories based on the NSF guidelines. The controlled vocabulary was informed by numerous sources: an internally-developed DMP checklist (UMN Libraries, 2015) and DMP resources developed at other institutions, including Purdue University Libraries (2011), Columbia University Libraries (2014), Johns Hopkins Libraries (2014), Cornell University Libraries

(Wright & Andrews, 2015), as well as the two previously published DMP content analyses (Curry, Kim, & Qin, 2013; Mischo, Schlembach, & O'Donnell, 2014).

We also referred to the draft analytical rubric under development by Whitmire et al. (2014). Our review instrument was created in a Google Form. The full instrument is available as an appendix and as a supplemental file to this article, and the complete dataset generated from the instrument is also available, including but not limited to the data sharing results presented in this paper (Johnston et al., 2015). However, unlike DART, the instrument is intended to review and analyze the content of the plan as-is. The instrument is not intended to critique the plan, create subjective measures of quality, or to provide feedback directly to the researchers. Therefore, our results and conclusions reflect trends that emerged from the *content* of the plans rather than the *quality* of the plans or how well they align with the NSF proposal guidelines.

Limitations

Our sample was limited by two factors: 1) we did not obtain access to the NSF grants database, Fastlane; and 2) our sample of analyzed DMPs were self-selected by the PI. Lack of access to the Fastlane database limited the number of DMPs that were collected in this study as well as limited any details about which NSF directorate to which the DMP was submitted. Instead, direct email solicitation was used which did limit our sample. For the purposes of this analysis the data management practices were the most important and the primary concern is that the sample accurately represents the departmental distribution from the DMPs available.

The methodology also has several limitations. Because the DMPs were written narratives and therefore qualitative data, the contents were potentially subject to interpretation and difficult to consistently code. To mitigate this limitation, the DMPs were reviewed independently by two graduate research assistants that support the libraries' data management and curation services. When each plan was reviewed twice, the authors compared the result of the analyses. If incongruities occurred then the DMP was reviewed again by one of the authors who made a final decision on how the plan should be classified. Finally, the authors selected the DMP as the unit of analysis, rather than each dataset described in the plan. Therefore, the review does not accommodate the description of multiple data sets in a single DMP. To the extent possible, the analysis captures all the methods of sharing that were listed but would not be able to link a specific type of data to a specific sharing method. Instead, the review captured qualitative information from the plan in a free-text quote, to minimize the complexity of the analysis.

RESULTS

College and Department Analysis

The collected DMPs were analyzed by college, department, and specific data management practices. Of the 12 colleges and administrative units that received NSF awards in the specified date range, 7 were represented in this sample (Table 1). With the exception of CSE, most colleges and units were slightly underrepresented in this sample. However, the distribution of our sample across the colleges and units does represent the distribution of NSF grants awarded in the date range with no notable outliers.

College or Unit	DMPs Available	% of Total Available	DMPs Collected	% of Total Collected
Science and Engineering	312	69.33%	146	80.22%
Biological Sciences	46	10.22%	15	8.24%
Liberal Arts	37	8.22%	8	4.40%
Food, Agricultural, and Natural Resource Sciences	17	3.78%	7	3.85%
Academic Health Center	11	2.44%	3	1.65%
Education and Human Development	9	2.00%	2	1.10%
Public Affairs	8	1.78%	1	0.55%
Design	3	0.67%	0	0.00%
Carlson School of Management	3	0.67%	0	0.00%
Administration	2	0.44%	0	0.00%
Public Health	1	0.22%	0	0.00%
Law	1	0.22%	0	0.00%
Total DMPs	450	100%	182	100%

Table 1. DMPs Available vs. Collected, by College: University of Minnesota NSF Grants Funded between 1/18/11-6/25/14

The CSE made up over 80% of the sample and the representation from each CSE department was compared to the DMPs awarded (Table 2). The Earth Sciences department was overrepresented (12.82% of NSF grants awarded vs. 21.92% of the sample), as was the Chemistry department (7.37% awarded vs. 11.64% of the sample). The Computer Science & Engineering department was underrepresented (16.35% awarded vs. 8.22% of the sample). All other departments were represented in the sample within 4 percentage points of how they were represented in the group of NSF grants awarded in the date range.

CSE Department or Unit	DMPs Available	% of Total Available	DMPs Collected	% of Total Collected
Earth Sciences	40	12.82%	32	21.92%
Mathematics	51	16.35%	22	15.07%
Chemistry	23	7.37%	17	11.64%
Electrical Engineering	39	12.50%	13	8.90%
Chemical Engineering & Material Science	24	7.69%	13	8.90%
Computer Science & Engineering	51	16.35%	12	8.22%
Biomedical Engineering	13	4.17%	8	5.48%
Mechanical Engineering	20	6.41%	7	4.79%
Physics	16	5.13%	7	4.79%
Aerospace Engineering	4	1.28%	7	4.79%
Civil Engineering	13	4.17%	4	2.74%
Industrial & Systems Engineering	3	0.96%	2	1.37%
St. Anthony Falls Laboratory	7	2.24%	1	0.68%
Astronomy	5	1.60%	1	0.68%
Structural Engineering	2	0.64%	0	0.00%
Geomechanics	1	0.32%	0	0.00%
Total CSE DMPs	312		146	

Table 2. DMPs Available vs. Collected, by CSE Department: University of Minnesota NSF Grants Funded Between 1/18/11-6/25/14

Mathematics made up a significant, but not disproportionate, percentage of our sample. It was expected that Mathematics would have fewer data to manage than other disciplines and their DMPs may state that no data would be collected for a particular grant (Division of Mathematical Sciences, n.d.). There was some concern that the Mathematics DMPs in the sample would influence the results of this review in such a way that would minimize certain data management practices in other departments. To mitigate this effect, if a DMP stated that no data would be produced from the grant, they were given “n/a” classifications in several of our analysis questions.

Data Sharing Methods

Data sharing was mentioned in 96% (176) of the DMPs (Figure 2).

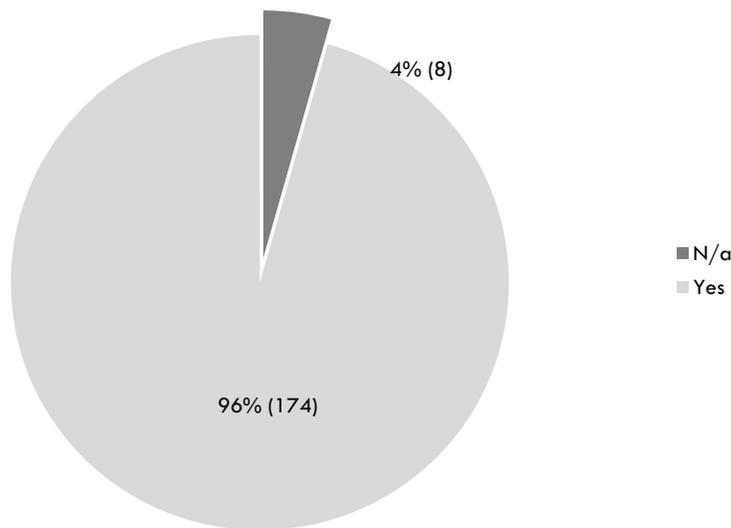


Figure 2. Does the DMP mention data sharing? (n=182)

A single DMP often includes several venues for sharing. In the sample, sharing methods of various types were mentioned 415 times and fell into nine major categories: publications (specific), publications (non-specific), conference presentations, theses or dissertations, personal websites, other websites, disciplinary repositories, local institutional repositories (IRs), and sharing by request (Figure 3, following page).

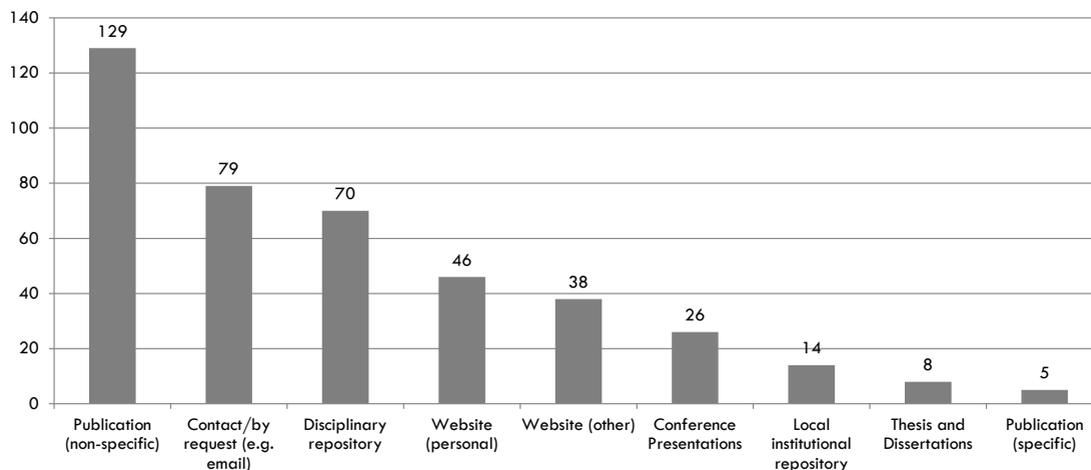


Figure 3. Methods of Data Sharing (n=145)

The most common venue for sharing data that appeared in 74% (135) of the 182 plans was the traditional journal publication (named or non-specific), or more accurately, via components of a journal publication (tables, graphs, images, etc.). This practice is frequently described in DMPs using general terms, such as sharing or disseminating the results of the research project through publication in peer-reviewed journals and conference presentation. Some DMPs mentioned specific publication outlets (e.g. the *Journal of Molecular Spectroscopy*) or publishers (e.g. American Chemical Society journals), but most did not. Other traditional publication venues for sharing data included theses and dissertations (usually when the grant proposal funded graduate research) and conference proceedings. Newly emerging data publication formats, such as the article-type known as a “data descriptor” in Nature’s Scientific Data journal, were not mentioned in our sample.

Providing data on request, the second most common form of sharing, was mentioned in 43% (79) of the plans. However, in all but 3 cases, this method was in addition to other forms of sharing. For example, some plans indicated that after the research is published, interested parties could request data directly from the PI.

Websites were another common venue for sharing. Personal websites most often included dedicated project websites or the PI’s personal website. A few DMPs mentioned that some data would be posted on a student’s website, and several specified that only a select few types of research data would be posted (e.g. MATLAB files or software code but not laboratory notebooks). Some researchers planned to track page views and downloads of their data,

while others only guaranteed that the files would be accessible for a limited duration (e.g. maintained for one year after the project is completed).

Other websites (non-personal) typically included project websites managed by another university. These websites were usually described as “central” or “partner” websites. Some DMPs referred to Github or Sourceforge, which were typically categorized as “other websites” unless the DMP indicated that they were considered a disciplinary repository for their field (e.g. arXiv.org, mentioned in 15 DMPs). Similarly, some DMPs referred to websites that host curriculum materials and seemed distinct from a disciplinary data repository.

The local institutional repository refers to the University Digital Conservancy (UDC) (<http://conservancy.umn.edu>), the IR at the University of Minnesota. If another university’s IR was mentioned it was coded as “other” website.

Access Levels Based on Sharing Method

Each method of sharing listed in the DMP sample was approximated to provide a different level of access to different audiences. Table 3 shows our pairing of the sharing method with the audience. The two “public” access categories emerged from language used in the DMPs.

Sharing Method	Level of Access
website (personal or other)	public/unrestricted
public repository (disciplinary or local)	public/unrestricted
available by request	public/available by request
publication (specific or non-specific)	researchers in any field
theses and dissertations	researchers in any field
conference presentations	researchers in PI’s field

Table 3. Sharing Method and the Associated Level of Access

About 40% of all 415 references to data sharing practices would potentially make the underlying data publicly accessible. PIs referenced a disciplinary repository 70 times (e.g. GenBank, Dryad, or the Magnetism Information Consortium/MagIC), a website (personal or otherwise) 84 times, and the local IR 14 times (Figure 4, following page).

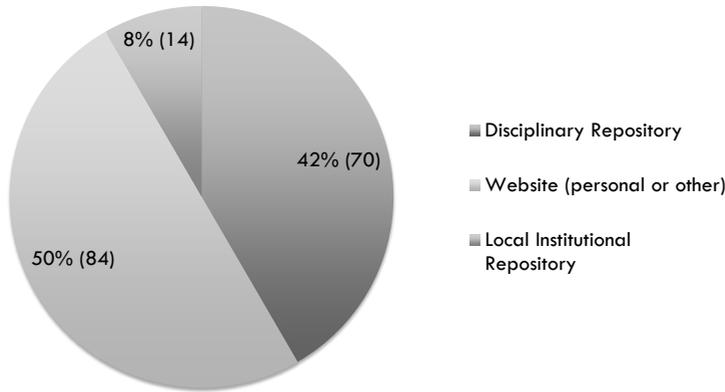


Figure 4. Proportion of DMP Data Sharing Practices That Would Likely Result in Publicly Accessible Data (by method, n=168)

On the other hand, about 60% of all data sharing strategies (n=247) would potentially make data inaccessible for certain audiences. For example, most traditional publications are subscription-based, and, therefore, those who do not subscribe to the journal or are unaffiliated with a research university would not have open, free access to the publicly-funded research data. Plans mentioned a specific or non-specific publication 134 times, conference presentations 26 times, and theses or dissertations 8 times. Finally, sharing through direct request to the PI, which has been shown to be an unreliable method of access (Savage & Vickers, 2009), was mentioned 79 times (Figure 5).

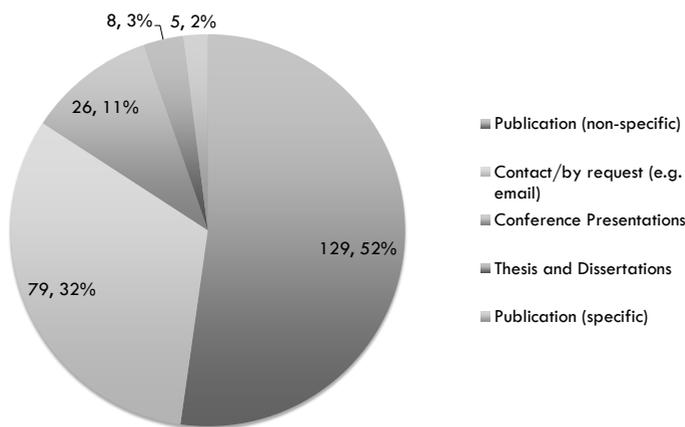


Figure 5. Proportion of DMP Data Sharing Practices That Would Likely Result in Limited Access or Inaccessible Data for Certain Audiences (by method, n=247)

Audience for Reuse

DMPs also often mentioned the intended audience for data sharing (Figure 6), and again, some plans referenced more than one audience. Of the 182 plans reviewed, 131 plans named at least one intended audience for a total of 202 mentions.

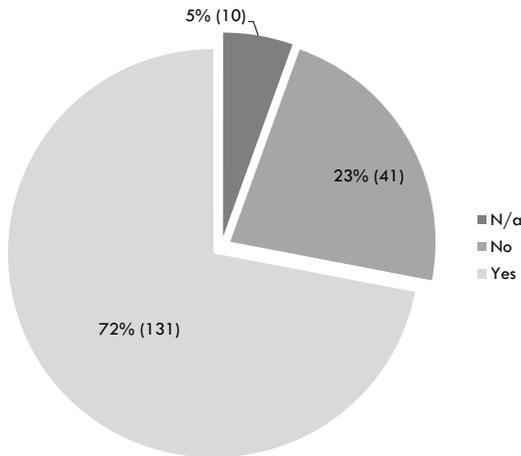


Figure 6. Does the DMP mention an audience for sharing data? (n=182)

Direct references to target audiences for data sharing (n=202) fell into the four categories listed in Table 3 plus two additional categories that emerged from the DMPs: project team and students (Figure 7).

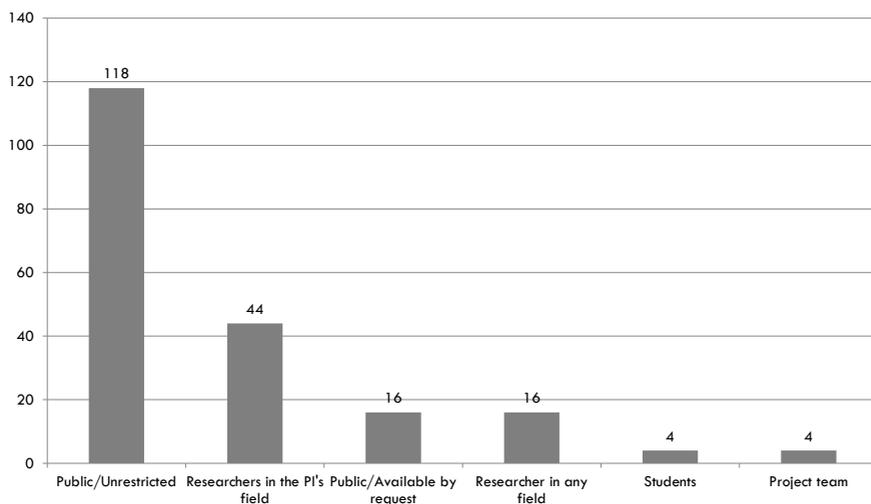


Figure 7. Categories of Intended Audiences for Sharing (n=202)

Of all the audience types mentioned in the sample, 58.4% (118) fall into the public/unrestricted category while 41.6% (84) fall into a more specific audience category: peers, students (either participants in a course or students supported by the grant), members of the project team, or anyone who requests the data.

Sharing Timeline and Retention Period

Less than half of the PIs include a timeline for sharing (43.4%, 79), and even fewer (29.7%, 54) specify a period of data retention (Figure 8 and 9 respectively).

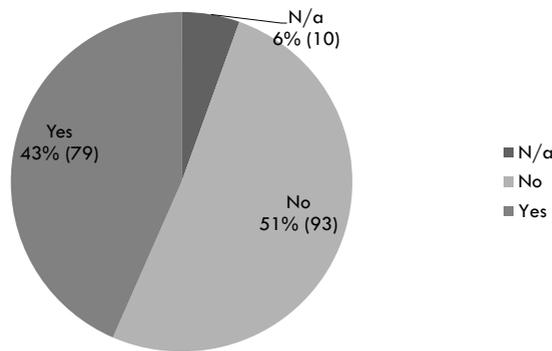


Figure 8. Timeline for Data Sharing Mentioned? (n=182)

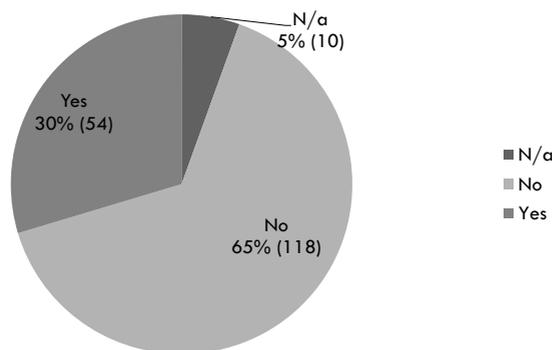


Figure 9. Time Period of Data Retention Mentioned? (n=182)

The point in the research lifecycle at which a PI plans to share their data is highly variable and difficult to categorize. A general grouping of the 79 free-text quotes that indicated when PIs would share revealed 91 timelines for sharing. These were categorized into seven groups, as shown in Figure 10. This inspection showed that most PIs that indicated a timeline are willing to share data after the research is published.

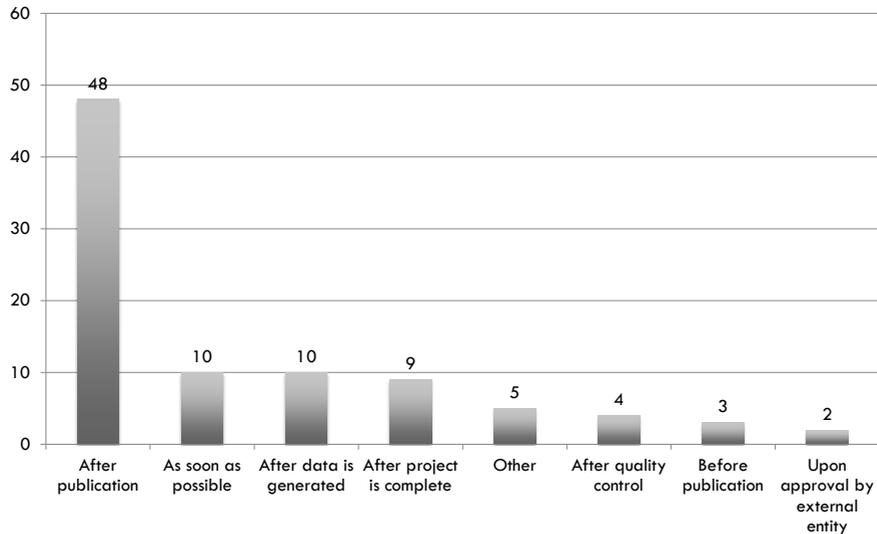


Figure 10. Distribution of Timelines Mentioned for Data Sharing (n=91)

Data retention periods vary widely as well, but PIs in the sample most commonly plan to keep their data for either three years or indefinitely (Figure 11, following page). Three years is the minimum retention period for the NSF Engineering Directorate (Directorate of Engineering, n.d.), and a few DMPs refer to these NSF guidelines rather than specifying a length of time. However, it is unclear when most retention periods are intended to start and likely differs based on the nature of the project and advice from the NSF directorate or division.

Private Data

Even within a sample that typically does not focus on human subjects research, nearly 18% (33) of the DMPs included one or more mentions of private or sensitive data. In addition, the examples of private data spanned a number of different data types (Figure 12), but the most common were personally identifiable information (PII) from human subjects research (12%, 22), and proprietary data (3%, 6). A couple idiosyncratic concerns surfaced as well including sensitive locations and cultural data. In such cases, the DMP indicated that specific sensitive information would be removed from the final dataset prior to sharing.

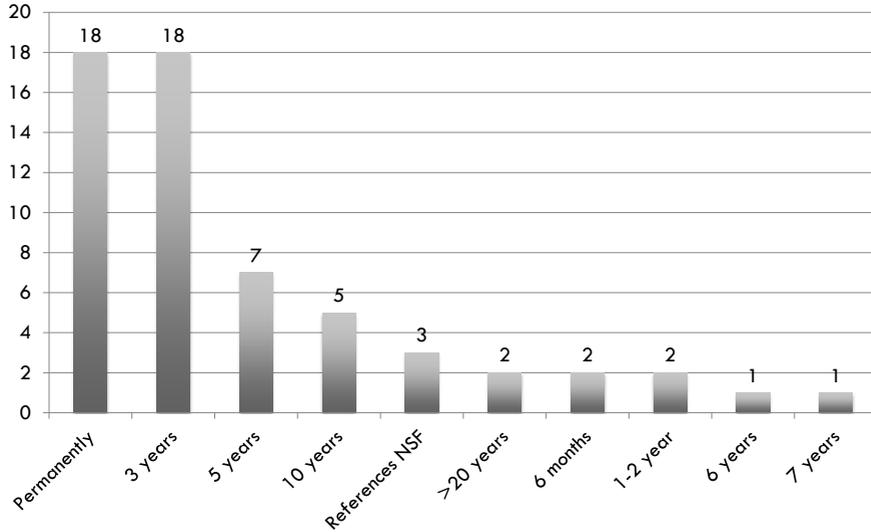


Figure 11. Distribution of Retention Periods Mentioned (n=59)

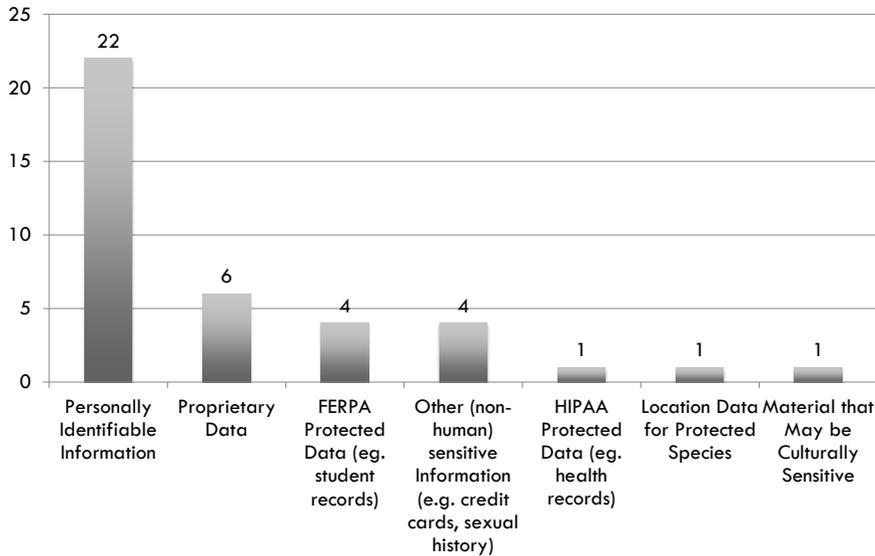


Figure 12. Categories of Private or Sensitive Data Mentioned (n=39)

Additional access concerns emerged in 29% (52) the sample regarding data ownership and/or intellectual property concerns. For example, DMPs mentioned data that may result in patentable information and thus be withheld until patent applications have been filed (Figure 13, following page). When researchers explained to whom they would disclose these

inventions, however, the plans varied between the University and the NSF. For example, one researcher stated that “Data acquired [are] subject to University of Minnesota intellectual property management policies. If discoveries/inventions [are] made, data requests will be granted after disclosures and filings are made,” while another states that in addition to University intellectual property guidelines, “...any inventions must be disclosed to NSF before filing [patents].”

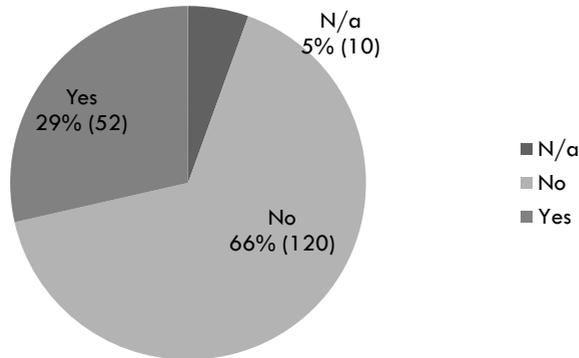


Figure 13. Data Ownership and/or Intellectual Property Rights Mentioned (n=182)

Long-Term Sharing and Archiving

Long-term data archiving plans may have an impact on how researchers share their data over time. Data archiving plans were included in 80% (145) of the DMPs and ranged from well-defined digital archives to more ad hoc, individual techniques (Figure 14, following page).

Digital archives designed to share and preserve data were mentioned in 47% of the plans including locally-run data repositories (n=24), institutional repositories (n=22), and disciplinary data repositories (n=40), such as the National Center for Biotechnology Information (NCBI) database, NASA data archives, the Inter-University Consortium for Political and Social Research (ICPSR), and the NOAA National Geophysical Data Center. Individual archiving techniques appeared in 25% (46) of DMPs such as storing the data in external hard drives or moving data to a remote server to be archived after the conclusion of the project.

Many DMPs proposed the same practices for sharing and archiving, often using a variation of the language: “data will be archived in the same way that it is stored.” This approach was mentioned in 32% (58) of the DMPs in our study, and in at least 11.5% (21) of the plans, this was the only archiving approach mentioned.

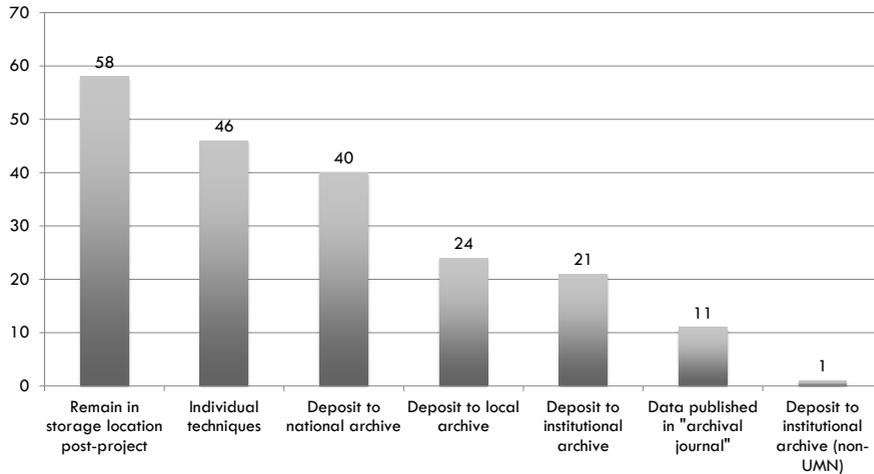


Figure 14. Categories of How PIs Plan to Archive Their Data (n=201)

University Services

University services, including data services available from the Libraries, were mentioned in 36% (65) of plans. The UDC, our campus IR, was mentioned in 11.5% (21) of the plans, while the University Libraries was mentioned in 2.2% (4) of the plans. Most PIs who mentioned the IR did so in the context of preservation, and almost none referred to it as a venue for sharing and access, even though the repository is public and its holdings are exposed to web crawlers. A variety of other campus services for data analysis, storage, and training were mentioned as well, including the local high-performance computing center (n=17), local file storage providers (n=3), and secure data training offered by the Office of Information Technology (n=1).

DISCUSSION

The 182 DMPs analyzed in this paper, with the exception of those that did not report any data to manage, represent a small but important sample of proposed data management practices at the University of Minnesota. The distribution of the sample across colleges and departments aligns with the actual distribution of NSF grants from that time period and provides a representative sample of the proposed data management practices on our campus, particularly in STEM disciplines. Overall, the results of our study find that DMPs illustrate a variety of data sharing approaches, timelines for retention and sharing, concerns with private data, and opportunities for additional services and partnerships.

Sharing Strategies and Access

Sharing strategies were found in nearly all of the DMPs in this sample and were even included in some DMPs for projects that would not generate data. This is most likely a reflection of the NSF guidelines which state that the primary purpose of the DMP is to “describe how the proposal will conform to NSF policy on the dissemination and sharing of research results” (NSF, 2014) which emphasizes public distribution. However, the analysis of granular, individual sharing plans indicates that there are widespread differences regarding the accessibility afforded by different methods of sharing. When an audience was mentioned in a DMP, it included the general public nearly 60% of the time. However, when a sharing method was mentioned, it would only make the data publicly accessible only about 40% of the time.

We suspect that this discrepancy has to do with PI perceptions of the various sharing methods. This study confirms what Mischo et al. (2014) and others have shown: data sharing happens primarily through traditional publishing channels. Researchers largely view journal publications and conference proceedings as a form of public sharing, even though most journals and conference proceedings are pay-to-view. This not only affects access to the general public and any researchers at institutions that do not subscribe to the publication, but has been shown to disproportionately affect low-income countries as well (Aronson, 2004). Furthermore, language in this DMP sample indicates that PIs rely on these publishing channels to provide even more services than are typically available; for example, 12 DMPs identified “archival publications” as their chief archiving strategy.

The DMPs showed that certain data types were not included in the sharing plan. For example, even though laboratory and field notebooks were often mentioned elsewhere in the DMPs as a data type, a method for data storage, and a means of data documentation, they were conspicuously absent from the DMPs’ data sharing strategies. Unfortunately, this may indicate that there is a wealth of original data and descriptive information that is potentially not shared. Electronic lab notebooks (ELNs) present a possible solution to this problem by allowing PIs and members of their research team export the shareable observations and data to a sharing-friendly format, yet ELNs were only mentioned nine times as a data type.

Timelines for Sharing and Retention

Timelines for sharing and retention were mentioned in fewer than half the DMPs. The absence of defined retention periods may reflect the NSF’s lack of guidance in this area since when a directorate does provide concrete guidelines, in the case of the Engineering

Directorate's minimum retention requirement of three years, this study found that many PIs did include the minimum retention period or referred to it indirectly. Since the PIs in this study appear to be receptive to well-defined guidelines for retention there may be an opportunity for data repositories to fill that gap and provide guidance for PIs that seek data sharing solutions with longer retention periods. Further analysis is needed, though, on those DMPs that did include a retention period, to determine whether the retention period reasonably aligns with the stated archival techniques, which may be different than the PI anticipated.

Private Data

Private and sensitive data present some serious challenges for sharing and access, and DMPs with private data occurred more frequently than expected. While some of these DMPs came from fields associated with the Social, Behavioral, and Economic Directorate of the NSF, they also came from engineering, computational sciences, and natural sciences. Privacy concerns may also appear in chemistry and chemical engineering in the form of proprietary data concerns. These findings demonstrate that while sensitive data are often associated with the health and social sciences, any service for the management and sharing of sensitive data must encompass a variety of disciplines and concerns.

NEXT STEPS: OPPORTUNITIES FOR SERVICES AND PARTNERSHIPS

The University Libraries offer many research data services, including templates and DMP training. Since 2010, over 774 faculty, researchers, and students have attended data management training, including a workshop titled "Creating a Data Management Plan for a Grant Application." Recently, the Libraries launched a data repository and curation service that includes new boilerplate language for PIs to include library services in their DMPs. The results of this study will contribute to the ongoing evolution of library data services and help frame existing services in the context of data practices on campus.

Sharing data through publication is a widespread practice in this sample of DMPs. While this is a limiting method or the overall public accessibility of data, it also presents an outreach opportunity for libraries. Data specialists and librarians can take on a more active role in identifying data sets for sharing and preservation, such as using publication alerts to prompt PIs about data sharing options. Such services would be relatively easy to implement, leverage the relationships between liaisons and faculty, and help build awareness across campus of good data sharing practices.

The importance of lab notebooks to the research process cannot be overstated, but if the medium does not accommodate good sharing practices, much of the shareable information

contained in them will remain inaccessible. The frequent mention of lab notebooks in DMPs raises the question of how to facilitate better sharing of this content and might suggest the value of ELNs for data sharing and access due to their digital nature. This may provide another favorable data point for institutions to better support ELNs at a campus-wide level.

An existing library service, the UDC, was mentioned in 8% (15) DMPs, but it is uncertain if any of the projects ultimately deposited research data or publications into the IR. This study has opened a discussion on how the repository staff can develop active collaborations with PIs who identify the IR as a potential sharing and archiving platform. The new Data Repository for the University of Minnesota is built on the existing IR and is designed to allow researchers to self-deposit their data for review by a data curation specialist before being finalized for publication in the IR. With better coordination, IR staff could provide occasional training sessions with grant recipients to encourage consistent data management practices throughout the project and provide guidance on the IR's policies for acceptance, including appropriate levels of data documentation and the exclusion of non-private or protected information from in the data files. Collaboration at the beginning of a research project would help facilitate ongoing, active discussions about these submission requirements that PIs and research staff could keep in mind as the data is collected, managed, and ultimately published through the IR.

To further support the promotion of library data services, the Libraries should rely on collaborative relationships between subject liaisons and researchers. For example, this study obtained a representative sample through strong liaison relationships with departments and colleges. This method of direct email solicitation had its limitations, including a potentially smaller pool of DMPs, but one advantage was the promotion of existing University Libraries' data management services (Figure 1). By asking liaison librarians to send these requests directly to their departments, the study provided a potential opportunity for them to get involved with the PIs research data management issues.

On the other hand, the variety of existing campus services presents a challenge. The library liaison network may act as an efficient referral system within the Libraries, but the results of this study identified a need to build more extensive referral networks outside the library and with other campus units. The Libraries must adapt existing referral models to connect PIs with communities, beyond the Libraries, that provide expertise in sensitive data management or platforms like Github.

Other potential partnerships could have more urgent goals of preventing data loss and obsolescence. As this study and others (Mischo et al., 2014; Curty et al., 2013) have shown,

personal and project websites are often used to house research data, but these are not ideal for long-term sharing and archiving (Goodman et al., 2014). It may be possible to work with the local web server administrators to actively identify PIs who use their websites for data archiving or develop other strategies to mitigate this long-term data storage and sharing problem.

CONCLUSION

This study included DMPs from NSF grants awarded to the University of Minnesota from 2011-2014. The methods of analysis were granular enough to uncover a range of data management and sharing approaches that PIs plan to use in their research. Data sharing strategies of researchers are inconsistent overall, and more education and intervention is needed to ensure that researchers implement the sharing provisions in their plans to the fullest extent possible. Specifically, there was evidence in the DMPs that demonstrates the need for a better understanding of what constitutes sharing to the public. Data services in the Libraries and campus-wide can address specific data sharing concerns, such as retention guidelines, data access issues through inadequate sharing venues such as journal publications and personal websites, and better support of researchers with private data concerns. Thus, the short-term goal will be to see more campus services mentioned in DMPs. This evolution of DMP support will take on even greater importance, beyond NSF, as more federal funding agencies are required to implement additional data sharing requirements. Finally, Libraries must be ready to expand their services from helping PIs plan to share their data to strategically facilitating the successful sharing and ultimate reuse of data.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the graduate students who supported this study: Anders Swendsrud, who helped create the analysis instrument, and Chris Storino and John McGrory, who reviewed the data management plans.

REFERENCES

Adler, P. (2015). *AHRQ, NASA, USDA release plans for public access to funded research*. Community Updates, Association of Research Libraries. Retrieved from <http://www.arl.org/news/community-updates/3532-ahrq-nasa-usda-release-plans-for-public-access-to-funded-research>

- Akers, K. G., & Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Digital Curation*, 8(2), 5-26. <http://dx.doi.org/10.2218/ijdc.v8i2.263>
- Aronson, B. (2004). Improving online access to medical information for low-income countries. *New England Journal of Medicine*, 350(10), 966-968. <http://dx.doi.org/10.1056/NEJMp048009>
- Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., ... Wouters, P. (2004). Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, 3, 135-152. <http://dx.doi.org/10.2481/dsj.3.135>
- Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63, 1059-1078. <http://dx.doi.org/10.1002/asi.22634>
- Columbia University Libraries. (2014). Data Management Plan Self-Assessment Questionnaire, adapted from Purdue University Libraries (2011). *Data management plan templates - Scholarly communications program*. Retrieved from http://scholcomm.columbia.edu/wp-content/uploads/2014/04/DMPSelfAssessment_v01.pdf
- Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 368(1926), 4023-4038. <http://dx.doi.org/10.1098/rsta.2010.0165>
- Curty, R., Kim, Y., & Qin, J. (2013). What have scientists planned for data sharing and reuse? A content analysis of NSF awardees' data management plans. Presented at the *Research Data Access & Preservation Summit 2013*, Baltimore, MD. Retrieved from <http://surface.syr.edu/ischoolstudents/2/>
- Data Repository for the U of M (DRUM). (2015). University of Minnesota Digital Conservancy. Retrieved from <http://hdl.handle.net/11299/166578>
- Diekema, A. R., Wesolek, A., & Walters, C. D. (2014). The NSF/NIH effect: Surveying the effect of data management requirements on faculty, sponsored programs, and institutional repositories. *Journal of Academic Librarianship*, 40(3-4), 322-331. <http://dx.doi.org/10.1016/j.acalib.2014.04.010>
- Directorate of Engineering. (n.d.). *Data management for NSF Engineering Directorate proposals and awards*. National Science Foundation. Retrieved from http://nsf.gov/eng/general/ENG_DMP_Policy.pdf
- Division of Mathematical Sciences (DMS), Directorate of Mathematical and Physical Sciences. (n.d.). *Advice to PIs on data management plans*. National Science Foundation. Retrieved from <http://www.nsf.gov/bfa/dias/policy/dmpdocs/dms.pdf>
- Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., ... Slavkovic, A. (2014). Ten simple rules for the care and feeding of scientific data. *PLoS Computational Biology*, 10(4), e1003542. <http://dx.doi.org/10.1371/journal.pcbi.1003542>
- Johns Hopkins Sheridan Libraries. (2014). *Questionnaire to help with the creation of a data management plan*. JHU Data Management Services. Retrieved from <http://dmp.data.jhu.edu/assistance/nsf-data-management-plans>

Johnston, L. R., Bishoff, C., McGrory, J., Storino, C., & Swendsrud, A. (2015). *Analyzed data management plans (DMPs) from successful University of Minnesota grants from the National Science Foundation, 2011-2014* [dataset]. Retrieved from the Data Repository for the University of Minnesota. <http://dx.doi.org/10.13020/D6TG6Z>

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 e-Science Librarianship*, 3(1), 31-43. <http://dx.doi.org/10.7191/jeslib.2014.1060>

National Science Foundation. (2010). *Dissemination and sharing of research results*. Retrieved from <http://www.nsf.gov/bfa/dias/policy/dmp.jsp>

National Science Foundation. (2014). Chapter II, Section C.2.j: Special information and supplementary documentation. *Grant proposal guide*. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_2.jsp

Parham, S. W., Carlson, J., Hswe, P. M., Rolando, L., Westra, B. & Whitmire, A. (2015). Development of an analytic rubric to facilitate and standardize the review of NSF data management plans. Poster presented at the *International Digital Curation Conference 2015*, London, England, UK. Retrieved from <http://dmpresearch.library.oregonstate.edu/presentations>

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://dx.doi.org/10.1002/bult.2012.1720390113>

Piowar, H. A. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLoS ONE*, 6, e18657. <http://dx.doi.org/10.1371/journal.pone.0018657>

Purdue University Libraries. (2011). *Data Management Plan Self-Assessment Questionnaire*. Purdue University, West Lafayette IN. Retrieved from https://www.purdue.edu/research/vpr/rschdev/documents/DMP_Self-Assess_14Feb2011.pdf

Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PLoS One*, 4(9), e7078. <http://dx.doi.org/10.1371/journal.pone.0007078>

Swan, A., & Brown, S. (2008). *To share or not to share: Publication and quality assurance of data research outputs*. Research Information Network. Retrieved from <http://www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs>

Tenopir C., Allard S., Douglass K., Aydinoglu A. U., Wu, L., Reid, E., ... Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, 6, e21101. <http://dx.doi.org/10.1371/journal.pone.0021101>

University of Minnesota Libraries (2014). *Vision, mission, and goals*. Retrieved from <https://www.lib.umn.edu/about/goals>

- University of Minnesota Libraries. (2015). U of M DMP template (Google Doc). *Creating a data management plan*. Retrieved from <https://www.lib.umn.edu/datamanagement/DMP>
- Van den Eynden, V. & Bishop, L. (2014). *Incentives and motivations for sharing research data, a researcher's perspective. A Knowledge Exchange report*. Retrieved from <http://knowledge-exchange.info/Default.aspx?ID=733>
- Vines, T. H., Andrew, R. L., Bock, D. G., Franklin, M. T., Gilbert, K. J., Kane, N. C., ... Yeaman, S. (2013). Mandated data archiving greatly improves access to research data. *FASEB Journal*, 27,1304-1308. <http://dx.doi.org/10.1096/fj.12-218164>
- Whitmire, A. L., Carlson, J., Hswe, P. M., Parham, S. W., Rolando, L., & Westra, B. (2014). Project overview. *Data management plans as a research tool (DART) project*. Retrieved from <http://dmpresearch.library.oregonstate.edu/project-overview>
- Whitmire, A. L., Carlson, J., Hswe, P. M., Parham, S.W., Rolando, L., & Westra, B. (2014, Unpublished Draft). Rubric for assessment of NSF data management plans. Personal Communication.
- Wright, S. J., & Andrews, C. (2015). Developing a for-credit course to teach data information literacy skills: A case study in natural resources. In J. Carlson & L. Johnston (Eds.), *Data information literacy: Librarians, data, and the education of a new generation of researchers* (pp. 94-96). West Lafayette, IN: Purdue University Press.