Journal of eScience Librarianship

Volume 3 | Issue 1 Article 3

2014

An Analysis of Data Management Plans in University of Illinois National Science Foundation Grant Proposals

William H. Mischo
University of Illinois at Urbana-Champaign, w-mischo@illinois.edu

Mary C. Schlembach
University of Illinois at Urbana-Champaign, schlemba@illinois.edu

Megan N. O'Donnell
Iowa State University, mno@iastate.edu

Recommended Citation

Mischo, William H.; Schlembach, Mary C.; and O'Donnell, Megan N.. (2014). "An Analysis of Data Management Plans in University of Illinois National Science Foundation Grant Proposals." *Journal of eScience Librarianship* 3(1): Article 3. http://dx.doi.org/10.7191/jeslib.2014.1060

JESLIB 2014; 3(1): 31-43
doi:10.7191/jeslib.2014.1060

Journal of eScience Librarianship
putting the pieces together: theory and practice

An Analysis of Data Management Plans in University of Illinois National Science Foundation Grant Proposals

William H. Mischo,[1] Mary C. Schlembach,[1] Megan N. O'Donnell[2]

[1] University of Illinois at Urbana-Champaign, Champaign, IL, USA
[2] Iowa State University, Ames, IA, USA

Abstract

The University of Illinois at Urbana-Champaign (UIUC) Library conducted an analysis of 1,260 Data Management Plans (DMP) submitted in National Science Foundation (NSF) proposals from July 2011 through November 2013. Each DMP was assigned controlled vocabulary and keyword terms which summarized the proposed data management mechanisms for storing and sharing data. A database composed of the proposal's title, PI, PI's department and college, NSF grant number, funded status, assigned DMP vocabulary, and keyword terms was constructed. As of May 2014, a total of 298 of these UIUC proposals had been funded by the NSF. Our analysis of this sample revealed no significant statistical differences in proposed data storage and reuse venues between funded and unfunded proposals. However, there was a statistically higher frequency of use of the campus institutional repository and disciplinary repositories in proposals submitted after October 2012.

Introduction

On January 18, 2011, the National Science Foundation (NSF) started requiring the submission of a Data Management Plan (DMP) in all NSF grant proposals. In July 2011, the University of Illinois at Urbana-Champaign Library and the campus Committee on Data Stewardship, working in conjunction with the campus Office of Sponsored Programs and Research Administration (OSPRA) began an ongoing analysis of the Data Management Plans (DMPs) in newly submitted NSF grant proposals. Our Office of Sponsored Programs and Research Administration (OSPRA) serves as the Authorized Organizational Representative (AOR) to NSF. AORs determine authorization for access and signatory permissions to an institution's NSF grants and grant applications. In addition to approval by the OSPRA office, approval for the analysis and publication of the data in this study was given by the Office of the Vice-Chancellor for Research and the University of Illinois Institutional Review Board. Detailed confidential identifying information is not included in the analysis as compiled data is at the research department or college-level only.

This analysis was undertaken to provide the Illinois campus and Library with information on the DMPs being submitted by Illinois researchers. In particular, the analysis allows us to classify the DMPs in grant applications with regard to their proposed data storage

Correspondence to William Mischo: w-mischo@illinois.edu
Keywords: National Science Foundation grants, data management plans, eScience, libraries, institutional repositories, discipline repositories

venues and data reuse mechanisms. One major accomplishment of the analysis has been to develop a controlled vocabulary of subject descriptors and to identify keywords used by NSF proposal preparers in the DMP text.

The analysis also provides us with data on the use of DMP templates, developed both by the University of Illinois Library and the national library community. A number of university libraries have developed custom data management templates (Schlembach and Brach, 2012), and a consortium of libraries has developed the original DMPTool and the DMPTool2 (https://dmptool.org/partners_list). This detailed analysis has allowed us to engage in a dialog with university administrators regarding the creation of a campus-wide research data service and to develop campus-wide tools and services that can be used by Illinois researchers to manage their data, provide access to it through dataset publication, and to develop best practices and standard approaches for data curation and management.

While the DMP study is ongoing, this paper reports on results obtained from an analysis of 1,726 NSF proposals submitted between July 2011 and November 2013. After eliminating multi-institutional collaborative proposals that did not contain the overarching proposal DMP and supplemental, travel, and update grants, the analysis was conducted over 1,260 actual proposals. A relational database with records containing the grant proposal's title, PIs, the PI's department and college, NSF grant number, funded status, assigned DMP descriptors, and keywords from the DMP was constructed for this study.

The large number of proposals, gathered over a 28 month time period, provided the project team with an adequate sample to test the comprehensiveness of the controlled vocabulary, identify and normalize the natural language keywords, and conduct a longitudinal study of the differences between funded and unfunded proposals. The data gathered in our analysis allowed us to assess whether there were any data storage venues, such as local institutional or disciplinary repositories, which were statistically utilized more often in funded proposals.

Background

NSF DMP Requirements

Since January 2011 the National Science Foundation (NSF) has required a two-page supplementary "data management plan" (DMP) to be submitted as part of the grant-proposal process.[1] The minimum requirements of a DMP vary by NSF subject directorate and, in some cases, divisions within a directorate, but all of them require proposals to "describe how the proposal will conform to NSF policy on the dissemination and sharing of research results."[2] In the cases where the grant proposal will generate no data, such as theoretical work or for travel grants, it is acceptable for the Principal Investigator (PI) to submit a DMP that states that no data will be generated.

One of the issues with the current NSF DMP requirement is that very broad language is used within the guidelines. DMP requirements rely on "communities of interest" to govern the standards and best practices. In cases where disciplines are widely varied, or span more than one NSF subject directorate, or lack disciplinary repositories, standards and best practices are difficult to establish. NSF guidelines also indicate that funding agencies expect common practices to evolve out of compliance with the DMP requirements. This does not currently provide clarity to proposal writers. However, this is not a problem unique to NSF. The U.S. Department of Energy, U.S Environmental Protection Agency, Sloan Foundation, and other funders also have very imprecise DMP requirements (Dietrich, Adamus, Miner, &

[1] http://www.nsf.gov/nsb/meetings/2010/0504/minutes.pdf pg. 17 under "CSB Task Force on Data Policies (DP)"
[2] http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Steinhart, 2012).

There is also some confusion over the types of data that should be included in a DMP. *Chapter II, section d subsection i of the NSF Grant Proposal Guide*[3] states that "the products of research, including preservation, documentation, and sharing of data, samples, physical collections, curriculum materials and other related research and education products" should be documented in a data management plan. However, there are additional guidelines for each Directorate. These guidelines often contradict the definition of data laid out in the Grant Proposal Guide. For example, the Directorate for Biological Sciences (BIO) notes that "physical objects such as laboratory samples" are not within the overarching NSF definition of data. The NSF Data Management & Sharing Frequently Asked Questions (FAQs) webpage extends the definition of data to publications and models – formats/types which are not discussed in the Grant Proposal Guide.[4]

While the DMP requirements strongly encourage data sharing, it is primarily a data management plan, not a data sharing plan (Borgman, 2012). Historically PIs acted as "gatekeepers" with their data, letting only specific colleagues gain access. The establishment of both disciplinary and institutional data repositories has permitted researchers to more easily share data. And yet data sharing remains low across disciplines (Borgman, 2012; Cragin, Palmer, Carlson, & Witt, 2010; Mischo & Schlembach, 2011).

Literature Review

There are few published studies that examine the contents of submitted NSF Data Management Plans. NSF funding is extremely competitive. Of the approximately fifty thousand grant applications NSF received in 2013, only 24% of them were funded (National Science Foundation, 2013). Funding rates differ among Directorates. The lowest 2012 funding rate was 18% from the Engineering Directorate, while the Office of Polar Programs was highest at 36% (National Science Foundation, 2013). Submitted grant proposals are reviewed and evaluated on criteria of Intellectual Merit and Broader Impacts. DMPs are evaluated on the same criteria and as such, should be considered an important part of the application process.

Because the DMP requirement is relatively new, and PIs are unfamiliar with writing DMPs, they often experience a high degree of uncertainty about what should be included in a DMP plan and often reach out to colleagues for help. A survey at the Georgia Institute of Technology, initiated in the fall of 2010 before the NSF DMP requirement went into effect, found that 40% of the respondents thought DMPs were unnecessary and 47% didn't know enough about them to include or create one (Parham et al., 2012). Steinhart et al. (2012) surveyed researchers at Cornell University about their understanding of data management and the NSF requirements and found that the choice "I'm not sure" was chosen more than 20% of the time on questions where it was an option. Researchers seem to grasp the importance of data management but remain divided over how challenging it is to create and execute data management plans (Curty et al., 2013).

PIs may choose to share DMPs with their peers. Parham and Doty (2012), who were able to analyze 181 DMPs from Georgia Institute of Technology researchers, found that about a third of the respondents were reusing text from data management plans that had been shared with other proposal writers. The same study also found that most of the sharing was occurring between pairs of researchers and that half of the sharing was between faculty in different schools (Parham & Doty, 2012). While most researchers seem to want help with DMPs, there are

[3] http://www.nsf.gov/publications/pub_summ.jsp?ods_key=gpg
[4] http://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp#1

mixed findings about what kind of help they prefer and where they obtain it from (O'Donnell & Bowen, 2014; Parham, Bodnar, & Fuchs, 2012; Steinhart, Chen, Arguillas, Dietrich & Kramer, 2012).

Previous studies of NSF DMPs have been conducted via surveys of researchers who have submitted NSF proposals. Curty, Kim, & Qin (2013) attempted to obtain copies of NSF DMPs from their survey respondents but found that the majority of PIs who took part in their study were unwilling or unable to share their DMPs with the authors. The study obtained 69 DMPs -less than half - of the 169 survey respondents. (Curty, Kim, & Qin, 2013).

Detailed knowledge of data types and formats generated by NSF grant-funded research is still largely unknown. Surveys of researchers have revealed that the majority of what they produce and plan to share is in text format (Parham et al., 2012; Steinhart et al., 2012) but other common formats include databases, spreadsheets, code, and images. Steinhart et al. (2012) noted that nearly half of their survey respondents indicated that they were generating three or more different data types.

Campus Outreach and Education

The University of Illinois at Urbana-Champaign is a Carnegie Research University with many top-ranked science and engineering programs. The Grainger Engineering Library Information Center promoted DMP overview materials highlighting requirements from the NSF Engineering Directorate. Library staff also created a NSF grant Data Management Plan template designed to help prepare DMPs. Similar overview materials and templates were developed for the chemistry, life sciences, geology, and physics libraries based on their respective NSF Directorate DMP requirements. The DMP templates were presented by Grainger librarians to College of Engineering departmental and IT staff, business managers, department heads, research officers, and key research faculty. In the course of these discussions, it became clear that various individuals, in addition to the investigators themselves, were partially responsible for preparing the DMP sections. The templates were made available via the various departmental library web sites and in College of Engineering distributions.

Methodology

The University of Illinois at Urbana-Champaign Library conducted an analysis of the DMPs submitted with NSF proposals between July 2011 and November 2013. A total of 1,726 NSF grant proposals from Illinois were submitted through the Fastlane system[5] during this period. Grant proposals were excluded from analysis if they were supplements (n=64), updates (n=76), preliminary proposals (n=69), or planning grant proposals (n=9) as these proposal types do not require a DMP. Additionally, collaborative, multi-institutional submitted grant proposals only require one DMP regardless of how many institutions are submitting. For this reason, all collaborative grants from Illinois without DMPs (n=215) were also limited. There were 33 proposals that could not be examined for various reasons including uploading the wrong document, missing DMPs, withdrawn grants, etc. The remaining 1,260 proposals were considered "valid" for the purpose of evaluating their data management plans.

A team made up of staff reviewed each submitted grant DMP in order to derive and construct the controlled vocabulary terms that would be assigned. The project team identified 11 data storage classifications as assigned terms. The majority of the categories addressed questions of where the data will be stored, on what type of machine/media, where the machine/media is located, and how the data can be found and accessed. The remaining categories classify non-digital

[5] https://www.fastlane.nsf.gov/fastlane.jsp

Table 1: Categories assigned to Data Management Plans

Category	Definition	Examples
PI Server	Computers, servers, hard drives and workstations that the PIs (and/or their staff) use to store project data.	*Laboratory server, external hard drive, group computer, flash drives, etc.*
PI Website	Websites usually edited or ran by the PI or a group they belong to	*Lab website, project website, wiki, PI's website, online databases, etc.*
Campus	Services located on or provided by the Illinois campus	*Illinois Institutional repository, academic computing center Netfiles, National center for Supercomputing Center (NCSA), campus institutes, etc.*
Department	When a department was specifically mentioned as providing a storage or hosting resource	*Departmental website and/or server, dept. backup service or a web address traced to an academic department.*
Remote	Services and sites not located on the Illinois campus.	*Governmental repository services, non-Illinois affiliated museums or institutes, etc.*
Disciplinary	Disciplinary data repositories	*GenBank, arXiv, ICPSR, Dryad, Protein Data Bank, etc.*
Cloud	Storage services using cloud technology	*Google Documents, Google Code, Box.com, DropBox, Amazon Cloud, etc.*
Publication	Traditional scholarly outputs	*Journal articles, workshops, conference presentations, poster sessions, etc.*
Analog	Physical records not including specimens, samples, or artifacts	*Lab notebooks, photographs, and paper files.*
Specimen	Physical specimens, samples or artifacts	*DNA samples, pottery fragments, fecal samples, etc.*
Optical Disc	DVD, CD, and Blu-ray discs	*DVD, CD-ROM, Blu-ray "backups."*
Template	Used the template written by physical science and engineering librarians	*[specific phrases were used to detect usage]*
Not Specified	The text of the DMP was not specific enough to record many details	*"a website", "a server", etc.*
Collaborative	Grants from more than one institution	*[indicated in proposal cover sheets & titles]*
No Data	The research will produce no data products.	*Theoretical studies (such as mathematics), travel grants, educational materials, workshop planning sessions, etc. which do not generate data.*
Funded	If the grant proposal was approved for funding by NSF	*Status change in Fastlane*

or specialized data types and formats and address DMPs which did not provide enough details for further analysis. Some of the categories overlap; this was intentional and unavoidable given the general language of most of the documents examined. Two of the hardest categories to determine were *PI Server* and *PI Website*. A distinction between the two was made being that *PI server* covers storage mediums and machines such as external hard drives, local servers, and personal computers while *PI Website* was applied to DMPs that mentioned storing/ sharing data on websites that they, their lab, or department manages. We believe this distinction is important for assessing data

Table 2: Summary of DMP Category Results

All Examined DMPs n = 1260		
Category	**Number**	**Percent**
PI Server	503	39.9%
PI Website	529	41.9%
Campus	667	52.9%
Department	142	11.2%
Remote	353	28.0%
Disciplinary	275	21.8%
Cloud	63	5.0%
Publication	556	44.1%
Analog	131	10.4%
Specimens	111	8.8%
Optical Disc	56	4.0%
Template used	250	19.8%
Not Specified	66	5.2%
No Data	103	8.2%

sharing: offline data is effectively not being "shared," only stored. While there is overlap in these terms, the team felt that each category was sufficiently independent to warrant its own separate category (Table 1).

These controlled vocabulary or category terms were assigned to each DMP and stored in the supporting relational database for the project. Multiple controlled vocabulary terms could be assigned to a specific DMP. The minimum number of categories assigned was 1 and the maximum was 9. For the 1,260 DMPs reviewed in this project there were an average of 3.2 vocabulary terms per DMP.

Library and Information Science graduate assistants working at the Grainger Engineering Library Information Center read through the DMPs of each submitted NSF proposal and assigned appropriate categories based on the textual content. The category terms were selected and stored in the supporting relational database which also contained a free text field with terms taken from the DMP. The graduate assistants also noted if they recognized wording from the Library developed templates in the submitted DMPs. While automation of the DMP content analysis was possible the wide variety of terminology, vague phrases, and frequent use of discipline and local acronyms would have significantly narrowed and skewed the results. Table 2 summarizes the frequencies of the assigned DMP controlled vocabulary terms.

Assigned Keyword Terms

The assigned keyword terms taken from the DMPs served to complement the controlled descriptors. For example, about one-third (31.3%) of all DMPs indicated that data would be deposited into Remote or Disciplinary repositories (there was some overlap in the usage of these two terms). A variety of specific repositories were mentioned in the DMPs and entered into the keywords field (Table 3).

Interestingly, neither FigShare nor DataCite were mentioned in any of the reviewed pro-

Table 3: DMP Named Repositories

arXiv	61
GenBank	55
National Center for Supercomputing Applications (NCSA) XSEDE	55
NanoHub	34
Dryad	22
National Center for Biotechnology Information (NCBI)	21
Inter-university Consortium for Political and Social Research (ICPSR)	17
Sustainable Environment Actionable Data (SEAD)	15
Box.com	12
GitHub	11
Dropbox	9
DataOne	6

Table 4: NSF proposals by College or Research Facility

Departments with ≥ 25 Proposals		**College/Unit Total**
College of Engineering		**615**
Aerospace Engineering	28	
Civil & Environmental Engineering	118	
Computer Science	106	
Electrical & Computer Engineering	61	
Industrial & Enterprise Systems Engineering	47	
Materials Science & Engineering	30	
Mechanical Science & Engineering	131	
Physics	39	
College of Liberal Arts & Sciences		**458**
Anthropology	38	
Astronomy	25	
Chemistry	43	
Mathematics	111	
College of Agriculture (ACES)		**67**
Crop Sciences	30	
State Water, Geological & Natural History Surveys (Prairie Research Institute)		**27**

posal DMPs. Also, some disciplines do not have established disciplinary repositories. As our study continues, we anticipate seeing references to additional disciplinary repositories and associated services, including Figshare and DataCite being utilized in DMPs.

The project team also looked for any indication of file type or dataset format when reviewing the DMPs. Only 87 proposals actually mentioned any file type or format and these were entered into the keywords field.

Discipline Analysis

While nearly every academic unit from the University of Illinois campus, as well as some administrative units, research institutes, and the Illinois Natural History Survey, are represented in the study the majority of the DMPs examined were authored by PIs from the College of Engineering (49%) and science-related departments in the College of Liberal Arts and Sciences (36%). Only 135 (11%) of the grant proposals were unaffiliated with the College of Engineering or the College of Liberal Arts and Sciences. Because of this, our results primarily represent engineering, computational, and physical or life science research proposals. Table 4 shows units with 25 or more proposals from the Colleges of Engineering, Liberal Arts and Sciences, and Agriculture and Consumer Environmental Studies (ACES). Not all departments within each college submitted over 25 proposals. The State Surveys (geology, natural history, and water) also submitted over 25 grant proposals. The departments of physics and computer science are in the College of Engineering at Illinois.

Distribution of grant proposals by college or unit including departments which submitted 25 or more proposals during the study period. Multidisciplinary proposals were counted once for each unit.

DMP Category Analysis

Servers and Websites

Nearly 40 percent (503) of the DMPs make reference to local storage mediums, such as a PI server. These ranged from portable hard drives and flash drives to "hard drives," "computers," and "group servers." A small number of proposals, 52 (4%), mention optical discs, but always in conjunction with other data storage options.[6] Cloud storage was mentioned infrequently (5%), and specimens were mentioned in only 111 DMPs (8.8%). The majority of the proposals with specimens were associated with life and medical sciences research but 19 were from the College of Engineering, 11 from Chemistry, and 7 from Anthropology. A small number of the proposals, 66 (5.2%), were not specific enough for us to analyze with our terminology.

A total of 667 proposals (52.9%) mention centralized campus resources as a data storage or preservation sites. There were 276 proposals (21.9%) which included the University of Illinois institutional repository as a data deposit resource. DMPs which mentioned these resources used wording such as "UIUC servers," "departmental storage," or referred to the institutional repository, Campus Information Technologies and Educational Services (CITES) managed storage or services, or websites hosted on the *illinois.edu* domain. Only 142 (11.2%) specifically mentioned departmentally managed resources. A large number of proposals (796 or 63.1%) indicate that data will be made available on web sites. The category "PI website" is defined as websites established or created by PIs to promote or distribute their work and includes web pages both on and off the *illinois.edu* domain.

Template and Repository Usage

A total of 250 (19.8%) proposals used word-

[6] Two proposals only mentioned "Optical Disc" and "Analog" – a combination which seems to be the least fulfilling of NSF's "data sharing" requirement.

Table 5: Indicators for DMP Template and Illinois Institutional Repository by College

College or Academic Unit	Template	Percentage	Illinois IR	Percentage
College of Agriculture (ACES)	9	3.60%	12	4.35%
Applied Health Sciences (AHS)	2	0.80%	2	0.72%
College of Engineering (COE)	190	76%	203	73.55%
Liberal Arts & Sciences (LAS)	45	18%	51	18.48%
Veterinary Medicine	1	0.40%	2	0.72%
Graduate School of Library and Information Science (GSLIS)	2	0.80%	3	1.09%
Other	6	2.40%	9	3.26%
TOTAL	249*		276*	

* 6 proposals are joint ventures between two different units. These proposals were counted in both units.

ing from the Grainger Library DMP template. The template specified the Illinois institutional repository as a storage location and described library-based metadata services. A total of 276 (21.9%) DMPs, including those using the Grainger Library template, specified the institutional repository as a data deposit and sharing resource. The majority of the template users were from the College of Engineering (76%, n=190). Table 5 indicates the number and percentages of DMPs citing the template and specifying the Illinois institutional repository.

Scholarly Publication

During the analysis we encountered a high frequency (44.1%) of DMPs that specifically mentioned traditional scholarly outputs in their data management plan. DMPs indicating journal articles, conferences, meetings, workshops, and posters were assigned to the "publication" category. We decided to treat all traditional scholarly communication outputs as one category. Very few DMPs were explicit as to how these traditional scholarly products would disseminate data or data sharing methodologies. Interestingly, this behavior was not restricted to only one college or department indicating that there is confusion across campus about how exactly data sharing and distribution is to occur and if these venues meet NSF's requirements. It is possible that this confusion stems from misunderstanding the difference between a scholarly paper, which describes the research results, and data generated during the research process. This confusion is likely tied to the current NSF focus on PIs supplying DMPs that indicate how processed data – rather than raw – will be made available for sharing.

Analysis of Funded vs. Unfunded Proposals

In the DMP sample studied, there were 298 grant proposals that had been funded by NSF as of May 2014. The analysis team was interested in identifying any trends or consistencies among grant-awarded DMPs. Our research objective was to discover any statistically valid frequency of DMPs indicating specific repository, cloud, or institutional IT server storage venues in funded proposals than in the unfunded proposals. Using the DMP analysis database with the assigned controlled vocabulary terms, we sought to determine if there were any significant differences in proposed stor-

Table 6: DMP Categories assigned to Funded Proposals

Funded Proposals as of May 2014 n = 298 out of 1,260 DMPs			
Category	**Funded**	**Percent Funded**	**Unfund-ed**
PI Server	102	34%	401
PI Website	125	42%	404
Campus	136	45%	531
Department	23	7%	119
Remote	68	22%	285
Disciplinary	54	18%	221
Cloud	11	3%	52
Publication	110	37%	446
Analog	15	5%	116
Specimens	16	5%	95
Optical Disc	10	3%	46
Template	50	16%	200
Not Specified	14	4%	52
No Data	45	15%	58

Table 7: Funded vs. Unfunded Grants by Proposed Storage

Type of Proposed Storage Mechanism	**Funded**	**Unfunded**	**Chi-Square Value**
PI Server / Website	183	569	0.7
Illinois Institutional Repository	62	197	0.02
Campus Storage Services	139	474	0.74
Departmental Server	24	102	1.67
Disciplinary / Cloud Service	67	241	0.85

Table 8: Longitudinal Repository Value

Proposed Data Storage Type	Before October 2012 n=622	After October 2012 n=638	Chi-Square Value
Institutional Repository	108	166	4.59
Disciplinary / Cloud Services	121	182	4.33

age venues or mechanisms between the unfunded proposals and the funded proposals (Table 6).

Of the 136 proposals indicating a campus storage solution, 59 (or 43%) indicated that processed datasets would be deposited in the institutional repository. A significant percentage (76%) indicated storage would be on a PI server or website.

To test whether there are any significant differences between the DMP characteristics of the funded vs. unfunded proposals, we employed the chi-square non-parametric test of significance. Using a 2 x 2 table to compare the frequencies of the assigned DMP categories, this test allows us to determine any significant differences between the funded and unfunded proposals. For the .05 significance level, the critical value of the chi-square statistic with one degree of freedom is 3.84.

We tested the frequencies among funded and unfunded proposals of five DMP proposed storage mechanisms: PI server or websites, institutional repository, campus storage, departmental servers, and disciplinary repositories storage. We removed the 45 funded proposals that indicated the grant would produce no data before conducting the test. Table 7 shows the chi-square values for these five types of services.

The obtained chi-square values are not high enough to indicate that actual population differences exist. The results of our analysis showed that there are no significant differences between funded and unfunded proposals with respect to these four proposed storage venues. For the DMPs included in this study, there was no advantage – in terms of being funded – for proposals specifying disciplinary repositories or the institutional repository as venues for data storage and access.

Longitudinal Analysis

Our final analysis was to conduct a longitudinal study of the proposal DMPs examined in this project. The initial group of proposals were submitted between July 2011 and November 2013, or 28 months. Therefore, the team began investigating differences between the earlier proposals prepared before October 1, 2012 (first 15-month period) and after October 1, 2012 (second 15-month period). This division also divided the proposals into two fairly equal sets of around 610 proposals each. We then calculated the chi-square values for two categories: the proposed use of the institutional repository and disciplinary repository services. Table 8 shows the results. In both of these cases the chi-square values are statistically significant and indicate that more recent proposals are specifying the use of the University's institutional repository and disciplinary repositories at a higher frequency. This may in fact indicate that the Library's DMP assistance and education efforts are having some effect and that PIs are becoming more aware of data preservation and available institutional and/or disciplinary repositories.

Discussion and Conclusion

This DMP analysis looked at 1,260 NSF pro-

posals from the University of Illinois at Urbana-Champaign dating from July 2011 to November 2013. Permission for this study and release of data was given by the University of Illinois campus Office of Sponsored Programs and Research Administration (OSPRA), the Office of the Vice-Chancellor for Research and the University of Illinois Institutional Review Board. Each DMP was assigned controlled vocabulary and keyword terms that reflected the grant's proposed data storage venues and the proposed mechanisms for sharing and reuse of data. One of the primary goals of this study project was to better understand what researchers were proposing in their data management plans. The project team identified a fairly comprehensive set of descriptors that can be used as a framework for describing grant proposal DMPs.

This project clearly identified the data repository and storage solutions that researchers proposed to meet the NSF DMP requirement. Knowledge of where researchers intend to archive data is important to institutions concerned with funding agency-compliance requirements. In addition, libraries need to assess trends in storage technologies and aid in developing new solutions for researchers with data management needs.

The large number of DMPs which mentioned publications, conferences, and workshops as a method of data dissemination and sharing was surprising. This, we believe, is partially due to the vagueness of the NSF DMP guidelines, but is also a side effect of the NSF's focus on the sharing of processed data – as opposed to raw data – and the PI's natural tendency to associate processed data with publications. Perhaps we should have expected this attempt to fit *de facto* practices of scholarly communication into the data management plan. Clearly, there is work to be done on the education of NSF grant writers on data, data stewardship needs, and local and disciplinary data management opportunities.

This project also looked at whether there were data storage venues that were proposed statistically more frequently in the funded proposals than in the unfunded proposals. In particular, we examined whether proposals that indicated the use of an institutional repository, a disciplinary repository, or a PI or departmental storage site were more likely to be funded. We found that there were no statistically significant differences between the specific storage venues or reuse mechanisms proposed within funded and unfunded proposals. This indicates that researchers are in the beginnings of the DMP lifecycle and that the expected communities of practice and best practices have yet to emerge.

We also looked for statistically significant differences in the DMPs of the early proposals compared to more recent proposals. We learned that in later proposals (since October 2012) there was a statistically significant higher use of the Illinois institutional repository and disciplinary or cloud storage solutions. It would appear that researchers are responding to the library's educational and assistance efforts with respect to storage mechanisms.

This study found that, Illinois' disciplinary and institutional repository resources for data storage and processed data deposit are being underutilized in NSF proposals. At the same time, we know that campus IRs are typically best suited for discrete, static, and processed datasets, not large datasets or dynamic, active datasets. Data management is an institutional-wide issue requiring collaborative working relationships between multiple stakeholders. It is critical that campuses and other institutions awarded NSF grants either develop or access key infrastructure services that will give researchers enhanced data management capabilities and provide mechanisms for compliance with federal grant requirements and mandates.

References

Borgman, Christine L. "The conundrum of sharing research data." *Journal of the American Society for Information Science and Technology*, 63, no. 6 (2012): 1059-1078, http://dx.doi.org/10.1002/asi.22634

Cragin, Melissa H., Carole L. Palmer, Jacob R. Carlson, and Michael Witt. "Data sharing, small science and institutional repositories." *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 368, no. 1926 (2010): 4023-4038, http://dx.doi.org/10.1098/rsta.2010.0165

Curty, Renata, Youngseek Kim, and Jian Qin. 2013. "What have Scientists Planned for Data Sharing and Reuse? A Content Analysis of NSF Awardees' Data Management Plans." In *Research Data Access & Preservation Summit*. Baltimore. http://surface.syr.edu/ischoolstudents/2/

Dietrich, Dianne, Trisha Adamus, Alison Miner, and Gail Steinhart. "De-Mystifying the Data Management Requirements of Research Funders." *Issues in Science and Technology Librarianship*, Summer 2012:1-16, http://dx.doi.org/10.5062/F44M92G2

Mischo, William H., and Mary C. Schlembach. "Open Access Issues and Engineering Faculty Attitudes and Practices." *Journal of Library Administration*, 51, no. 5-6 (2011): 432-454, http://dx.doi.org/10.1080/01930826.2011.589349

National Science Foundation. (2013). *Report to the National Science Board on the National Science Foundation's Merit Review Process Fiscal Year 2012* (p. 71). Retrieved from http://www.nsf.gov/nsb/publications/2013/nsb1333.pdf

O'Donnell, Megan N., and Bonnie S. Bowen (2014). *Data Management Brownbag Evaluation Summary* (p. 5). Ames. Retrieved from http://lib.dr.iastate.edu/libreports/12/

Parham, Susan W., Jon Bodnar, and Sara Fuchs. "Supporting tomorrow's research Assessing faculty data curation needs at Georgia Tech." *College & Research Libraries News*, 78, no. 1 (2012): 10-13.

Parham, Susan W., and Chris Doty. "NSF DMP Content Analysis: What Are Researchers Saying?" *Bulletin of the American Society for Information Science and Technology*, 39, no. 1 (2012): 37-38, http://dx.doi.org/10.1002/bult.2012.1720390113

Schlembach, Mary C., and Carol A. Brach, C. A. "Research Data Management and the Role of Libraries." *Special Issues in Data Management, ACS Symposium Series*, Vol. 1110 (2012): 129-144, http://dx.doi.org/10.1021/bk-2012-1110.ch008

Steinhart, Gail, Eric Chen, Florio Arguillas, Dianne Dietrich, and Stefan Kramer. "Prepared to Plan? A Snapshot of Researcher Readiness to Address Data Management Planning Requirements." *Journal of eScience Librarianship*, 1, no. 2 (2012):63-78, http://dx.doi.org/10.7191/jeslib.2012.1008

Disclosure: The authors report no conflicts of interest.

ISSN 2161-3974