1-1-2017

# Quality Evaluation of Data Management Plans at a Research University

James E. Van Loon
*Wayne State University*, jevanloon@wayne.edu

Katherine G. Akers
*Wayne State University*

Cole Hudson
*Wayne State University*

Alexandra Sarkozy
*Wayne State University*

*Article*

# Quality evaluation of data management plans at a research university

**Authors:**
**James E. Van Loon**
Wayne State University, USA
**Katherine G. Akers**
Wayne State University, USA
**Cole Hudson**
Wayne State University, USA
**Alexandra Sarkozy**
Wayne State University, USA

**Abstract**
With the emergence of the National Science Foundation (NSF) requirement for data management plans (DMPs), academic librarians have increasingly aided researchers in developing DMPs and disseminating research data. To determine the overall quality of DMPs at Wayne State University, the Library System's Research Data Services (RDS) team evaluated the content of 119 DMPs from NSF grant proposals submitted between 2012 and 2014. The results of our content analysis indicate that, while most researchers understand the need to share data, many DMPs fail to adequately describe the data generated by the project, how data will be managed during the project, or how data will be preserved and shared after the completion of the project. Our results also show that DMP deficiencies vary across academic units, suggesting the need for differentiated outreach services to improve the strength of DMPs in future NSF grant proposals.

**Keywords**
Data management, data sharing, research data, National Science Foundation, evaluation, quality

## Introduction

Researchers are increasingly asked to provide access to their research data. Two key pieces of policy have set the tone concerning research data sharing in recent years: the National Science Foundation (NSF) 2011 requirement for the inclusion of data management plans (DMPs) in all grant proposals, and the 2013 memo by the Office of Science and Technology Policy requiring all major federal funding agencies to facilitate access to the publications and data resulting from federally funded research. As such, other federal funding agencies, including the National

**Corresponding author:**
James E. Van Loon, University Libraries, Wayne State University, 5265 Cass Avenue, Detroit, MI 48202
Email: jevanloon@wayne.edu

Institutes of Health, the Department of Energy, and the National Aeronautics and Space Administration now require or will soon require DMPs.

According to NSF guidelines, a DMP is a supplementary document of no more than two pages that describes how the proposal will conform to the funding agency's policy on providing access to research data (National Science Foundation, 2014). The DMP is reviewed as part of the intellectual merit or broader impact of each NSF proposal. Although the content requirements for DMPs vary slightly across different NSF directorates, DMP elements expected for all directorates include data types and formats, methods of data sharing, and policies for data reuse and redistribution.

Wayne State University is a "doctoral university: highest research activity" according to the Carnegie Classification of Institutions of Higher Education (Indiana University Center for Postsecondary Research, 2015). Wayne State University has an annual research expenditure of over $245 million (University Research Corridor, 2016) and received almost $12 million in NSF research grants in fiscal year 2015 (National Science Foundation, 2016). In 2013, a team of librarians and specialists established Research Data Services (RDS) to provide outreach, consultation, and training on research data management and sharing to Wayne State University faculty and research support staff. To further understand faculty research data management practices and to direct the future efforts of the RDS team, we analyzed the content of DMPs submitted by Wayne State researchers, focusing solely on NSF proposals due to the volume of NSF funding at our institution and the relative maturity of DMP requirements for this agency. The objectives of our study were to (1) evaluate their overall quality and adherence to NSF guidance, and (2) determine whether academic units differ in their adherence to NSF guidelines.

## Literature review: content analysis of DMPs

The overall quality of NSF DMPs has been evaluated in previous studies. Curty and colleagues (2013) used an online survey to assess attitudes and practices around data management planning among 966 NSF awardees from across the country and then analyzed the content of 68 DMPs volunteered by a subset of these researchers. They found several weaknesses in the DMPs, including dependence on informal or personal methods of sharing data (e.g., emailing upon request) and failure to address metadata standards and policies for data reuse/redistribution. As part of a pilot project to provide data management services to NSF applicants at the University of Michigan, Nicholls and colleagues (2014) acquired 104 DMPs from successful proposals from engineering faculty and analyzed how well the DMPs conformed to NSF guidance. They concluded that although most DMPs were of acceptable quality, many lacked required elements, such as identification of the individuals responsible for data management and specification of the period of data retention. Bishoff and Johnston (2015) analyzed the content of 182 DMPs solicited from researchers at the University of Minnesota and found significant variation across DMPs in data sharing methods, the intended audience for sharing, and data preservation strategies.

Other studies have focused on evaluating NSF DMPs to specifically assess researchers' methods of data preservation and sharing, including the use of an institutional repository (IR) to provide access to data. Parham and Doty (2012) reviewed the content of 181 DMPs at the Georgia Institute of Technology, focusing on whether researchers indicated that they would use

the IR to share their research data. They often found outdated or inaccurate references to the IR, presumably due to researchers' practice of sharing "boilerplate" DMP language across academic departments, suggesting the need to develop consistent language about repository services for research data and to target IR awareness efforts to specific departments. Also, Mischo and colleagues (2014) examined 1,260 DMPs at the University of Illinois and found no significant association between data storage methods and proposal funding success, although they discovered an increasing reliance on their IR as a venue for research data preservation over time.

Recently, the Data Management Plans as a Research Tool (DART) project, led by Rolando and collaborators (2015), developed and tested an evaluation rubric for NSF DMPs to create a robust and standardized assessment tool for DMPs to enable cross-institutional comparisons. An early version of the DART rubric was used by Samuel and colleagues (2015) to assess 29 DMPs from engineering faculty at the University of Michigan. They found that the overall quality of DMPs varied greatly and identified elements that were often missing from DMPs, including clear roles and responsibilities for data management, metadata standards for describing research data, and policies for protecting intellectual property rights.

## Motivation for the present study

Although other researchers have evaluated the overall quality or specific elements of DMPs, we evaluated the quality of NSF DMPs at Wayne State University to (1) characterize the content of DMPs created by researchers at our institution and (2) identify significant variations in DMP content between academic units. Another potential outcome of this study was knowledge of specific and chronic deficiencies in DMPs that might help our team in developing tailored outreach and education for WSU faculty, administrators, research support staff, and other librarians.

## Methodology

We approached Wayne State University's Sponsored Program Administration (SPA) office with a proposal to study NSF DMP quality in 2014. SPA was receptive to our proposal and provided read-only access to the pre-award administrative system and support for compiling the DMP sample. Our study fell within the scope of program evaluation/quality improvement activities as defined by Wayne State's Institutional Review Board (IRB) and thus did not require IRB approval.

We compiled all funded NSF proposals between 2012 and 2014 and a roughly equal number of unfunded NSF proposals. After omitting proposals containing no DMP or for which the DMP content was minimal (e.g., conference or travel proposals), our final sample consisted of 119 DMPs from five WSU academic units as summarized in Table 1. To maintain confidentiality of NSF proposal content, the DMPs were secured on a password-protected, internal library server for the duration of the study.

**Table 1.** Final sample of DMPs.

| Academic unit | Number of DMPs |
|---|---|
| College of Education[a] | 2 |
| College of Engineering[b] | 61 |
| College of Liberal Arts and Sciences[c] | 50 |
| Law School | 1 |
| School of Medicine[d] | 5 |
| Total | 119 |

[a] Departments: Teacher Education (1), Theoretical & Behavioral Foundations (1)

[b] Departments: Biomedical Engineering (2), Chemical Engineering (14), Civil Engineering (6), Computer Science (13), Electrical & Computer Engineering (11), Engineering Technology (3), Industrial & Systems Engineering (5), Mechanical Engineering (7)

[c] Departments: Biological Sciences (4), Chemistry (22), Geology (5), Mathematics (8), Physics (11)

[d] Departments: Anatomy (1), Pediatrics (1), Pharmacology (1), Physiology (2)

The DMPs were evaluated using a modified version of a rubric previously used by researchers at the University of Michigan (Nicholls et al., 2014; Samuel et al., 2015). Our rubric (Appendix 1) consisted of 15 items addressing the inclusion of information requested by the NSF (National Science Foundation, 2014) and other common pieces of information often found in DMPs. Two evaluators independently applied the rubric to each DMP, and any inconsistencies between evaluators were discussed and ultimately reconciled. Descriptive statistics for each rubric item were calculated for the full sample and separately for two major subgroups: the College of Engineering and the College of Liberal Arts and Sciences. Furthermore, we examined statistically significant differences in DMPs between the College of Engineering and the College of Liberal Arts and Sciences using Chi-square tests, with statistical significance set at $p < 0.05$.

## Results

### *Overall quality of DMPs*

Table 2 summarizes the proportion of DMPs containing each recommended element for the full sample and the two major subgroups: the College of Engineering and the College of Liberal Arts and Sciences. For the full sample, nearly half of the DMPs (49%) specified the individual(s) responsible for data management/sharing. A minority of DMPs (8%) specified the total amount of data that would be generated or the rate of data generation. Most DMPs (81%) characterized data in terms of either its type (e.g., mass spectrometry data, scanning electron microscope

images) or format (e.g., file extensions, name(s) of software used to collect the data). Less than half of the DMPs (38%) mentioned specific metadata standards or methods of data description (e.g., codebook, readme file). More than half of the DMPs (60%) discussed data back-up during the active project period. A vast majority of DMPs (92%) expressed an intention to share at least some data after completion of the project, but less than half (43%) specified the duration of data preservation.

We further addressed the specific methods by which researchers intended to share their data. For the full sample, the most frequently specified method of data sharing was posting data on personal websites/databases (51%; Table 2). The second most common methods were providing data upon request (e.g., by email; 24%) or depositing data in a dedicated data repository (24%). 13% of DMPs mentioned sharing data through supplemental materials submitted alongside journal articles. Interestingly, a substantial proportion of DMPs (20%) stated that research data would be shared via journal articles (not as supplemental material) or conference presentations, indicating that some researchers do not distinguish between their results (i.e., summary data in tables and graphs) and the underlying data that support their results (i.e., individual-level or "raw" data in various file formats).

We also evaluated DMP content related to policies for data sharing. For the full sample, less than half of the DMPs (42%) mentioned policies for intellectual property, and only about one in five DMPs included statements about policies for data reuse/redistribution or protecting sensitive information.

**Table 2.** Elements contained in DMPs

| DMP element | Full sample | College of Engineering | College of Liberal Arts and Sciences |
|---|---|---|---|
| Basic elements | | | |
| 1. Responsible individual | 49% | 44% | 50% |
| 2. Amount of data | 8% | 11% | 6% |
| 3. Expected types/formats | 81% | 87% | 72% |
| 4. Description/metadata | 38% | 36% | 36% |
| 5. Data backup | 60% | 59% | 60% |
| 6. Intention to share data | 92% | 93% | 90% |
| 7. Duration of data preservation | 43% | 56% | 26% |
| Method of data sharing | | | |
| 8. Email on request | 24% | 18% | 30% |

**Table 2.** Elements contained in DMPs

| DMP element | Full sample | College of Engineering | College of Liberal Arts and Sciences |
|---|---|---|---|
| 9. Personal website or database | 51% | 57% | 50% |
| 10. Journal articles or conferences | 20% | 28% | 14% |
| 11. Supplemental material | 13% | 2% | 28% |
| 12. Data repository | 24% | 18% | 28% |
| Data sharing policies | | | |
| 13. Reuse or redistribution | 19% | 30% | 8% |
| 14. Sensitive information | 20% | 26% | 6% |
| 15. Intellectual property | 42% | 64% | 18% |

Table 3 shows a breakdown of how researchers characterized the data generated by their project. For the full sample, less than half of the DMPs (42%) included both general (i.e., data types, such as mass spectrometry data or scanning electron microscope images) and specific (i.e., data format, such as file extensions or the name(s) of software used to collect the data) descriptions of the expected data, and smaller proportions of DMPs included either general or specific descriptions of data (32% and 7%, respectively) but not both. A substantial proportion of DMPs (19%) completely lacked a description of the data to be generated.

**Table 3.** Characterization of data types/formats in DMPs

| Characterization of expected data types/formats | Full sample | College of Engineering | College of Liberal Arts and Sciences |
|---|---|---|---|
| Absent or unclear | 19% | 13% | 28% |
| General (i.e., type) | 32% | 29% | 34% |
| Specific (i.e., format) | 7% | 7% | 8% |
| Both general and specific | 42% | 51% | 30% |

*Differences in DMP content between engineering and liberal arts and sciences*

The full sample of DMPs (n = 119) contained two major subgroups: the College of Engineering (n = 61) and the College of Liberal Arts and Sciences (n = 50). Therefore, we analyzed differences in DMP content between these two academic units. Whereas 28% of DMPs from liberal arts and sciences expressed the intention to share data via journal supplemental materials, only 2% of DMPs from engineering expressed this intention ($\chi^2$ (1, n = 111) = 16.3, p < 0.001; Table 2). Slightly over half of engineering DMPs (56%) specified the duration of data preservation, but only 26% of liberal arts and sciences DMPs contained this element ($\chi^2$ (1, n = 111) = 9.9, p = 0.002). Furthermore, DMPs from engineering were more likely to describe policies for data reuse or redistribution (30%; $\chi^2$ (1, n = 111) = 7.9, p = 0.005) and safeguarding sensitive information (26%; $\chi^2$ (1, n = 111) = 7.9, p = 0.005) or intellectual property (64%; $\chi^2$ (1, n = 111) = 23.6, p < 0.001) than DMPs from liberal arts and sciences (reuse/redistribution: 8%, sensitive information: 6%, intellectual property: 18%). No other differences in DMP elements between the two major subgroups were statistically significant.

## Discussion

We found substantial variation in the quality of individual NSF DMPs from Wayne State University researchers. 92% of DMPs indicated that at least some data would be shared with others after the completion of the projects, which demonstrates that Wayne State researchers largely understand that the NSF expects broad data sharing. However, similar to previous studies (Curty et al., 2013; Nicholls et al., 2014; Bishoff and Johnston, 2015), we found that many DMPs failed to adequately describe the data that would be generated by the project, how data would be managed during the project, or how data would be preserved and shared with others after the completion of the project. In particular, we found that 51% of DMPs did not identify the individual(s) responsible for data management, which may be problematic for proposals involving multiple principal investigators or cross-institutional collaboration or for labs with high turnover rates for graduate students and research staff. Most DMPs (92%) did not provide an estimate of the total amount or expected rate of data generation, which is important for choosing the most appropriate data storage and preservation methods. 57% of DMPs did not specify the duration that data would be preserved after the project or policies governing how other researchers might reuse or redistribute their data, suggesting that researchers often do not carefully think about the lifespan of the data beyond the active period of the project. Furthermore, a majority of DMPs (62%) did not mention specific metadata standards or methods of data description methods, indicating that the data might not be easily discoverable by or understandable to other researchers in the long term.

In terms of data sharing methods, we found that researchers often rely on informal methods of providing access to data, such as sharing data through email upon request (24%) or through personal or project-specific websites or databases (51%). Only 24% of researchers stated that they would deposit data into a dedicated data repository. Informal data sharing methods, particularly sharing via email upon request, have been found to be less reliable for long-term data access than the use of a dedicated repository. Vines et al. (2014) found that the odds of successfully receiving data in response to an email request fell at the rate of 17% per year and that the chances of locating working email addresses for authors also dropped by 7% per year.

Thessen et al. (2016) found that more than one-third of email requests for datasets received no response and that the overall success rate for email requests was 40%. They also found that sharing upon request was inefficient, requiring an average of 7.8 emails between the requester and data holder to negotiate a successful data transfer. Furthermore, Savage and Vickers (2009) found that only 10% of datasets requested by email were successfully received. Therefore, our RDS team will work to make Wayne State University researchers aware of the disadvantages of informal data sharing methods and encourage them to use more reliable and persistent methods of data sharing.

Interestingly, similar to previous findings by Bishoff and Johnston (2015), we discovered that a substantial proportion of DMPs (20%) stated that data would be shared via journal articles or conference presentations. In these cases, it was clear that researchers were not referring to sharing data through supplemental files accompanying journal articles; rather, they considered the publication of journal articles themselves as a way to share data. Although it is certainly expected that the results of research (i.e., interpreted, summary data in graphs and tables) would be shared through journal articles and conference presentations, these are not valid avenues of sharing the actual data underlying those results (i.e., uninterpreted, individual-level data in a variety of file formats). We believe that this may stem from a tendency for researchers to use the terms "data" and "results" interchangeably, which suggests that researchers could benefit from greater awareness of the NSF and Office of Management and Budget definitions of "research data".

Most NSF proposals in our sample originated from two academic units (the College of Engineering and the College of Liberal Arts and Sciences), allowing us to examine differences in DMP content between engineering and basic science researchers. DMPs from engineering researchers were less likely to mention data sharing through supplemental materials accompanying journals articles compared with DMPs from liberal arts and sciences faculty. This finding suggests the need to improve awareness among engineering researchers of the possibility of sharing research data via this method, which we will incorporate into future outreach efforts. Also, DMPs from liberal arts and sciences researchers were less likely to specify the duration of data preservation and to describe policies for reuse/redistribution and protecting sensitive information and intellectual property rights compared with DMPs from engineering faculty. These findings indicate a need to inform liberal arts and sciences faculty about the importance of thinking about the lifespan of their data beyond the period of the project and considering whether steps should be taken to safeguard aspects of their data while also allowing the broadest access possible.

## Conclusion

By employing content analysis, we have characterized the level of quality and the variation between different academic units in NSF DMPs written by Wayne State researchers. We find that many DMPs provide an incomplete or ambiguous description of how research data will be managed and shared with others, suggesting that there is substantial room for improvement in DMP quality at our institution. Furthermore, we found several differences in DMP content between proposals from engineering versus liberal arts and sciences. These results indicate a need for the library to provide greater outreach, education, and consultation on developing strong

DMPs and best practices in research data management and dissemination, and suggest that these efforts should be tailored to the needs and practices of particular groups of researchers.

Finally, we note that performing a DMP quality evaluation at our university has been a valuable experience for our RDS team, providing an opportunity to increase our knowledge of the grant application and data management planning process, to foster relationships between our team and university administrators and other research support staff, and to create a DMP-related workshop for other librarians.

**Acknowledgement**

The authors wish to thank Gail Ryan and Tim Foley of the Wayne State University Office of Sponsored Programs Administration for their support and guidance in this work.

**Declaration of conflicting interest**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## References

Bishoff C and 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. DOI: 10.7710/2162-3309.1231.

Curty R, Kim Y and Qin J (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 2013*, Baltimore, MD. Available at: http://surface.syr.edu/ischoolstudents/2 (accessed 2 May 2016).

Indiana University Center for Postsecondary Research (2015) *Carnegie Classification of Institutions of Higher Education, 2015 Edition*. Available at: http://carnegieclassifications.iu.edu/ (accessed 2 May 2016).

Mischo WH, Schlembach MC and O'Donnell MN (2014) An Analysis of Data Management Plans in University of Illinois National Science Foundation Grant Proposals. *Journal of eScience Librarianship* 3(1): 31-43. DOI: 10.7191/jeslib.2014.1060.

National Science Foundation (2014) *Chapter II - Proposal Preparation Instructions*. Available at: http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_2.jsp#dmp (accessed 2 May 2016).

National Science Foundation (2016) *Award Summary: by Top Institutions*. Available at: http://dellweb.bfa.nsf.gov/Top50Inst2/default.asp (accessed 2 May 2016).

Nicholls NH, Samuel SM, Lalwani LN, et al. (2014) Resources to Support Faculty Writing Data Management Plans: Lessons Learned from an Engineering Pilot. *International Journal of Digital Curation* 9(1): 242-252. DOI: 10.2218/ijdc.v9i1.315.

Parham SW and 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. DOI: 10.1002/bult.2012.1720390113.

Rolando L, Carlson J, Hswe P, et al. (2015) Data Management Plans as a Research Tool. *Bulletin of the American Society for Information Science and Technology* 41(5): 43-45. DOI: 10.1002/bult.2015.1720410510.

Samuel SM, Grochowski PF, Lalwani LN, et al. (2015) Analyzing Data Management Plans: Where Librarians Can Make a Difference. In: *122nd ASEE Annual Conference & Exposition*, Seattle, WA. Available at: https://www.asee.org/public/conferences/56/papers/12072/view (accessed 2 May 2016).

Savage CJ and Vickers AJ (2009) Empirical study of data sharing by authors publishing in PLoS journals. *PloS one* 4(9): e7078. DOI: 10.1371/journal.pone.0007078.

Thessen AE, McGinnis S and North EW. (2016) Lessons learned while building the Deepwater Horizon Database: Toward improved data sharing in coastal science. *Computers and Geosciences* 87: 84-90. DOI: 10.1016/j.cageo.2015.12.001.

University Research Corridor (2016) *About the University Research Corridor*. Available at: http://urcmich.org/about/ (accessed 2 May 2016).

Vines TH, Albert AYK, Andrew RL, et al. (2014) The availability of research data declines rapidly with article age. *Current Biology* 24(1): 94-97. DOI: 10.1016/j.cub.2013.11.014.

**Appendix 1. Wayne State University DMP evaluation rubric.**

*Basic DMP elements*

1. Are the individual(s) responsible for data management specifically named (or referred to as "the PI")?
   1 = yes
   0 = no/not clear
2. Is the total amount of expected data and/or expected rate of data generation specified?
   1 = yes
   0 = no/not clear
3. Are the file formats of expected data specified (e.g., file extensions, name of data collection software)?
   0 = no/not clear
   1 = general description (e.g., mass spectrometry data)
   2 = specific description (e.g., file extensions, software used)
   3 = both general and specific description
4. Will specific metadata standards and/or other description methods (e.g., readme files, codebooks, and lab notebooks) be used?
   1 = yes
   0 = no/not clear
5. Is a method of data backup (e.g., RAID, remote backup, external hard drive) specified?
   1 = yes
   0 = no/not clear
6. Will any data and/or code be made accessible after the study?
   1 = yes
   0 = no/not clear
7. Is the duration of data/code preservation specified?
   1 = yes
   0 = no/not clear

*Method of data sharing*

8. Will data/code be provided (e.g., emailed) upon request?
   1 = yes
   0 = no/not clear
9. Will data/code be posted on personal or project-specific website or database?
   1 = yes
   0 = no/not clear
10. Will data be shared via journal articles or conference presentations?
    1 = yes

0 = no/not clear

11. Will data/code be submitted to journals as supplemental material?
    1 = yes
    0 = no/not clear
12. Will data be deposited in a dedicated data repository/archive?
    1 = yes
    0 = no/not clear

*Data sharing policies*

13. Are policies for data re-use or redistribution specified?
    1 = yes
    0 = no/not clear
14. Do policies for data access and sharing specify protections against disclosure of sensitive information?
    1 = yes
    0 = no/not clear
15. Do policies for data access and sharing specify protections for safeguarding intellectual property rights?
    1 = yes
    0 = no/not clear