2017-02-24

An Analysis of Data Management Plans from the University of Michigan

Carlson, Jake

An Analysis of Data Management Plans from the University of Michigan

Jake Carlson
Research Data Services Manager
University of Michigan Library

Feb 24, 2017

Table of Contents

Summary

This report contains the findings of a content analysis of 100 data management plans (DMPs) from grants submitted to the National Science Foundation (NSF) written by researchers at the University of Michigan (U-M). The intent behind this analysis was to better understand how researchers at the U-M have interpreted and responded to the NSF's requirement to submit a DMP that describes how they will manage, share and archive the data generated over the course of their funded research. The results will be used to identify potential areas in need of support and to consider how the University of Michigan and the library in particular, could respond to these areas through providing services.

General Observations

It was clear from reviewing the 100 DMPs that many researchers at the University of Michigan do not appear to fully understand what they are being asked to do or what information they are supposed to provide in their DMP. The DMP requirement is a relatively new one and researchers may still be adjusting to having to consider the issues it raises. Particularly in communities where data sharing is not yet commonly practiced, researchers may not have a complete understanding of terms, concepts or rationale behind the data management plan requirement.

The DMPs I reviewed varied widely in their quality, amount of detail provided and particular focus. Some DMPs were clearly written with the intent of providing the NSF with the information that it wanted. Others appeared to be written in haste or with the intent to fulfill the letter of the data management requirement, but not the spirit. Then there were others that did not seem to fully connect with what the NSF was asking them to do. For instance, some researchers took the 5 components required by the NSF and used them as headings in their DMP, but then did not actually address the components themselves in these sections.

Overall, researchers tended to focus more on the management aspects of the DMP requirement over the sharing aspects. Multiple researchers focused on describing their data storage architecture and infrastructure in their DMPs. This may not be particularly surprising as researchers have already had to develop a means to administer their data for their own purposes prior to addressing the DMP requirement. Similarly, many of the researchers who will be generating data from human subjects or data that would be considered sensitive appeared to have copied parts of their IRB application and pasted it into their DMP. This too is not all that surprising as the IRB also asks for information about how data will be stored and secured over the course of the research project.

This is not to say that sharing data was not addressed in the DMPs I reviewed. Multiple DMPs focused on how they would prepare their data for sharing and make it available to others in their research community. Some of the researchers that did address sharing in their DMPs appeared to have some prior experience in considering how their data would be shared outside of project personnel. For example, some of these DMPs identified a specific data repository as the means through which they would share their data and included text on how they would prepare their data. Others mentioned

potentially applying for patents or working with the Technology Transfer office and making their data available based on the determination of these offices.

General Recommendations

In making the data management plan requirement more meaningful, researchers will need more support from funding agencies, the university and the library. My list of general recommendations to each is below. More specific recommendations based on each component of the NSF's DMP requirement are included in the later sections of the report.

Recommendations for funding agencies:

The National Science Foundation (NSF) has not provided a particularly clear or detailed description of the responsibilities of researchers in managing, sharing or archiving the data generated from funded projects. Based on the findings of this review I would recommend that any funding agency seeking to institute and support a requirement for researchers to manage, share or archive their data consider taking the following actions:

- In addition to providing a clear explanation of the requirements (including definitions of key terms and concepts), agencies should provide a rationale behind the requirements in their guidance. For example, it may not be clear to the researcher why they would need to include the file formats of their data set in a DMP. Instructions provided by the agency could include a statement on the need to encourage the use of community supported, open formats and to help repositories better plan and prepare to receive the data.

- It was apparent from reading many DMPs that some researchers simply saw this as a checkbox that they had to fill rather than as a useful exercise to inform their work. Funding agencies should encourage researchers to think of and develop the DMP as a living document that can be used to guide the evolution of the data set. Restructuring the data management plan requirement to be more in tune with the researcher's workflow or the lifecycle of the data set(s) being generated could promote the DMP's continued use. Funding agencies should also consider including progress made in working through the DMPs as an explicit component of reporting to the funding agency.

- Funding agencies should consider sharing the information collected from data management plans in aggregate to better inform the repository, publishing, curation and other communities who provide support for data management, sharing and preservation. Having access to the content of DMPs would help these communities better understand the nature and types of data being generated and to better anticipate and respond to researcher needs. If a specific repository is named in the DMP, consider sharing the DMP with the repository.

Recommendations for the University of Michigan:

Based on my review of DMPs produced at the University of Michigan I believe that researchers would benefit from additional and coordinated support in developing and executing DMPs. I recommend the following actions for unit at U-M that provide support relevant to the DMP requirement:

- Consider how support for developing and carrying out a DMP could be more fully incorporated into the lifecycle of a grant and into the systems through which grants are implemented, tracked and concluded. For example, the option to request assistance in developing a DMP could be included as a part of the Proposal Approval Form (PAF) and proposal checklist.

- Investigate how we could strengthen the connections between service units on campus that provide support, resources and guidance to researchers in managing, sharing and archiving their data over the course of the lifecycle of the research. This would include units that oversee the grants process such as the Research Administrators Network (RAN), IT units that provide systems used to store, back-up, secure manage and share data such as Advanced Research Computing (ARC), and the Library which provides services to support sharing and archiving research data.

- Explore the feasibility of sharing DMPs, either in aggregate or individually, with the service agencies on campus (IT units, the library, etc.) named by the researcher to provide support over the course of the grant. This would aid these service agencies in anticipating and responding to the needs of the researcher over the course of the grant.

Recommendations for the University of Michigan Library:

The library provides services to support researchers in managing, sharing and archiving their data (https://www.lib.umich.edu/research-data-services). These services include consultation on developing a DMP that not only addresses the requirements of the funding agency but takes into consideration the local practices and needs of the researcher. The library launched its data repository, Deep Blue Data, in September 2016 as a means for researchers to share and archive their data in accordance with funding agency requirements. I recommend the following actions for the library:

- Explore how the library could further promote our research data services in providing assistance, particularly at key points in the research lifecycle. Can we better attune when we offer workshops on developing a DMP or promote our consultation services to when grant proposals are due? For researchers who have identified Deep Blue Data as the destination for their data, can we identify check in points to see if the researcher has questions or would like to schedule a consultation?

- Research to what extent researchers have “closed the loop” at the end of their grant by submitting their to the data repository they identified or by making their data available in the

means they specified in their DMP. If they have not submitted their data, identify the challenges and barriers they faced that prevented them from following through on their DMP.

Project Background

This review of DMPs is a component of the Data management plan As Research Tool (DART) Project. The DART project was funded by the Institute for Museum and Library Services (IMLS) in 2013 to develop a means for librarians to review data management plans as a means to inform data management services in libraries. Oregon State University is the lead institution for the DART project, which includes librarians from Georgia Institute of Technology, Penn State University, the University of Oregon and the University of Michigan[1].

The DART project is centered on two goals:

1. Developing an analytic rubric to standardize the review of data management plans as a means to inform targeted expansion or development of research data services at research libraries;
2. Producing a study utilizing the rubric as a means to analyze the content of data management plan at five universities.

To accomplish these goals, participating librarians reviewed the data management plan requirement of the National Science Foundation (NSF). Our review included the instructions of how to develop a data management plan (DMP) and guidance in how to do so provided by the NSF, the more specific guidance developed by the individual directorates within the NSF, and supplemental material such as research articles describing researchers experiences in writing DMPs and guidance provided by librarians. We developed an initial rubric for evaluating DMPs through analysis of these documents with an eye towards implications and impacts for library services.

The next stage of the project was to test the rubric which was done in two phases. Both phases required building a collection of data management plans written by researchers to use as a means to test the rubric's effectiveness. To develop this collection, each librarian gathered 100 DMPs from their respective institutions. In the first phase of our test of the rubric, we each selected 5 DMPs from each institution to form an initial collection of 25 DMPs. Each librarian on this project reviewed these 25 DMPs separately using the draft rubric to test out the content and structure of the rubric and to identify areas needing further discussion or development. We then met several times as a group over Skype to review our responses and discuss our experiences using the rubric. These discussions provided us with the opportunity to compare results, to raise questions or ask for clarification for areas that were not clear and to further refine our protocols. We then produced a new draft of the DART rubric based on our discussions. We also used this phase to test inter-rater reliability. We used intra-class correlation (ICC) to

[1] More information about the DART project can be found at: https://osf.io/kh2y6/wiki/home/

assess inter-rater reliability[2,3,4] and ultimately were able to achieve a median ICC score of 0.76, indicating a strong agreement between raters.

Phase two of the testing stage of the DART project was for each librarian in the project to use the DART rubric to review the 100 DMPs they had gathered from their institution. Each librarian worked separately in this analysis and only reviewed DMPs from their institution. We used the Qualtrics survey tool as our means of capturing our responses which allowed us to easily compile and compare our results. We conducted an analysis of our results and published our findings in the International Journal of Digital Curation (IJDC)[5]. The draft rubric and other supporting materials are openly available online through the Open Science Framework[6].

The final phase of the project is to further refine the DART rubric based on additional observations and suggestions for improvement made by project librarians. Using the refined rubric, I revisited the 100 DMPs and rescored some of my responses from phase two. Once this phase is complete the final rubric will be published and made available for anyone to use.

Data Management Plans from the University of Michigan

This report presents the findings of the review of DMPs generated by researchers from the University of Michigan during the third phase of testing for the DART rubric. DMPs were collected from the College of Literature, Sciences and the Arts (LSA) and from the College of Engineering (ENG). I then selected 50 DMPs produced by researchers from LSA and 50 DMPs produced by researchers from ENG. The selected DMPs were from applications that were submitted to the NSF between 2011 and 2014. Table 1 displays the breakdown of DMPs that were reviewed by the year they were written.

In the tables in this report, DMPs from LSA are represented in the blue columns, DMPs from ENG are represented in the red columns, and the white columns represent the full results from both LSA and ENG.

---

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

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

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

[5] Parham, S. W., Carlson, J., Hswe, P., Westra, B., & Whitmire, A. (2016). Using Data Management Plans to Explore Variability in Research Data Management Practices Across Domains. International Journal of Digital Curation, 11(1), 53–67. https://doi.org/10.2218/ijdc.v11i1.423.

[6] The DART Project: using data management plans as a research tool. https://osf.io/kh2y6/

Table 1 – The composition of DMPs selected for the DART project by year

DMP Year	LSA	ENG	All
2011	3	7	10
2012	16	11	27
2013	29	2	31
2014	2	30	32
Total	50	50	100

The breakdown of the 100 DMPs used in this study by the NSF directorate they were submitted to is shown in table 2. The DMPs reviewed in this study are concentrated in two of the NSF's seven directorates: the Mathematical & Physical Sciences (MPS) directorate for researchers in LSA and the Engineering directorate for researchers in ENG.

Table 2 - NSF Directorate of Reviewed DMPs

	LSA	% LSA	ENG	% ENG	All	% All
BIO - Biological Sciences	6	12%	1	2%	7	7%
CISE - Computer & Information Science & Engineering	3	6%	11	22%	14	14%
EHR - Education & Human Resources	1	2%	1	2%	2	2%
ENG – Engineering	3	6%	30	60%	33	33%
GEO – Geosciences	4	8%	2	4%	6	6%
MPS - Mathematical & Physical Sciences	21	42%	5	10%	26	26%
SBE - Social, Behavioral & Economic Sciences	10	20%	0	0%	10	10%
Unknown	2	4%	0	0%	2	2%
Total	50	100%	50	100%	100	100%

Of the 100 DMPs I reviewed, six of them indicated their research would not produce data as defined by the NSF. Therefore, these six were not analyzed any further.

Table 3 – Will the project produce data?

	LSA	ENG	All
Yes	45	49	94
No	5	1	6
Total	50	50	100

Overview of the Rubric

All DMPs submitted to the NSF are required to address the following five components regardless of the directorate they are submitted to[7]:

- Types of Data
- Data and Metadata Standards
- Policies on Access and Sharing
- Policies on Use and Reuse
- Archiving

Many of the NSF directorates ask for additional information from researchers to augment or supplement the base five sections. The DART rubric does account for these additional directorate based requirements and they were scored as a part of the review we conducted. However in the interest of brevity and clarity, I am not including the directorate-based responses in this report.

There are three possible responses for each of the questions in the rubric that are based on the NSF requirements: "complete", "addressed issue but incomplete", or "did not address the issue." Definitions of these rankings are presented after each table, followed by excerpts from the DMPs I reviewed to illustrate how content was ranked.

It's important to note that the DART rubric is not intended to evaluate the quality of the content of a DMP as much as its completeness. Some of the answers that I deemed to be "complete" were not necessarily good answers in my opinion, but they answered what the NSF had asked of them in a way that met the criteria of the rubric. Similarly, some of the responses that I deemed to be "addressed but incomplete" described what I might consider to be a good course of action but either left some questions about the researcher's understanding of what was being asked of them, the intent of the researcher, or how the researcher would actually respond.

In addition to questions based on the NSF requirements, we also created supplemental questions for the DART project rubric. These supplemental questions are meant to capture information of interest in reviewed to librarians in providing services, even though these areas of information are not explicitly required by the NSF. Responses from these questions are included in the report within relevant sections where appropriate and are noted as being supplemental.

Section 1 – Data

In the first section of a Data Management Plan, researchers are asked to describe what type of data will be generated, collected or captured as a part of the research project being conducted. In reviewing DMPs, we were looking for a clearly expressed, general description of the data that would be generated or collected over the course of the project. The results are presented in table 4.

[7] As stated in the NSF's Grant Proposal Guide:
https://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_2.jsp#dmp

Table 4 - Describes what types of data will be captured, created or collected

	Score LSA	% LSA	Score ENG	% ENG	Score All	% All
Complete / detailed	27	54%	34	68%	61	61%
Addressed issue, but incomplete	14	28%	8	16%	22	22%
Did not address the issue	4	8%	7	14%	11	11%
(No Data)	5	10%	1	2%	6	6%
Totals	50	100%	50	100%	100	100%

Complete: Clearly defines data type(s). E.g. text, spreadsheets, images, 3D models, software, audio files, video files, reports, surveys, patient records, samples, final or intermediate numerical results from theoretical calculations, etc. Researcher could also define data as: observational, experimental, simulation, model output or assimilation.

> "Raw computational data mainly consist of MD trajectories from the CHARMM and GROMACS, structure-activity matrix from the in-house QSAR, output from molecular docking and many in-house scripts and programs for simulation setup and analysis. Experimental raw data will include images, numerical data, and video recordings from various biophysical measurements such as NMR, CD, DSC, ITC, AFM, SEM, TEM, ThT, SPR, and confocal microscope." (from a DMP submitted to the MPS directorate)

Addressed but Incomplete: Some details about data types are included, but DMP is missing details or wouldn't be well understood by someone outside of the project.

> "We plan to produce data that explain reasons for the throughput variability in reentrant production lines and develop methods for its alleviation. These data will be obtained by mathematical and numerical analyses of the equations that describe production systems at hand. The results of these analyses will be mathematical theorems, numerical facts, continuous improvement procedures, and insights." (from a DMP submitted to the ENG directorate)

Did not address the issue: No details are included, or fails to adequately describe data types.

> "We do not anticipate that there will be any significant intellectual property issues involved with the acquisition of the data. In the event that discoveries or inventions are made in direct connection with this project, access to the data will be further governed by University of Michigan's policies pertaining to intellectual property, record retention, and data management. The PI will share all the theoretical achievements created during this project through publishing journal articles and also make all the numerical test experiments and their results widely available and usable." [This was the entire DMP] (from a DMP submitted to the MPS directorate)

Observations

Overall, the information provided in this section was rated to be the most complete of the five. This is a logical outcome as researchers are certainly used to considering what data they will generate over the course of their research and in describing their data to their funders and to others in their field. Still, only 61% of the reviewed DMPs at the University of Michigan were scored as being complete, with DMPs from the College of Engineering performing a bit better than LSA.

The results seem to indicate that researchers may not be fully familiar or clear on the requirements of the NSF in developing the data section of the DMP, or they may not be used to having to summarize their data before their research projects even begin. I did not review the content of the grant narrative which may have included information about the data, as this analysis was focused solely on the content of DMPs.

Recommendations

- Reference materials could be created to better explain this DMP requirement and what is expected from researchers for each section of the DMP with examples to illustrate how the requirements could be addressed. These materials could be made available to researchers by embedding them into the grant submission workflow.

- Brief workshops (one hour or less) on completing DMPs could be offered to researchers two weeks or so before important calls for proposals for major NSF grants are due.

Section 2 – Standards for Data and Metadata

Section 2 of the DMP requirement from the NSF asks researchers to describe the standards that they will apply to their data set. The expected application of standards generally covers two distinct areas: the formats of the data set and the metadata applied to the data.

Formats -

For the format portion of the requirement, we were looking for specific information about the formats of the data that would be generated in the project. DMPs that included specific format types or extensions were rated as being complete, whereas DMPs that mentioned the broad type of data files to be produced were marked as addressed but incomplete. Table 5 the extent to which researchers identified the data formats that they will employ for the data generated or collected over the course of the project.

Table 5 - Describes the data formats created or used during project

	Score LSA	% LSA	Score ENG	% ENG	Score All	% All
Complete / detailed	23	46%	28	56%	51	51%
Addressed issue, but incomplete	10	20%	9	18%	19	19%
Did not address the issue	12	24%	12	24%	24	24%
(No Data)	5	10%	1	2%	6	6%
Total	50	100%	50	100%	100	100%

Complete – Clearly describes file format standard(s) for the data.

> "Raw data will be in ASCII format from two different radiation detectors, a multimeter, and six different mass spectrometers. Processed data (after normalization of the samples to standard data) will be in Excel or .txt or .dat formats (data tables noting sample name, location, date, chemical data, and isotopic compositions)" (from a DMP submitted to an Unknown Directorate)

> "X-ray data will be stored as *.cif files. HPLC data will be stored as data files (*.dat), GC data will be stored as Schimadzu GC data files (*.gcd), GC/MS data will be stored as Schimadzu GC/MS solution data files (*.qdg), LC/MS and MS data will be stored as raw data files (*.raw); IR data will be stored in ACSII format (*.acs); optical rotation data will be stored as PID files (*.pid)." (from a DMP submitted to the MPS Directorate)

Addressed but incomplete – Describes some but not all file formats, or file format standards for the data. Where standards do not exist, does not propose how this will be addressed.

> "During physical characterization of materials, data in the form of spectra and images will be recorded via computerized data acquisition software. During structural characterization, data will be recorded via special instrument-specific software." (from a DMP submitted to the ENG directorate)

> "The data gathered from standard experimental instruments are already standardized by the vendors and will be kept in the same format and used as is." (from a DMP submitted to the MPS directorate)

Did not address – Does not include any information about data format standards.

> "The format of data generated is routine for the methods employed." (from a DMP submitted to the MPS directorate)

> "The data format includes digital data recorded by computers and instruments." (from a DMP submitted to the ENG directorate)

Metadata

For the metadata portion of the review, we were not necessarily looking for the application of a specific metadata standard (see table 7 below). Instead, we sought to determine to what extent researchers would provide contextual information about the data for others to be able to understand it.

Table 6 reports the extent to which researchers identified the metadata standards or metadata formats that will be applied to the data produced by the proposed project. Table 7 reports the number of researchers who listed a specific metadata standard in their DMP.

Table 6 – Describes the metadata standards or metadata formats that will be applied to the data

	Score LSA	% LSA	Score ENG	% ENG	Score All	% All
Complete / detailed	13	26%	6	12%	19	20%
Addressed issue, but incomplete	10	20%	19	38%	29	31%
Did not address the issue	22	44%	24	48%	46	49%
(No Data)	5	10%	1	2%	6	6%
Total	50	100%	50	100%	100	100%

Complete - The metadata standard that will be followed is clearly stated and described. If no disciplinary standard exists, a project-specific approach is clearly described. DMPs that provided details about the specific contextual information that would be captured were generally rated as being complete.

> (with metadata standard) "The Dublin Core will be used as the standard for metadata. The metadata set mainly consists of fifteen elements, for example, title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights; and these element set has been ratified as a national standard (i.e., ANSI/NISO Standard Z39.85) as well as an international standard (i.e., ISO Standard 15836)… According to the Dublin Core standard, thus metadata will be created as soon as the data is collected or produced, and this will facilitate not only efficiently managing the data by ourselves but also rapidly sharing the data with others." (from a DMP submitted to the ENG directorate)

> (without metadata standard) "Metadata will include time, date, and location of measurement, device measured, equipment used, and personnel present. Data will be identified by the date it was gathered, the name of the investigator who obtained it, and a short descriptor. A detailed description of the data and experimental conditions will be included in an accompanying laboratory notebook." (from a DMP submitted to the ENG directorate)

Addressed but Incomplete – The metadata standard that will be followed is vaguely stated. If no disciplinary standard exists, a project-specific approach is vaguely described. DMPs that indicated that

contextual information would be captured or created but did not provide details or specifics on the nature of the contextual information were generally rated as addressed but incomplete.

> "Meta-data explaining the system configuration for these experiments will be generated. Documents with proofs for various methods and algorithms used to provide concurrency semantics will also be generated." (from a DMP submitted to the CISE directorate)

> "Fabrication data is typically recorded in laboratory notebooks or process run-sheets that are maintained electronically. Experimental characterization data is similarly recorded through quantitative, pictorial, or quantitative observations in notebooks or electronic files." (from a DMP submitted to the ENG directorate)

Did not address the issue – No metadata standard is stated and no project-specific approach is described. Mostly those that fell into this category did not mention metadata or documentation at all. Other DMPs just did not report any meaningful information.

> "All experimental procedures will be recorded daily on laboratory notebooks by every members of [name of professor's] group." (from a DMP submitted to the MPS directorate)

Table 7 – Did the DMP mention a specific metadata standard?

	Score LSA	% LSA	Score ENG	% ENG	Score All	% All
Yes	8	16%	3	6%	11	11%
No	37	74%	46	92%	83	83%
(No Data)	5	10%	1	2%	6	6%
Total	50	100%	50	100%	94	100%

Although the application of a metadata standard is preferable, very few of the DMPs reviewed from the University of Michigan, only 11%, included a specific metadata standard. A few DMPs mentioned more than one metadata standard. As we did not include a full review of the submitted grant I did not have the context to determine if relevant metadata standards existed for the data being developed in each proposal, nor was I able to determine if a general metadata standard, such as Dublin Core, might be appropriate for the data set.

The metadata standards mentioned in DMPs from the University of Michigan are:

- FGDC & CSDGM – used for geospatial data
- EML [Ecological Markup Language] (mentioned in two separate DMPs)
- "VO [virtual observatory] compliant identifiers" – used for astronomical data
- FITS (mentioned in three separate DMPs) – used for astronomical data.
- ARC-LTER, ACADIS
- Dublin Core (mentioned in three separate DMPs)
- DDI [Data Documentation Initiative] – used for social science data
- Darwin Core

Observations

The instructions on developing a DMP provided by the NSF in their Grant Proposal Guide and in their data sharing policy in their Awards and Administration Guide focus on sharing data with other researchers. However, the NSF guidance does not require that researchers directly identify which research communities the data would be of use to or how the data would be prepared in ways that facilitate its use by the identified research community. Instead, the NSF asks that researchers include statements about what "standards" they will employ for their data. Asking researchers to identify standards for formatting and describing their data in their DMP appears to serve as a proxy for asking researchers to identify the communities of practice that would benefit from having access to their data.

Some of the instructions and guidance provided by the NSF directorates do mention the need to connect the data to a community of practice. The Geosciences directorate provides guidance for developing data management plans for four particular areas of study: Ocean Sciences, Earth Sciences, Atmospheric and Geospace Science and the Polar Sciences. These guides generally provide more directed and explicit instructions for researchers in completing their DMPs. The Biological Sciences directorate includes a statement on connecting data sharing to research communities:

> "BIO recognizes that each biology sub-discipline may have its own data management standards, and that accepted norms are changing as biology increasingly collaborates with other scientific disciplines. Therefore, each DMP should be appropriate for the data being generated and reflect the best practices and standards in the area of research being proposed." (https://www.nsf.gov/bio/pubs/BIODMP061511.pdf - p.1 2013)

Some of the DMPs reviewed in this study mentioned the community of practice who would find their data to be of value; however most of the DMPs reviewed did not mention a particular audience for their data. As most researchers were not asked explicitly to identify a community of practice this is not an unexpected outcome. However, without making the connection between identifying a research community who would benefit from having access to the data and employing standards relevant to the norms and practices of that community, the purpose of including standards becomes harder to understand and address.

From looking at the DMPs generated at the University of Michigan, many researchers did not appear to understand what was being asked of them in this section.

Researchers mentioned the formats of their data more frequently than the metadata that they were planning on using. However, researchers rarely included a rationale behind why they selected the particular formats for their data, nor did they link their decision to use a particular format in terms of it being a community standard. Instead, most of the researchers simply provided a list of the formats they intended to use without connecting them to management, sharing or archiving activities in their DMP.

Close to half of the reviewed DMPs did not address metadata. Even among those who did mention metadata, many did not provide any indication of a particular standard or any details of what metadata was to be collected or why. Other DMPs conflated documentation with metadata, listing the means by which they would track their work without indicating how what they were tracking would be structured and made useful to those seeking to understand and use their data. From this review it is not clear that researchers have an understanding of what metadata is or designed to do. Furthermore, it's not likely that researchers have an understanding of how to generate metadata that would aid in the discovery,

understanding or re-use of their data, especially without a defined connection to a particular community of practice as an intended audience for the data.

Recommendations

- Reframe the "standards" section of the DMP to further emphasize effective data sharing. Ask the researcher to identify communities of practice that would be a likely audience for the data and then describe how the researcher would prepare the data set to be discovered, understood, trusted and used by those communities. In particular, consider what someone from that community would need to know about the data and how that information will be generated and incorporated into the data set.

- In cases where metadata standards do not yet exist, are not sufficiently developed or are not appropriate for use for whatever reason, ask the researcher to describe the information that will be collected that could be used as metadata.

- Make a clear distinction between documentation and metadata. In addition to asking the researcher to describe the metadata that will be generated for the data set, ask about the types and nature of the documentation and other contextual information that will support discovery, understanding, trust and re-use of the data set.

Section 3: Policies for access and sharing

In some ways this section is the heart of the Data Management Plan as the NSF states that it wants to know how the proposal "will conform to NSF policy on the dissemination and sharing of research results" (NSF GPG, https://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_2.jsp#dmp). However, the NSF did not provide much in the way of direction to researchers as to how they should comply with this requirement, leaving it up to them to decide how and when to share data.

The two aspects that the NSF stated they wanted researchers to include are when the data would be made available to others outside of the project and details about how the data would be made available. The success of researchers from the University of Michigan in complying with these requirements are shown in Table 8 (when) and Table 9 (how). Table 10 provides more specifics on the means by which researchers at the University of Michigan plan on sharing their data from a supplemental question asked in the DART rubric.

Table 8 - Describes when the data will be made publicly available

	Score LSA	% LSA	Score ENG	% ENG	Score All	% All
Complete / detailed	9	18%	21	42%	30	30%
Addressed issue, but incomplete	18	36%	11	22%	29	29%
Did not address the issue	18	36%	17	34%	35	35%
(No Data)	5	10%	1	2%	6	6%
Total	50	100%	50	100%	94	100%

Complete – Clearly specifies when the data will be made available to people outside of the project.

> "We plan to make the data available toward the end of Year 3, by which time the data will be cleaned up and ready to be used for other researchers. We would like to withhold the data for a little longer if our data analysis has not been completed by then." (from a DMP submitted to the SBE directorate)

> "The PIs will make copies of data available to co- investigators, students, and others by request within 45 days from receipt of the request unless a longer period is necessary for protection of intellectual property or publication of data in a journal article." (from a DMP submitted to the ENG directorate)

Addressed but Incomplete - Verifies that the data will be made available outside of the project but does not identify when, such as a time frame (e.g., duration of the project, or for a period after the conclusion of the project).

> "Data will be available for access and sharing as soon as is reasonably possible. In the event that discoveries or inventions are made in direct connection with this data, access will be granted upon request once appropriate invention disclosures and/or provisional patent filings are made." (from a DMP submitted to the ENG directorate)

Did not address – Does not address when the data will be made available outside of the project.

> "The data produced from the proposed research and published in the peer-reviewed literature will be available for other researchers." (from a DMP submitted to the MPS directorate)

Table 9 - Provides details on how the data will be made publicly available

	Score LSA	% LSA	Score ENG	% ENG	Score All	% All
Complete / detailed	30	60%	28	56%	58	58%
Addressed issue, but incomplete	11	22%	17	34%	28	28%
Did not address the issue	4	8%	4	8%	8	8%
(No Data)	5	10%	1	2%	6	6%
Total	50	100%	50	100%	100	100%%

Complete - Includes specific details on the means by which the data will be made available. E.g., this may include a publically accessible data repository or a description of how the researcher or a 3rd party will provide access.

> "Since the excel formats could become unreadable over time as software systems change, final versions of all datasets will also be exported to and made available as ASCII and/or CSV data files, with accompanying command/syntax files, so future users will be able to access the data." (from a DMP submitted to the GEO directorate)

> "Access to the data, if requested, will be provided through direct access to our data-storage facilities. For this, we will provide guest-access to our intranet, and data, I/O-routines, and other access-protocols will be provided for access, retrieval, and modifying the simulation results." (from a DMP submitted to the ENG directorate)

> "To facilitate the sharing of data, the project data will be made publicly available via a university-controlled repository, Deep Blue." (from a DMP submitted to the ENG directorate)

Addressed but incomplete – Provides vague or limited information on how the data will be made available, or details about sharing can be inferred from the mention of a repository or archive that will be used for depositing the data.

> "Currently, to our knowledge, no national or international data set repositories are available for noble gas measurements in water samples. As such, we are currently exploring the possibility to create a data set repository in one of our departmental servers at the University of Michigan that will be made available to the public at large." (from a DMP submitted to the GEO directorate)

> "The vast majority of data that other researchers will be interested in is provided as supporting information freely available through the journal websites. Unpublished data is generally not made widely available for public view, but will be distributed for valid reasons upon request." (from a DMP submitted to the MPS directorate)

Did not address – No details are provided about how the data will be made available.

> "The PI will share all the theoretical achievements created during this project through publishing journal articles and also make all the numerical test experiments and their results widely available and usable." (from a DMP submitted to the MPS directorate)

> “These data of the research projects themselves will be handled by the respective PIs according to the standards of the field.” (from a DMP submitted to the MPS directorate)

Table 10 - How are the researchers planning to share the data? (Supplemental)

Note: Some DMPs listed multiple methods of sharing. The number of responses does not add up to 100 as a result. Some researchers who listed multiple methods for sharing their data indicated that they would use one method prior to an event, such as completing the project or publishing their work. and then another method afterwards. Other researchers who listed multiple methods did not indicate if they would switch methods based on an event and so this information was not recorded as a part of the rubric.

	Score LSA	% LSA	Score ENG	% ENG	Score All	% All
On request	15	30%	17	34%	32	32%
Journal / Supplement	13	26%	15	30%	28	28%
Personal website	6	12%	17	34%	23	23%
Data center or repository	21	42%	2	4%	23	23%
Institutional repository (Deep Blue)	9	18%	10	20%	19	19%
Other method (describe)	8	16%	11	22%	19	19%
Did not specify	3	6%	8	16%	11	11%
Conference proceedings	3	6%	1	2%	4	4%
Not planning to share data	3	6%	0	0%	3	3%
Book	1	2%	0	0%	1	1%
Thesis / Dissertation	0	0%	1	2%	1	1%
Total	82		82		164	

Observations -
It's interesting to note that researchers provided more complete answers about "how" they were planning on sharing the data than "when" in their DMPs. Researchers seem to have an understanding that how they share the data is a major component of the DMP. It may be easier to provide a more definite response to the question of how the data will be shared rather than when as the available options can more easily be identified and planned out than to project the timing of when findings will be known or published, or when the data will be ready to share.

Researchers from the College of Engineering provided a more complete response to the question of when they would share their data in their DMPs than Researchers from the College of LSA did. It is not clear what might be behind the differences in the completeness of the responses.

Statements on when the data would be made available included definite times (6 months, by the end of year 3), but more often was associated with a broadly defined event (after the project, upon publication of the paper, upon request, in a reasonable amount of time).

The publication of the findings was a frequent consideration for both how and when the data would be released. Several DMPs that were not explicit about when the data were to be released instead implied that their data would be shared as a part of the publication process (often as supplemental files). A few DMPs went as far as to state that their data sharing would take place through the charts, graphs and tables that would appear in their publications. This of course is not an effective plan for sharing data and

runs counter to the intent of the data management plan requirement. These statements may indicate an extreme reluctance on the part of the researcher to share their data. Thankfully, these extreme sentiments were relatively rare. As I did not have access to the reviewer feedback, I do not know if these statements were identified by the NSF as being unacceptable or whether or not the research was asked to modify their DMP.

In this study, researchers from the College of LSA were much more likely to list a data center or repository as a means through which they shared their data than researchers from the College of Engineering. Conversely, researchers from Engineering listed sharing their data through a personal website somewhat more than researchers from LSA did. It may be the Engineers have fewer data centers or repositories available to them than researchers from LSA do, or they simply are not aware of the existence of these options as a means for sharing their data. Alternatively, they may not want to turn over their data to a 3rd party for sharing, preferring instead to keep control over the data and how it is made available.

Recommendations

- Further research into the data sharing resources available to Engineering researchers and their preferences in making use of them. If resources are not available or desirable, how can the library position our services and the Deep Blue Data repository to connect with the College of Engineering more effectively?

- There is a need to close the loop on what the researcher wrote in their DMP with what the researcher has actually done to make their data available. Many of the projects represented in the DMPs should be at or close to the point where the data will need to be made available outside of the lab. Did the researcher follow through with their plan to share to share the data using the method described in their DMP? If so, did the process go as planned and were the results satisfactory, or where there unforeseen challenges? If not, what prevented the researcher from following through with their original plan? Is the data current shared through another method, or at all?

- Investigate how the library could better connect with researchers at, or more ideally well before, the time when they are preparing their findings for publication to provide consultation on available options for publishing or sharing their data in ways that would maintain the connection between data and publication and maximize the likely impact of sharing their data.

Section 4 – Policies for the Use and Reuse of the data

In addition to making research data accessible beyond the project itself, the NSF is interested in knowing what audiences for the data will be allowed to do with it. Section 4 addresses to what extent researchers described what a potential user of the data would be able to do with it. More specifically, the NSF asks researchers to explain in their DMP what rights or limitations a potential user of the data would have to reuse the data, to distribute the data and to produce derivatives of the data.

Table 11 - Describes the policies or provisions in place governing the use and reuse of the data

	Score LSA	% LSA	Score ENG	% ENG	Score All	% All
Complete / detailed	10	20%	5	10%	15	15%
Addressed issue, but incomplete	16	32%	21	42%	37	37%
Did not address the issue	19	38%	23	46%	42	42%
(No Data)	5	10%	1	2%	6	6%
Total	50	100%	50	100%	100	100%

Complete – Clearly explains the policies or guidelines in place governing future reuse of the data

> "The data will be released to the public domain under Creative Commons Zero which allows free use of the data without legal and technical impediments. Citation of the data is expected to follow community norms for scholarly communication." (from a DMP submitted to the BIO directorate)

> "The authors will retain rights to the data until the resulting publication is produced, typically within two years of generating the data. There will be no limitations on the re-use or re-distribution of the data produced by this project once published." (from a DMP submitted to the BIO directorate)

> "Data can be re-used and re-distributed with the obligation to inform the PI of further distribution and acknowledgment / of original data sources." (from a DMP submitted to the ENG directorate)

Addressed but incomplete – Provides a general overview of how data may or may not be reused, or the applicability of the policy can be inferred from general/ broad/ blanket statements about data being made open or being kept private, or policies can be inferred based on the sharing location.

> "Re-use, re-distribution and protection of derivative products will be governed by the policies of the University of Michigan, the university in collaboration with Google™, the State of Michigan, NSF-sponsored XSEDE™ and the United States federal government." (from a DMP submitted to the GEO directorate)

> "Generally, images and data published through the [project] website are made available for fair use as defined in the United States copyright laws." (from a DMP submitted to the BIO directorate)

> "There will be no problems or restrictions on the re-use of the data for the purposes of additional analyses that another researcher might wish to perform. No limitations will be imposed on the re-use of the data by a legitimate interested party." (from a DMP submitted to the ENG directorate)

Did not address – Does not address use/reuse of the data.

> "In the event that discoveries or inventions are made in direct connection with this project, access to the data will be further governed by University of Michigan's policies pertaining to

intellectual property, record retention, and data management." (from a DMP submitted to the ENG directorate)

"No issues regarding protection of privacy, confidentiality, intellectual property etc. are foreseen for this work, and the work does not require any subject testing." (from a DMP submitted to the CISE directorate)

Table 12 - Describes the policies or provisions for redistribution of the data

	Score LSA	% LSA	Score ENG	% ENG	Score All	% All
Complete / detailed	14	28%	5	10%	19	19%
Addressed issue, but incomplete	12	24%	19	38%	31	31%
Did not address the issue	19	38%	25	50%	44	44%
(No data)	5	10%	1	2%	6	6%
Total	50	100%	50	100%	100	100%

Complete – Clearly explains policies or guidelines for future redistributions of data

"Once the data are collected; there will be no limitations on sharing the data for re-use or re-distribution. The only exception is the data collected from industrial partners which have to be distributed in keeping with their requirements." (from a DMP submitted to the ENG directorate)

Addressed but incomplete - Provides a general overview of how and when data may be redistributed, or policies can be inferred based on the sharing location.

"The investigator will permit the re-use and re-distribution of our code and data case by case." (from a DMP submitted to the ENG directorate)

Did not address – Does not address redistribution of the data.

(no examples)

Table 13 - Describes policies or provisions for building off of the data, such as through the creation of derivatives

	Score LSA	% LSA	Score ENG	% ENG	Score All	% All
Complete / detailed	4	8%	0	0%	4	4%
Addressed issue, but incomplete	15	30%	13	26%	28	28%
Did not address the issue	26	52%	36	72%	62	62%
(No Data)	5	10%	1	2%	6	6%
Total	50	100%	50	100%	100	100%

Complete - Clearly describes guidelines or policies governing the creation of derivatives from the data generated by the project

> "No restrictions on re-use, distribution, or production of derivatives will be placed on the samples or the data following initial publication." (from a DMP submitted to the BIO directorate)

Addressed but incomplete – Provides a general overview of the creation of derivatives from project data, or policies can be inferred based on the sharing location.

> "By using the open literature for the dissemination of results and distributing the results as open source software under an approved open-source license, re-use, re-distribution and production of derivative works will be easily provided for." (from a DMP submitted to the ENG directorate)

Did not address – Does not discuss production of derivatives from the data.

> (no examples)

Observations

This was the section of the DMP that researchers seemed to have the most trouble responding. Very few researchers provided complete answers to the questions asked by the NSF. Those that did gave very short answers often indicating that would not make any restrictions in allowing others to reuse, redistribute or create derivatives from the data. The brevity of these responses made it difficult to determine if researchers fully understood the potential issues in making their data available to others or if they simply inserted a broadly-worded statement on making their data widely available and usable into their DMPs because they felt it was expected of them. For example, researchers frequently express a desire for others to cite their data set as a means to receive credit for their work and as a measure of its impact. In the DMPs I reviewed some researchers mentioned that others should cite their data set as a condition of use, but many others did not.

Multiple researchers mentioned U-M policy as the guiding principle behind how the data may be used, without providing much if any detail about the content of these policies and how it would affect the use, distribution and derivatives made from the data set. A number of researchers also mentioned state and federal government or policies made by the NSF as factors in how the data could be used. Though the DMPs 2 page limit may have dissuaded some from providing more details about IP issues, it may also be that researchers do not fully understand the IP issues over their data or what, if any, IP rights they could assert. It's not likely that grant reviewers would have an understanding of U-M policy towards acceptable use of data external from the DMP and so it's unclear as to what basis they would have to evaluate the quality or completeness of the researcher's statement.

Researchers may also be concerned about making statements in the DMP that are contradictory to university, state or federal policies. University policy on research data is spread across several Standard Practice Guides, mixed in with policy on administrative data, and not written in language that researchers would readily understand. State and federal policies may not be readily accessible and their relevance to specific situations may not be clear. These factors may inhibit researchers from feeling confident in making clear and complete statements on how their data could be used, distributed, or used to create derivatives. Researchers may also be unaware of the support that is available to them at U-M to help them make informed decisions on intellectual property issues for their data.

Recommendations

- More clarity from the NSF has to what IP rights researchers could claim and still meet the NSF requirement that data be shared with others.

- More clarity from the University of Michigan on their policy of sharing research data as required by the NSF and other funding agencies. What requirements or expectations does the university have regarding sharing research data and how can researchers demonstrate compliance? What questions are researchers able or expected to decide for themselves?

- More exploration about the depth of researcher's knowledge of intellectual property issues as they pertain to data. Are there gaps between what they should know and what they do know that need to be addressed? What knowledge and support do they need to make informed decisions about the use or re-use of their data?

- Develop training sessions and support materials for researchers on issues in sharing data to provide clarification and direction.

Section 5 – Plans for Archiving or Preservation of Access

Beyond simply making research data available to others, the NSF is interested in ensuring that the data are accessible for the long term (although most directorates do not provide a definition for what long term actually means). Section 5 of a DMP asks researchers to provide information on how their data set(s) will be archived.

Table 14 - Indicates whether or not the data will be archived

	Score LSA	% LSA	Score ENG	% ENG	Score All	% All
Complete / detailed	14	28%	19	38%	32	32%
Addressed issue, but incomplete	17	34%	22	44%	39	39%
Did not address the issue	14	28%	8	16%	23	23%
(No Data)	5	10%	1	2%	6	6%
Total	50	100%	50	100%	100	100%

Complete / Detailed: Clearly indicates whether or not data will be archived, including digital data and physical samples where applicable.

> "All records, including unpublished data, will be maintained for a minimum of three years after publication or conclusion of the work. These digital files will be maintained in three locations to ensure redundant access: (1) on the lab/computer equipment where the data were originally collected, (2) on desktop computer(s) on which the data were analyzed/processed and (3) backed up on our long term data storage server housed on the campus of the University." (from a DMP submitted to the ENG directorate)

Addressed but Incomplete: The researcher describes a plan for maintaining access to archived data but does not fully explain the plan, or the reviewer is left with questions about the plan or how it will be fulfilled.

> "Data will be stored on the [name of] website which maintains data sets for more than 500 researchers, commercial applications users, and others worldwide. Backup of the data will be maintained on the PI's website." (from a DMP submitted to an Unknown directorate)

> "We do not have any end date in mind for our data's availability. That means the data will be posted semi-permanently, as long as the PI is at the University of Michigan, and we will make every effort to make the data available in other web spaces if the PI should move to a different post." (from a DMP submitted to the SBE directorate)

> "Archiving and preservation of access to products from the proposed research is provided by and according to the policies of the University of Michigan system in connection with / Google™ and NSF-sponsored XSEDE™." (from a DMP submitted to the GEO directorate)

Did not address the issue: Makes no mention of intent to archive or preserve digital or physical data, despite the expected generation of such by the project. Or the mention is so vague as to not have any meaning.

> "The researchers will keep copies of the data, but the data cannot be made available for replication." (from a DMP submitted to the SBE directorate)

> "The results obtained in this project will be archived as papers in archival / journals and conference proceedings." (from a DMP submitted to the ENG directorate)

Not every directorate required researchers to address how they were planning on archiving their data and so Table 15 is to be considered supplementary.

Table 15 - How are the researchers planning to archive their data? (Supplemental)

	Score LSA	% LSA	Score ENG	% ENG	Score All	% All
Did not specify	18	36%	21	42%	39	39%
Institutional repository (Deep Blue or other)	10	20%	7	14%	17	17%
Personal computer or storage media (e.g. external drive)	5	10%	10	20%	15	15%
Data center or repository	11	22%	3	6%	14	14%
Other method (describe)	8	16%	6	12%	14	14%
Unit-, college- or university-level IT storage/servers	2	4%	10	20%	12	12%
Personal /research group website	2	4%	5	10%	7	7%
Journal / supplement	3	6%	3	6%	6	6%
Not planning to archive data	4	8%	0	0%	4	4%
Conference / proceedings	0	0%	2	4%	2	2%
Total	63		67		130	

Note: Some DMPs listed multiple methods of archiving their data. The number of responses does not add up to 94 as a result.

Another supplemental question that we included in reviewing DMPs was on the length of time the researcher intended to archive their data.

Table 16 - Did the researcher state how long they plan on preserving the data? (Supplemental)

	Score LSA	% LSA	Score ENG	% ENG	Score All	% All
Yes	8	16%	28	56%	36	36%
No	37	74%	21	42%	58	58%
(No Data)	5	10%	1	2%	6	6%
Total	50	100%	50	100%	94	100%

Observations

Many researchers did not seem to understand what was meant by archiving the data (and the NSF does not provide a definition). Backing up and storing the data was often mentioned as an archival solution. In addition, researchers frequently did not include a statement on continued access to their data.

Researchers often implied that their proposed solution to make their data available would also serve as a means to archive their data. If the researcher's sharing strategy was to make their data available through a server or other system under their control, they might mention that they would perform routine backups of their data. If the researcher's solution was to make their data available through a 3rd party, such as a data repository, they rarely described the archival services or measures taken by the repository. Archival services and capabilities vary from repository to repository, assuming the repository even provides archival services, and so it would be difficult for a reviewer to evaluate the feasibility of the researcher's archival strategy without this additional information.

As seen in Table 16, there is a wide disparity between researchers from the College of LSA and from the College of Engineering. More than half of Engineering researchers provided information about how long they would archive their data while only 16 percent of researchers from LSA included this information. There was also some variation between researchers from LSA with those in ENG over the amount of time listed for archiving data. The durations stated by researchers from LSA were wide ranging; anywhere from two years to indefinitely. Durations stated by researchers from ENG also included a wide range of time frames, but they were more clustered around a period of three years. This may be due to the guidance produced by the NSF's Engineering Directorate including a statement the: "Minimum data retention of research data is three years after conclusion of the award or three years after public release, whichever is later." These results indicate that researchers may respond when the directorates provide more detailed expectations for the treatment of their data. It is worth noting that a review of DMPs cannot reveal whether or not researchers followed through with what they stated in their plans. Researchers may have listed archiving their data for three years because they believed that this is what reviewers wanted to hear rather than what they actually they actually plan to do.

Recommendations

- Provide researchers a clearer and more complete understanding of what it means to archive data and the steps that are needed to archive data successfully. In particular, address how archiving data is different from long-term storage and making back-ups of the data.

- Work with researchers to reinforce the connection between archiving data and providing long-term access to the data. Ideally, this would include maintain the value and usability of the data through updating documentation, format migration and other preservation measures that go beyond the current DMP requirement.

- Work with researchers in evaluating the archival services provided by the data repositories they have identified as suitable for their sharing their data and determine if the services provided are sufficient for their needs. Continue to develop the capabilities of the Library's Deep Blue Data repository based on researcher needs and feedback.

Supplemental Questions: U-M services mentioned in DMPs

The rubric included supplementary questions to gather information about the types of services that were mentioned in DMPs. Table 17 captures how often services from the library were mentioned in DMPs and Table 18 notes how often other services available to U-M researchers were included.

Table 17 – Did the DMP make any reference to library services?

	Score LSA	% LSA	Score ENG	% ENG	Score All	% All
Yes	16	32%	11	22%	27	27%
No	29	58%	38	76%	67	67%
(No Data)	5	10%	1	2%	6	6%
Total	50	100%	50	100%	100	100%

Of the 27 references to the library services, 23 of them were to "Deep Blue", the library's institutional repository (the library's data repository, Deep Blue Data, was not released until September 2017). The references to Deep Blue were evenly split between DMPs from the College of LSA (12) and ENG (11). The other 4 references made to library services were all from LSA and included three general references to institutional repositories, and one general reference to library services.

Table 18 - Did the plan make any reference to specific campus services? If yes, describe.

	Score LSA	% LSA	Score ENG	% ENG	Score All	% All
Yes	18	36%	23	46%	41	41%
No	27	54%	26	52%	53	53%
(No Data)	5	10%	1	2%	6	6%
Total	50	100%	50	100%	94	100%

The campus services mentioned in DMPs were generally centered on the information technology services and resources provided by the university. References to data storage servers and systems was the primary type of information technology mentioned in DMPs, followed by systems used to back-up the data. Other information technology based resources that were mentioned included High Performance Computing (HPC), support for databases, and data management and sharing tools (iRODS, Box, Google Drive, Globus). Some DMPs did not mention specific services or resources and simply stated that they would be working with their campus, college or department based IT unit.

Outside of information technology based services, a number of researchers mentioned the campus IRB and the ethical training required for human subject based research: Program for Education and Evaluation in Responsible Research and Scholarship (PEERRS). Interestingly, two of the DMPs that referred to PEERRS were not engaged in human subjects based research. Other researchers mentioned CTools (the campus based learning management system) and the Office of Technology Transfer.

Observations
From this review it appears that the library is already recognized by some researchers as providing services to support sharing and archiving data. It's not clear just from reviewing the DMP how the researcher may have learned of Deep Blue as a service that would satisfy funding agency requirements or their level of understanding regarding the support available for guiding them in preparing their data sets for deposit.

Recommendations

- The library should explore how we could further promote the use of our Deep Blue Data repository across campus. We can also help researchers identify a disciplinary repository that is suited to the needs of the data generated by a particular field.

- The library should take steps to connect with researchers who indicate that they want to make use of Deep Blue Data to guide them in preparing their data for its eventual deposit. Our suite of research data services are designed to help researchers move through the data lifecycle, from the beginning (when DMPs are developed) to the end (when data are deposited into a data repository for sharing and preservation). We need to close the loop with researchers to not only ensure that their data are deposited, but that they are deposited in ways that enable them to be discovered, understood and used by others.

- The library should reach out to other campus agencies that provide support for data management, sharing and curation and determine how our services compliment or fit together with their services. In particular we should explore how we could have the library's research data services represented further through these agencies and how in turn we could better represent and connect users to the services these agencies provide.

Discussion

The NSF Data Management Plan requirement and its Dissemination and Sharing of Research Results policy emphasize that researchers are expected to share their data with other researchers as a condition of receiving funding, but provide little to no guidance on how to manage, share or preserve data successfully. The NSF and its directorates instead rely on researchers following community driven best practices. From a high level perspective this is a sensible approach as the needs and norms surrounding data vary depending on the community and how they understand and use data. Some research communities have already invested a considerable amount of time and effort into considering the value of the data they produce to the field and how data ought to be managed, shared and preserved as an asset. Many other communities though are only beginning to engage in these discussions and have not yet fully considered the issues surrounding how they could share research data effectively. I believe that the NSF understands the situation and is taking steps to support research community engagement on

issues surrounding data sharing.[8] However in the interim the absence of standards, resources and training for data sharing in many communities means that researchers are left on their own to figure out how to prepare their data for its eventual dissemination and use by others.

I understand the reluctance of the NSF and other funding agencies to provide more direct or concrete instructions in their data management plan requirement so as not to appear that they are dictating to researchers how their data should be managed, shared and archived. However, I do believe that there is some middle ground that funding agencies ought to consider.

First, I believe that the NSF should reframe their data management requirement as a data sharing requirement as this would more accurately convey the intent of the requirement. The components of the data sharing requirement should be centered on the researcher providing enough information for another person in their field to be able to discover, understand and use the data set.

Second, I would ask that the NSF to provide more explanation to researchers about the particular components that are required and why they are important. Researchers may not understand what metadata is or how it applies to their data set. They may not understand why they are being asked to provide information about how the formats they use for their data. A brief explanation may go a long way in helping researchers better understand the requirements and in responding to them in meaningful ways.

Finally, I would encourage the NSF and other funding agencies to develop better connections and workflows with data repositories and other organizations that provide support to researchers in managing, sharing and archiving research data. Mutual exchanges of information would benefit both parties. For instance, data support agencies would benefit from learning more about the content of the data management plans being submitted to the NSF and when the could expect researchers to submit their data into the data repositories they support. The NSF would benefit from learning more about good practice in managing, sharing and preserving data and how this information might be communicated to researchers through incorporating it into the guidance provided by the NSF.

I concede that the suggestions I have made are not likely to be implemented by the NSF in the near future. However, there are actions that could be taken locally to further support researchers who are subject to the data management, sharing, and archiving requirements of the NSF and other funding agencies. First, this study explores only one facet of how researchers understand and respond to requirements on managing and sharing data and is incomplete by itself. Additional research should be conducted to increase our understanding of actual practices taken by researchers and to identify gaps in the support structures provided by the university to help researchers manage, share and archive data. In particular, we need to have a better understanding not just of what researchers tell funding agencies that they are going to do, but what they actually do once the grant has been awarded. Do they put the data management practices they described in their DMP into place once work begins? Do they share

[8] As evidenced through reports from the NSF such as “Today’s Data, Tomorrow’s Discoveries: Increasing Access to the Results of Research Funded by the National Science Foundation” (2015) https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf

their data with others as stated in their DMP? If not, what barriers did they encounter and can those barriers be mitigated somehow? Further study of researcher practices would help inform the library and other agencies at the university that provide support for data and enable them to provide more targeted services.

Second, the library provides a suite of services designed to support the data management, sharing and archiving needs of researchers across the lifecycle of their data. This includes providing consultations to researchers who are developing a data management plan for their project and a data repository, Deep Blue Data, where any researcher affiliated with the U-M can deposit their data to make it available to the world and ensure that it will be properly archived. The services provided by the library are only one component of the services needed by researchers to successfully manage, share and archive their data over the course of its lifecycle. The Office of Research and Sponsored Programs coordinates the submission of grants and monitors the award to be sure that the requirements made by the funding agency are fulfilled. The LSA Information Technology (LSAIT) unit, the Computer Aided Engineering Network (CAEN), Advanced Research Computing (ARC) and other IT units provide a wide range of information technology resources and support for researchers to store, backup, secure and otherwise manage their data. Many other campus units such as the Institutional Review Board, the Interuniversity Consortium for Political and Social Research (ICPSR) and the Technology Transfer office may also potentially play a role in supporting the management, sharing and archiving of research data depending on the nature of the research. Although each unit is certainly aware of the others and probably has at least a high-level understanding of the roles and services they provide, there is not a strong coordinated effort between support agencies to address researcher needs surrounding the data management plan requirements of funding agencies. One exception is the collaboration between the Engineering Library, CAEN and the Associate Dean for Research Office at the College of Engineering in offering a consulting service for researchers writing DMPs. This collaboration enables researchers to receive expert advice as needed in areas relevant to their specific research in crafting their DMPs. This program could be a model for extending collaborations across university agencies to provide needed support in managing, sharing and archiving data.

Conclusion

The introduction of the data management plan requirement by the NSF in 2011 was an important milestone in the open data movement, recognizing the importance of access to data to support the needs and practices of 21st century research. However, managing, sharing and archiving research data in ways that enable data sets to be discovered, trusted and used by others are not yet common practice in many research fields. Moreover, the resources and tools need to enable sharing and archiving data are still being developed and the institutions where research takes place.

It's clear that more work is needed by the NSF, universities and libraries to realize the benefits that having access to collections of well-curated data would provide researchers and to implement the services and support needed to create these collections. It's my hope that this analysis of how

researchers have responded to the data management plan requirement contributes to the continuing discussion and action towards addressing how we provide this support at the University of Michigan.