JLSC

ISSN 2162-3309 | JLSC is published by the Pacific University Libraries | http://jlsc-pub.org

Volume 3, Issue 3 (2015)

Assessing Research Data Management Practices of Faculty at Carnegie Mellon University

Steve Van Tuyl, Gabrielle Michalek

Van Tuyl, S., & Michalek, G. (2015). Assessing Research Data Management Practices of Faculty at Carnegie Mellon University. *Journal of Librarianship and Scholarly Communication,* 3(3), eP1258. http://dx.doi.org/10.7710/2162-3309.1258

External Data or Supplements:

Van Tuyl, S., Michalek, G. (2014). 2014 faculty data survey instrument [survey instrument]. Retrieved from Carnegie Mellon University Digital Collections. http://dx.doi.org/10.1184/RDL/01092015

ISSN 2162-3309 10.7710/2162-3309.1258

RESEARCH

Assessing Research Data Management Practices of Faculty at Carnegie Mellon University

Steve Van Tuyl
Data and Digital Repository Librarian, Oregon State University Libraries & Press
Gabrielle Michalek
Head Scholarly Publishing, Archives, and Data Services, University Libraries, Carnegie Mellon University

INTRODUCTION Recent changes to requirements for research data management by federal granting agencies and by other funding institutions have resulted in the emergence of institutional support for these requirements. At CMU, we sought to formalize assessment of research data management practices of researchers at the institution by launching a faculty survey and conducting a number of interviews with researchers. **METHODS** We submitted a survey on research data management practices to a sample of faculty including questions about data production, documentation, management, and sharing practices. The survey was coupled with in-depth interviews with a subset of faculty. We also make estimates of the amount of research data produced by faculty. **RESULTS** Survey and interview results suggest moderate level of awareness of the regulatory environment around research data management. Results also present a clear picture of the types and quantities of data being produced at CMU and how these differ among research domains. Researchers identified a number of services that they would find valuable including assistance with data management planning and backup/storage services. We attempt to estimate the amount of data produced and shared by researchers at CMU. **DISCUSSION** Results suggest that researchers may need and are amenable to assistance with research data management. Our estimates of the amount of data produced and shared have implications for decisions about data storage and preservation. **CONCLUSION** Our survey and interview results have offered significant guidance for building a suite of services for our institution.

Received: 04/29/2015 Accepted: 08/22/2015

Correspondence: Steve Van Tuyl, 121 The Valley Library, Oregon State University, Corvallis, OR 97331-4501, steve.vantuyl@oregonstate.edu

IMPLICATIONS FOR PRACTICE

1. Combining survey and interview methods can provide a helpful combination of breadth and depth when gathering information from faculty.
2. Many researchers are aware of regulatory mandates, but day-to-day research data management practices do not always align with best-practices
3. Preliminary estimates of the amount of data produced and shared by researchers have important implications for future discussion about storage and preservation of data.

INTRODUCTION

On February 22, 2013, the White House Office of Science and Technology Policy released a memo calling for increased public access to the results of federally funded research, including research publications and datasets (Holdren, 2013). This memo, coupled with additional recent and forthcoming requirements from major research funding agencies (e.g., National Science Foundation (NSF), National Institutes of Health (NIH), The Gates Foundation), signals a shift in policy at the federal and funder level.

At Carnegie Mellon University (CMU), as at most other research universities, the management of scientific data is largely left to individual researchers. Many universities, however, are implementing policies and service infrastructures that support researchers' needs to manage, document, and share their data. Given impending federal mandates for data sharing, we feel that it is important to coordinate the development of these services to avoid duplication of efforts across university units, but more importantly, to provide clarity and ease of compliance to researchers in our community.

In 2011, the University Libraries, Computing Services, and members of the Office of the VP for Research began discussing whether and how to provide data services on campus and who the service providers would be. As part of this initiative, we formed the Data Management Services Group (DMSG), which has conducted ongoing informal assessments of the data landscape at CMU and followed the evolution of data initiatives at institutions around the world. These informal assessments included discussions with faculty, administrators, and technology professionals at the university, a pilot faculty survey to explore research data management practices (unpublished), and an environmental scan of what services other institutions were beginning to provide in this area. By 2012, we had begun developing pilot services for researchers and students at CMU including workshops on research data management, exploratory discussions with research administrators about information sharing (e.g. sharing reports with DMSG on in process or recently funded grants), and

discussions with faculty who had previously expressed interest in potential research data management services (e.g. data repository, metadata assistance). This set of pilot services was meant to help inform future data services at CMU, rather than act as a full-blown suite of services. Information gathered during these pilot interactions with researchers indicated that our understanding of the breadth and depth of data needs at the university was limited and that further information would be required to define the direction and content of future services.

The research presented in this manuscript is focused on understanding the breadth and depth of data services needs of researchers at CMU. This includes understanding the modes by which faculty currently manage research data, where improvements could be made to data management practices, and how the DMSG could target new services to meet these needs. This paper reports on our investigation of the state of research data management at CMU as well as researchers' awareness of data mandates from funders and their attitudes towards university-provided data services. This information was gathered in two primary ways: a survey of a large faculty sample of to investigate broad trends in data management across fields and in-person interviews with a smaller faculty sample to more deeply probe both their data practices and their attitudes towards data services on campus. In addition, we augmented information gathered in the survey and interviews with data collected from university research administration to help contextualize some of our results.

LITERATURE REVIEW

Spurred, in large part, by mandates for data sharing and research data management (e.g. Holdren 2013), universities (specifically academic libraries) and scientific communities have been exploring data management and sharing practices of researchers to try to understand the extent to which a "data problem" exists. Research from Australia and the United Kingdom (UK) from before most of the data sharing requirements emerged in the US gave an indication of what might be expected in the United States (US). Beagrie, Beagrie, & Rowlands (2009) in the UK and Henty, Weaver, Bradbury, & Porter, S (2008) give indications that, across institutions, data management and sharing practices show similar patterns, with data management practices that work well for researchers but create issues for long term preservation, sharing, and reuse. These same studies also indicate that researchers are eager to understand more about data management best practices through training and outreach.

Over the past few years, academic librarians in the US have begun to conduct similar investigations into the data management practices of researchers, with similar results. Almost without exception, survey and interview research exploring data management and

sharing practices in academia exhibit striking homogeneity including: a general lack of data sharing through repositories and journals in favor of sharing via personal communication, researchers rarely creating metadata or other documentation for data, backup and storage practices that are of questionable efficacy and on questionable media, and concerns among researchers about the amount of time required to prepare data for sharing and the potential for misuse of data (Akers & Doty, 2013; Doty, 2012; Editorial, 2011; Parham, Bodnar, & Fuchs 2012; Steinhart, Chen, Arguillas, Dietrich, & Kramer, 2012; Cragin, Palmer, Carlson, & Witt, 2010; Peters & Dryden, 2011). On the other hand, the literature is similarly replete with survey and interview results suggesting that researchers are interested in training and guidance for research data management best practices (Akers & Doty, 2013; Parham et al., 2012; McClure, Level, Cranston, Oehlerts, & Culbertson, 2014, Wright et al., 2013).

These results indicate that while data management practices are not ideal in many cases, there is a desire on the part of researchers for outreach and training in best practices. This broad-scale agreement on the problems of research data management and desire for training in best practices do not, unfortunately, obviate the need for the exploration of local research data management practices and needs. Cragin et al. (2010) and Whitmire, Boock, & Sutton (in press) both note that there is a dichotomy between the need to provide overarching services in this area and the need to identify, at the domain-level, specific services needed by individual researchers.

There have been a wide variety of approaches used to identify the research data management practices and service needs of researchers at universities and academic institutions. Many researchers have used institution-wide faculty/researcher surveys to gather information about research data management practices at their institutions (e.g. Akers & Doty, 2013; Beagrie et al., 2009; Henty et al., 2008; Parham et al., 2012; Whitmire et al., in press) while other survey projects have focused on specific sub-groups within the university such as NSF awardees (Steinhart et al., 2012), STEM faculty (D'Ignazio & Qin, 2008), or at a teaching college (Scaramozzino, Ramirez, McGaughey, Ramírez, & McGaughey, 2012). An alternative set of approaches favor in-depth interviews or focus group with researchers such as those conducted through the Data Curation Profiles Toolkit (datacurationprofiles.org; Carlson, 2012; Cragin et al., 2010; Wright et al., 2013; Zilinski & Lorenz, 2012) or through similar methods such as focus groups (McClure et al., 2014) or other interview protocols (e.g. Peters & Dryden, 2011). A third option, taken by some in the literature, is to combine surveys and interviews to leverage the respective breadth and depth provided by these methods (e.g. Anderson et al., 2007; Beagrie et al., 2009).

Many of the information gathering methods described here have been used not only to gather information about the current state of research data management, but also to identify

researcher preferences and needs for new services to support data management. McClure et al. (2014) and Steinhart et al. (2012) asked researchers in focus groups about the types of services they would be interested in engaging with. Others, such as Wright et al. 2013, asked researchers to prioritize a list of potential services. Last, Zilinski and Lorenz (2012) use researcher interviews to help drive the selection of a new digital asset management system based on researcher data practices and needs.

METHODS

In this project we use a mixed method to collect information from researchers at CMU – a broad-scale survey of faculty coupled with a set of in-depth data interviews. Surveys of this type, and certainly in our specific case, tend to allow for many respondents to answer a few questions, resulting in broad coverage of responses but a lack of depth. Interviews, on the other hand, can offer a depth of response not typically available using surveys, though it is very difficult to achieve broad coverage of respondents. We chose to combine these two methods due to the successes and benefits we'd seen in the literature (see above) of using the methods individually and together. Surveys and interviews were conducted in spring and summer of 2014.

Faculty Survey

Our survey instrument was based, with permission, on the instrument used by Parham et al. (2012) to survey faculty at Georgia Institute of Technology. We modified the survey to be specific to the local conditions at CMU (e.g. data storage options) and to address specific questions posed by the DMSG and administration (Van Tuyl & Michalek, 2014). The survey included questions about: awareness of changes in research data management requirements in the federal regulatory environment, the amount of data produced and shared over the course of a "typical project," and general data management practices such as the use of data management plans, storage, and backup practices. For each question, respondents were asked to give a multiple choice or numeric response, but also had the option to provide free-text comments. The survey was administered using the online platform SurveyMonkey through which each subject received a unique invitation link allowing only one submission per subject.

Faculty Interview

The second element of our data collection was in-depth interviews with faculty members about both their current data practices and their opinions on elements of research data management including data creation, management, and sharing. The interview was divided into three major elements:

1. A general overview of the faculty member's research program, encompassing all of the research they conduct and oversee.
2. An in-depth discussion with the faculty member about a specific research project, including the types of data created and the data management practices in place for the project.
3. A rating by the faculty member of the usefulness of a variety of potential research data management services under consideration for implementation at CMU.

Our interview protocol was adapted from the Data Curation Profiles Toolkit (DCP Toolkit, datacurationprofiles.org) but modified to focus at the project level rather than the dataset level. We aimed for the interviews to be fairly broad in order to capture faculty opinions and current data practices beyond a single project. Interviews were conducted using a semi-structured interview format.

At each interview (with the exclusion of one), at least two librarians, including the liaison librarian for the interviewee's academic department, were present to ask questions, seek clarification, and record responses. Interviews were also audio recorded with the permission of the interviewees. Following the interview, the librarians involved created a summary document profiling the researcher's data management practices and attitudes towards potential services. A draft of this write-up was circulated to the researcher to offer an opportunity for clarification and correction as needed. Analysis of the information collected from interviews was informal and observational.

Faculty Sample

We worked with the CMU Office of Institutional Research to create a random sample of 500 faculty members at CMU who we contacted to participate in our survey. Our sample included faculty members of all types (research, teaching, and tenure-track) but was restricted to faculty on the Pittsburgh campus. We excluded faculty members from non-Pittsburgh CMU campuses (Qatar and Silicon Valley) and faculty from the Software Engineering Institute (a CMU-affiliated, DOD-sponsored, Federally Funded Research and Development Center) because they represent a different data management environment due to location or administrative structure. Faculty for the data interviews were recruited to represent a broad spectrum of research domains and project types and were selected based on personal knowledge of projects by members of the DMSG or a liaison librarian. Eleven faculty members completed data interviews. Due to the anonymous nature of the survey implementation, we do not have data to indicate if interview participants were also contacted as part of our survey sample or if they completed the survey.

University Project Data

Using CMU financial reporting systems we identified projects, from any funding source, that had been initiated in the preceding three years (May 2011-May 2014). These data were used to estimate the total number of recent or active projects at the university and the distribution of the number projects by award manager (i.e. the researcher or principal investigator). We included financial data only from academic units that were represented in our survey and interview samples. Given that many faculty have a significant number of additional, unfunded projects active at any time, the project data extracted from our financial database almost certainly underestimates the total number of projects active at the university.

Ethics Statement

The CMU Institutional Review Board approved the survey and interview protocol and methods. Survey and interview participants were made aware of the nature of the research and assented to participate using a click-through form for the survey and a signed consent form for the interviews.

RESULTS

Survey Results

Demographics. Our survey sample included 500 faculty representing about 37% of the faculty population at the university, but only 491 subjects received the survey due to incorrect email addresses or opting-out of all surveys. Of respondents, 104 (21%) answered part of the survey the survey and 69 (14%) answered all of the survey questions. In the results, we consider all responses for each question, including responses from both partially and fully completed surveys. Survey responses are assumed to be representative of the sampled population given the randomized nature of the survey sample. It was not possible to quantify self-selection bias in the responses, and survey results are presented with this caveat.

The Humanities and Social Sciences returned the most survey responses (26%), followed with roughly equal distribution of responses from the Sciences (19%) and Engineering (18%). Schools of Business and Computer Science both returned about 12% of the survey results, followed by much lower returns from Fine Arts and Public Policy (Table 1, following page).

Awareness of regulatory environment. In the survey we probed faculty awareness of the regulatory environment around research data management—specifically related to the data management plans (hereafter DMPs) and data sharing mandated by NSF, NIH, and

College	% of Responses
Humanities & Soc. Sci.	26%
Sciences	20%
Engineering	18%
Business	13%
Computer Science	13%
Fine Arts	5%
Public Policy	5%

Table 1. Affiliation of survey respondents across the Schools/Colleges at CMU (n = 86)

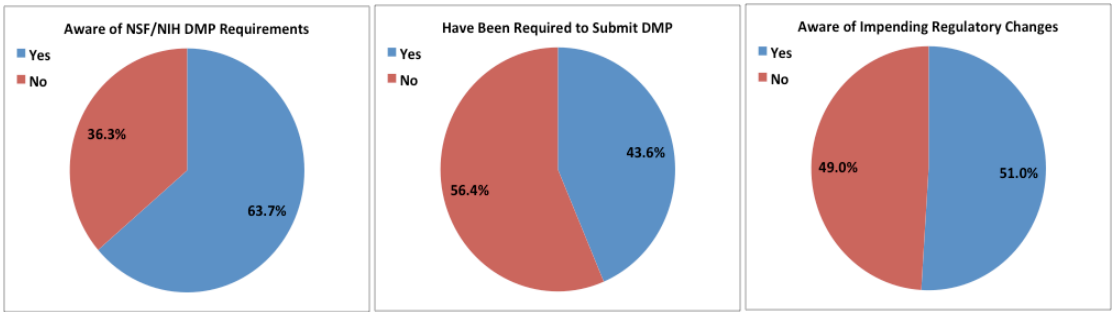

Figure 1. Awareness of current and impending regulatory frameworks for research data management in the US

other agencies. Across all colleges, 64% of respondents are aware of current requirements by funding agencies, but fewer respondents (51%) are aware of impending changes to the regulatory environment related to the 2013 OSTP memo. Remarkably only 44% of respondents have been required to submit a DMP as part of a grant (Figure 1).

At the college level, patterns of awareness of current requirements are uniformly high across all colleges, excluding the College of Fine Arts where there is little federal funding (Figure 2, following page). There is high variability across colleges in experience submitting DMPs as part of a grant. Responses from Engineering and Computer Science indicate that grant

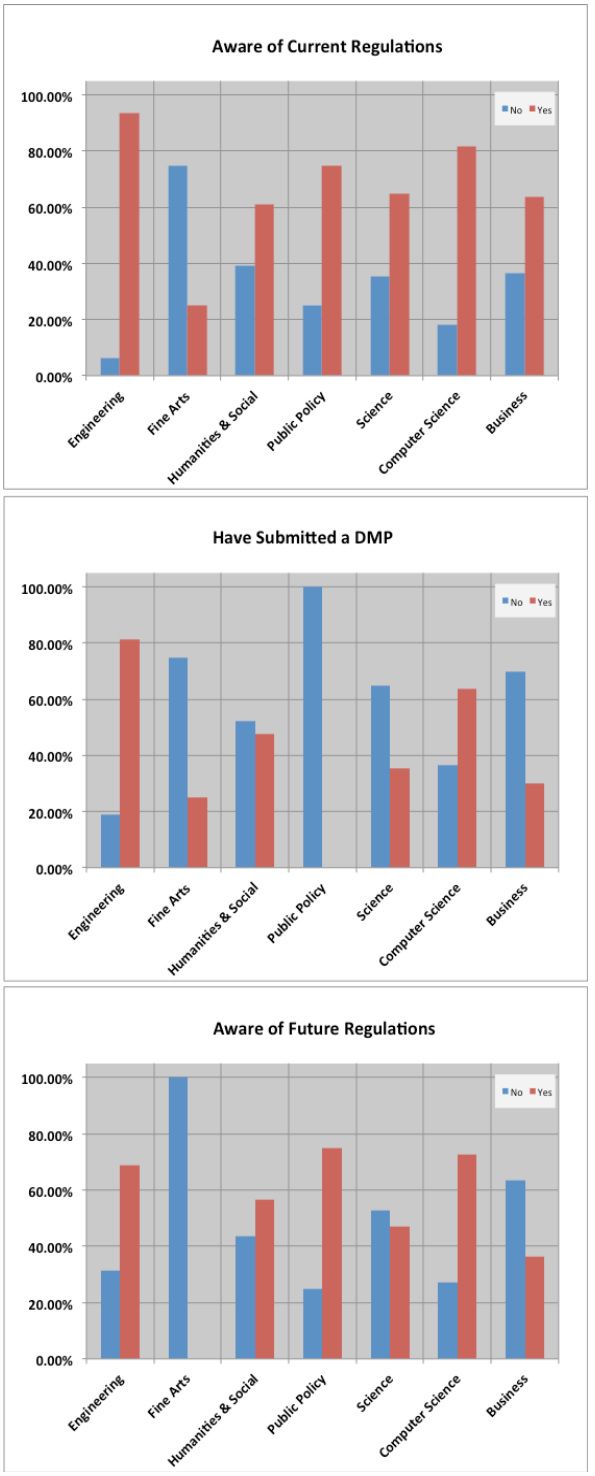

Figure 2. Awareness of current and impending regulatory frameworks for research data management in the US across colleges

applications from these colleges are strongly affected by DMP mandates. Conversely, DMP engagement from other colleges is somewhat limited, which is not surprising for some colleges (e.g. Fine Arts, Business) but is quite surprising for the sciences (e.g., mathematics, physics, biology, chemistry). Awareness of future regulations mostly matches the degree to which discipline-specific grant applications require DMPs—high awareness in Engineering and Computer Science, low awareness in Fine Arts and Business.

When queried about DMPs more broadly, fewer than 25% of researchers at CMU report having DMPs for all (7%) or most (15%) projects (regardless of funding source or mandates) while 27% and 39% report having DMPs for some or no projects, respectively.

Data production, use, and sharing. At the project-level, researchers at CMU produce varying amounts of research data ranging from just a few MBs to many TBs (Figure 3). We asked faculty to reflect on data for a "typical project," which we acknowledge leaves some ambiguity for how a faculty might have thought about a project, anywhere from an entire grant to a single experiment, or how they constructed an average if project size varies greatly even within their research program. That said, nearly 40% of projects at CMU produce less than 1 GB of data, and 18% of projects produce less than 50 MB of data. The fraction of projects producing between 1 GB and 10 TB is fairly stable at between 12% and 15%, and about 5% of projects generate more than 10 TB of data. Estimates of the amount of data shared outside of a project's immediate collaborators are, as expected, lower than the total amount of data produced by the project. For nearly 65% of projects, researchers report sharing less than 1 GB of data, and very few (< 5%) research projects currently share more than 1 TB of data (Figure 4, following page).

Faculty indicated that they use many types of files for their data, with the five most common formats being: data tables (e.g. Microsoft Excel, comma delimited text), documents (Microsoft Word, PDF), code (e.g. python, R, MATLAB), text files, and image formats (e.g. JPG, TIFF; Figure 5, page 12). Across colleges, uptake of various file formats varies slightly, though major format categories show some uniformity from college to college (Figure 6, page 12).

Data management. The major modes of storage for research data at CMU are desktop/laptop computers, external hard drives, cloud storage, and IT infrastructure maintained by the research group or the department (Figure 7, page 13). Data backup is primarily stored by three means—department-level IT servers, external hard drives, and cloud storage. There is relatively little use of other IT infrastructure offered at the university (e.g. college-level IT, central computing services) or external to the university (e.g. external/domain repositories) for backup. Slightly fewer than 5% of respondents indicated that they "Don't Back Up Data" or "Don't Know" how their data are backed up (Figure 8, page 13).

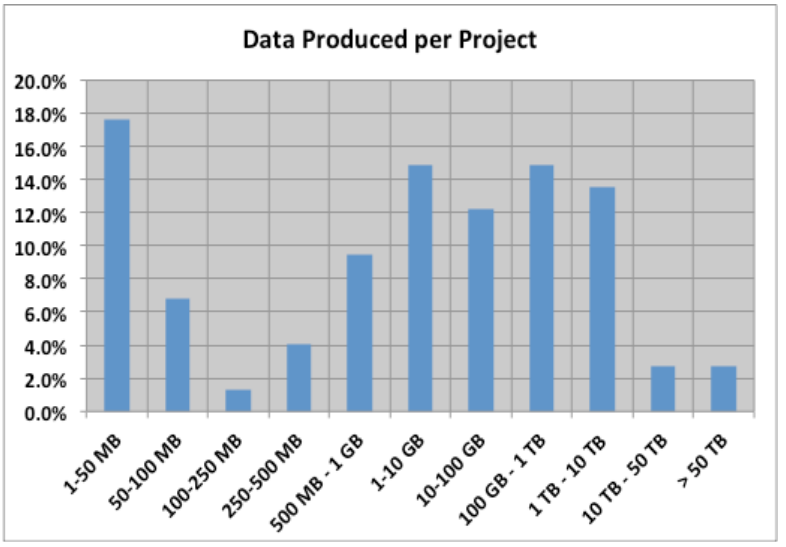

Figure 3. Frequency distribution of the amount of data produced per project. Project counts come from sponsored research reports produced by CMU and most likely under represent the number of projects due to undocumented unfunded research.

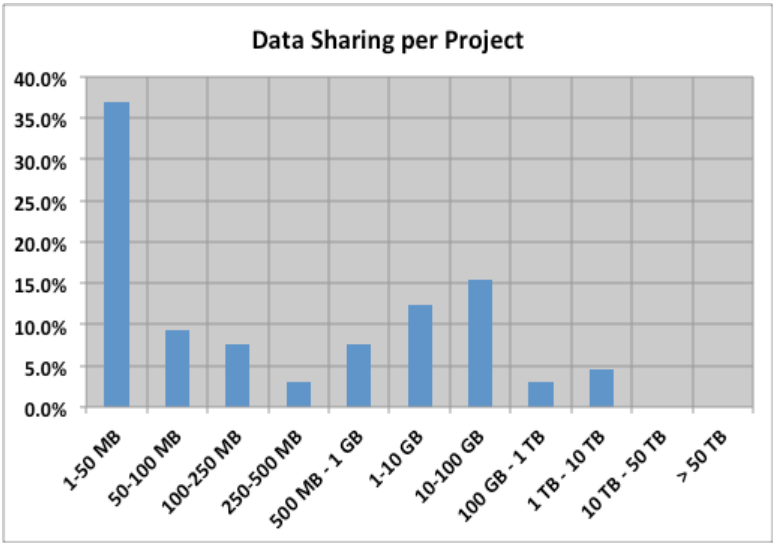

Figure 4. Frequency distribution of the amount of data shared per project. Project counts come from sponsored research reports produced by CMU and most likely under represent the number of projects due to undocumented unfunded research.

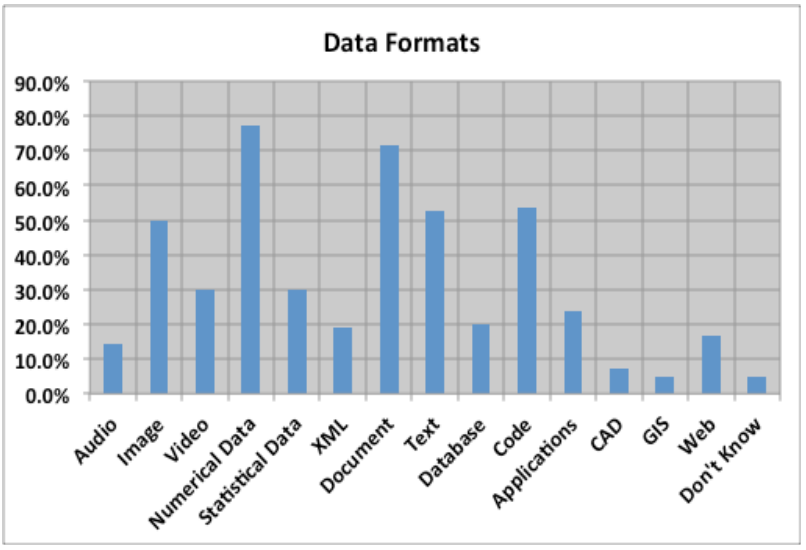

Figure 5. Data formats produced and used by researchers at CMU

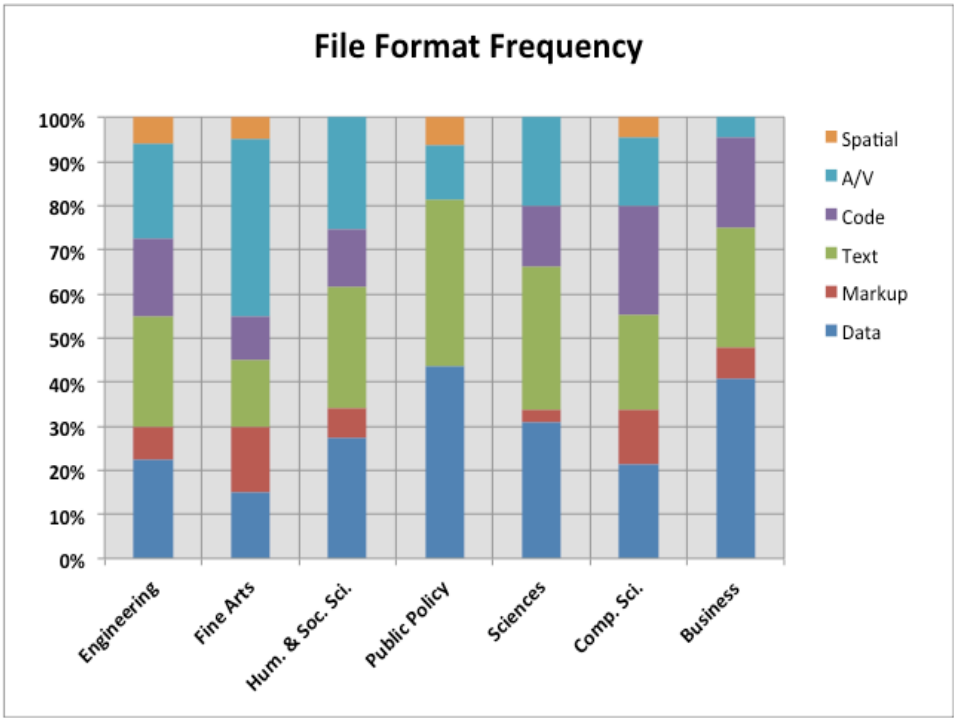

Figure 6. Most common categories of file format, by domain. Definitions of format categories are: Spatial—GIS, CAD; A/V—audio, video, image; Code—computer code, executable files; Text—documents and human-readable text; Markup—machine readable text such as HTML or XML; Data—numerical and statistical datasets such as Excel files, comma separated text, SPSS, R, Matlab, etc.

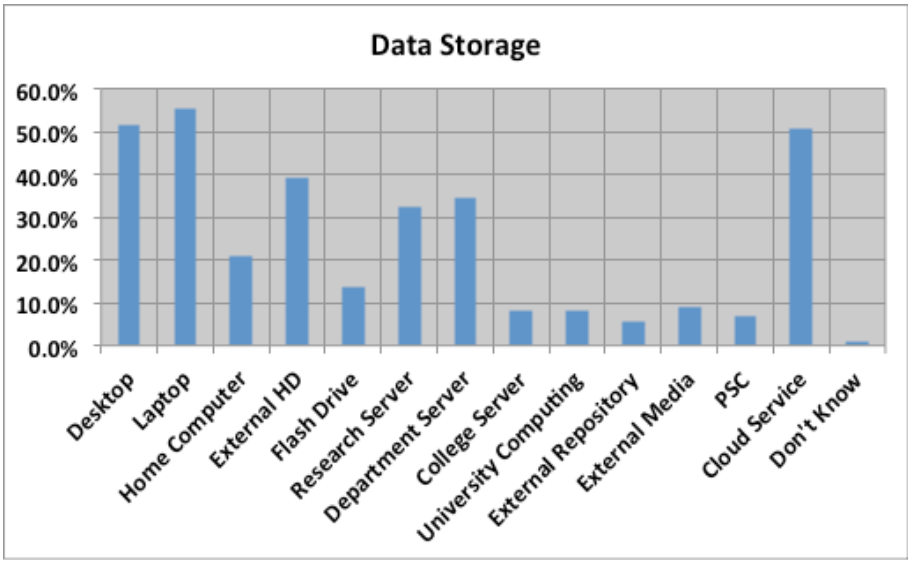

Figure 7. Uptake of major data storage media for research projects (PSC = Pittsburgh Supercomputing Center)

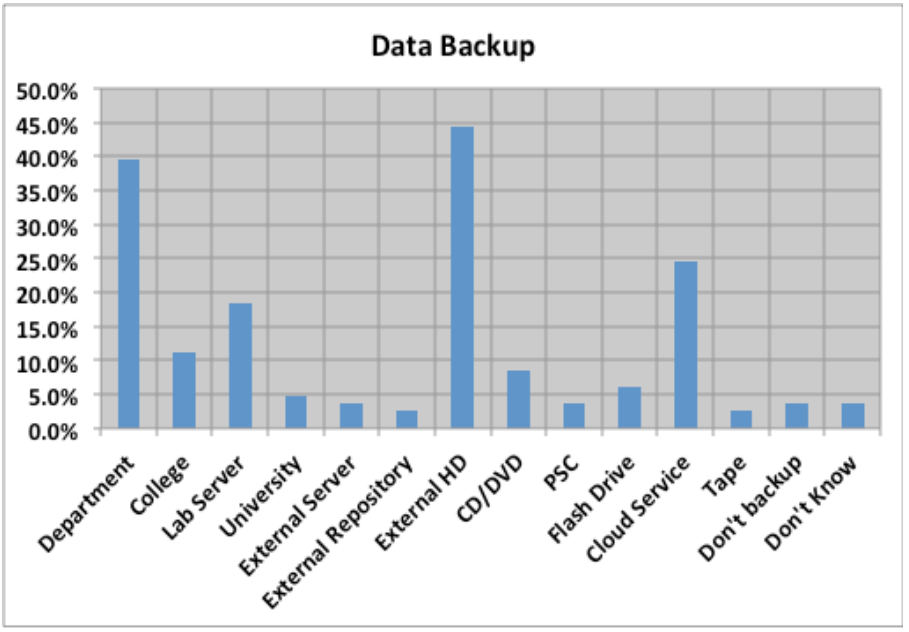

Figure 8. Use of various data backup solutions for research data at CMU

Interest in data services. Survey respondents' interest in potential research data management services varied considerably between different types of services (Table 2). The top services of interest were: help creating data management plans for grant proposals, services for long-term preservation and access to datasets, in-depth data management planning for the lifecycle of the project (i.e. operational data management planning), and providing guidance for identifying appropriate discipline-specific repositories for research data. In addition to the services listed in the survey, one respondent indicated interest in assistance estimating the costs of storing and preserving large datasets for planning and in particular to itemize in grant budgets.

University Project Data Results

Using the data we collected from the university financial reporting system we looked at how many grant-funded projects (i.e. a research proposal supported by a specific award such as a NIH RO1 award) researchers managed at once. Approximately 40% of researchers with grant-funded projects have funding for a single project, and about 80% of researchers

Service Category	Response Percent
DMPs - For Grant Proposals	73%
Preservation of Data	59%
DMPs - Operational	51%
Domain Repositories	45%
Intellectual Property	39%
Privacy and Confidentiality	38%
Documentation and Metadata	35%
Hosting Applications	32%
Impact Metrics	32%
Access Controls and Embargoing	28%
Connectivity and Linking	25%
Data Formats	16%

Table 2. Respondent preference for potential research data management services

have four or fewer funded projects. The median number of grant-funded projects at the university is 2, and the mean is 4.2 (standard deviation: 8.7). These numbers almost certainly underestimate the total number of research projects at the university since this tally does not include unfunded projects or projects with funding streams not considered "sponsored research" in the financial reporting system. Nonetheless, these project counts paint a picture for us of what project management is like for a researcher at any given time.

Estimating Data Needs

We next wanted to combine the estimates of data production from the faculty survey with our estimates of how many projects were ongoing during the previous three-year period to get a snapshot of the total amount of data being produced at the university over that time. By taking the response rates for each size range of data production on the survey (data created or shared per project) from our sample (n=86), we generalized this distribution across the total number of faculty (1,327) to estimate the total amount of data produced of each size on campus (Tables 3 and 4, following pages). Multiplying the total amount of data production or sharing per project by the median number of projects per researcher yields a rough estimate of the total amount of data produced and shared at the university during this time. The results of these calculations show that while the largest number of projects are producing datasets smaller than 1 TB (as reported earlier in our results) the largest amount of data are being generated by a few large data producers. Of the approximately 9,700 TB of data estimated generated for the reporting time period, only 230 TB are generated from the 80% of the projects producing less than 1 TB of data. The remaining 9,470 TB are generated by the remaining 20% of large data producing projects. The picture is similar for data sharing, with an estimated 670 TB of data shared by the largest 5% of data producers, while the remaining 95% of the projects share the remaining 70 TB of data.

Interview Results

The faculty we interviewed work in many different research domains: English, music theory, biology, psychology, chemical engineering, human-computer interaction, physics, public policy, environmental engineering, materials science, and language technologies. During the interviews we discussed projects that varied in size—both in terms of personnel and funding—and funding source (federal, private, unfunded), and discussed general research data management concerns.

Data management plans. Researcher use of DMPs, of any type, for research projects was quite limited. Formal data management plans were largely used by researchers who were mandated by funding agencies (e.g. NSF) to do so. These researchers universally had a sense

Estimates of Operational Data Currently Funded by Grants			
Amount of Data Per Project	**Percent Response**	**Faculty Count**	**Total Data (TB)**
1-50 MB	18%	233	0.006
50-100 MB	7%	90	0.007
100-250 MB	1%	18	0.003
250-500 MB	4%	54	0.020
500 MB - 1 GB	10%	126	0.094
1-10 GB	15%	197	1.085
10-100 GB	12%	161	8.877
100 GB - 1 TB	15%	197	108.491
1 TB - 10 TB	14%	179	986.284
10 TB - 50 TB	3%	36	1075.946
> 50 TB	3%	36	2689.865
		Total (TB)	**4870.7**
		Median # of Projects	**2**
		Grand Total (TB)	**9741.4**

Table 3. Snapshot estimate of total grant-funded data production at the university from 2011-2014

Estimates of Shared Data Currently Funded by Grants			
Amount of Data Per Project	**Percent Response**	**Faculty Count**	**Total Data (TB)**
1-50 MB	37%	490	0.012
50-100 MB	9%	122	0.009
100-250 MB	8%	102	0.018
250-500 MB	3%	41	0.015
500 MB - 1 GB	8%	102	0.077
1-10 GB	12%	163	0.898
10-100 GB	15%	204	11.228
100 GB - 1 TB	3%	41	22.457
1 TB - 10 TB	5%	61	336.854
10 TB - 50 TB	0%	0	0.000
> 50 TB	0%	0	0.000
		Total (TB)	**371.6**
		Median # of Projects	**2**
		Grand Total (TB)	**743.1**

Table 4. Snapshot estimate of total grant-funded data shared at the university from 2011-2014

of exasperation with these plans, expressing frustration that the plans were not meaningful or that they provided only a vague framework for data management.

Informal data plans or standards were more common and were set at the laboratory or project level. These informal plans typically involved indoctrinating new graduate students or post-docs to the local culture upon their matriculation into the lab/group, but typically did not provide fixed guidelines for data management; rather, they set general expectations that data should be properly managed.

Those researchers with projects mandating DMPs also expressed uncertainty about how seriously the DMPs are considered in the grant review process, and at least one interviewee expressed that, as a reviewer for a funding agency requiring DMPs, he does not consider them an important element of the grant proposal. Only some interviewees who had research funding that required writing a DMP were interested in receiving assistance with developing more meaningful research data management workflows and practices for the project or practices based on their mandated DMPs.

Data backup and storage. Data backup and storage practices largely followed the patterns found in the survey results, with most researchers storing and backing up data on local computers and external hard drives, with departmental IT units, or in cloud services. When questioned about whether they've interacted with IT units around data storage and backup (either at the department, college, or university level), those who had not done so indicated that they assumed that the institutional IT offerings would be too expensive, unreliable, or burdensome to use. In one specific case, a faculty member suggested that in comparing the stability of the more recent university infrastructure with a cloud service infrastructure like Amazon the university infrastructure might more likely result is data loss.

Metadata. Interviewees expressed a wide range of methods used for data documentation and metadata including extremely informal arrangements (e.g. a graduate student has complete control over all data documentation decisions) to strongly centralized arrangements (e.g. formal documentation or metadata standards, common rules for filenaming and organization). A common theme that emerged from the interviews is that as the number of individuals involved with a research project increases, the more likely the project is to have a formalized set of documentation standards.

Common themes when discussing metadata and documentation included:

- not knowing enough about metadata standards to create metadata
- the effort required to properly document datasets was disproportionate to the return on investment

- current documentation (simple readme files, etc.) was sufficient
- documentation in program code was sufficient for documenting output (results of the code)
- data was simple and thus did not require metadata
- data was 'self describing in some way' and thus did not require metadata
- data would not be usable even if it was well documented or the data creator would need to be involved in data re-use due to data complexity, so why create metadata

Data sharing practices. Data sharing practices and expectations varied widely across interviews, and it is difficult to describe a pattern across the faculty we spoke to. Generally, the largest amount of research data is shared within a research group or between project collaborators but less data is shared with those not involved with a project. Sharing within a research group is relatively common (though not always done), sharing outside of the research group is less common, and sharing with researchers outside of the specialized domain of research is even less common. Researchers showed mixed interest in sharing outside of their research groups and research domains.

When attempting to define policies for data sharing for research projects, interviewees commonly sought policy direction from their colleagues both within their domains of research and from those outside their domain who are perceived as leaders in data sharing. For those seeking guidance on setting data sharing policy for their projects or research groups, questions about which data to share (raw data? processed data?), embargo periods for data sharing, and terms of use (e.g. should the data creator be given authorship on resultant publications?) were quite common.

In a few cases, interviewees who had a strong theoretical understanding of why data sharing is important and why the regulatory landscape has changed still had a hard time understanding how data sharing could be functionally applied to their own research. Among these interviewees, the sense that the amount of overhead required to effectively share data with others (e.g., data formatting, metadata creation, versioning) was too high for the potential returns to the researcher. In some of these cases, interviewees also indicated that, since sharing was not currently important for their field of research, it would not be important in the future and that it would be hard to justify spending time and resources on.

Restrictions on sharing. More than half of the researchers interviewed indicated that they had some concerns about data sharing due to a need to publish on data before sharing, privacy, or ownership. In almost all of these cases, interviewees indicated that they have had little interaction with administrative units at the university who could provide advice

on resolving these questions (e.g., Office of General Counsel, Office of Research Integrity and Compliance). Cases where these administrative units were necessary for resolving an issue (human subjects research, contracts) were always referred to administrative units, but when more nuanced issues arose, the interviewees tended to try to resolve the concerns themselves. Some examples, given by interviewees, of this include interviewees assessing the point at which sufficient modifications to a proprietary dataset (purchased from an external source) allowed them to share the data with others or determining whether data had been sufficiently anonymized to protect personal information.

Services. The final element of the data interview asked researchers to rate the importance, for their research, of a variety of research data management services that CMU might provide. This list of services mirrored the list of services in the faculty survey. Services that interviewees most frequently ranked highest priority were assistance with creating data management plans for grant proposals, providing a data preservation platform, assistance with making datasets citable in other academic works, and providing metrics for use of research data (e.g. downloads, citations; Table 5). When both high and medium priority services are

Service	High + Med	High	Medium
Preservation	10	8	2
DMPs for Proposals	10	6	4
Credit/Citation	8	6	2
Metrics	7	6	1
Comprehensive DMPs	9	4	5
Discovery at CMU	4	4	0
Linking to/from data	6	4	2
Metadata	7	3	4
Discovery Broadly	6	3	3
Access Controls	6	3	3
Embargoing	7	3	4
Web Hosting	4	3	1

Table 5. Ranking of the importance of potential research data management services. Values represent the number of interviewees who ranked each service as High or Medium priority for their research. High + Med column is the sum of the High and Medium columns. Table is ranked on the High priority column.

considered, comprehensive data management planning, assistance with metadata and data documentation, and a need for embargoing datasets also emerge as needed services.

One element of service that wasn't originally included in our list of potential services presented in both the survey instrument and interviews, but that repeatedly emerged as an issue in faculty interviews, was a need to host web applications including those for data visualization or exploration. In many cases (approximately a third of interviews), researchers felt that, while providing access to their research data would be valuable to other researchers, providing methods or interfaces that allowed others to interact with their datasets was even more valuable. In all of these cases, the interviewee was concerned that access to the entire dataset might be overwhelming or not meaningful to a non-expert, especially with regards to the particulars of the data structures or formats, or analysis methods used for the data. In these cases, researchers have commonly built interfaces for interacting with the data, such as a tool to understanding the relationships among data points without actually manipulating the raw data yourself (e.g. social networks) or a tool to explore a meaningful subset of the entire dataset (e.g. vehicle safety inspection records for Allegheny County, PA). Simply sharing the data that reside in the background of these interactive tools would, the interviewees asserted, produce a burden on the researcher acquiring the data. Interviewees, in these cases, inquired about the university's ability to preserve these tools and applications in conjunction with the datasets that drive them.

DISCUSSION

Survey Results

The results of the faculty survey show us the broad patterns of the research data management at our institution. Many researchers have a moderate understanding that the regulatory landscape is changing to require better data management planning and to encourage data sharing and reuse, and researchers are interested in assistance for some aspects of these activities. We observed that implementation of formal data management planning is strongly associated with regulatory requirements. It is still unclear how helpful these requirements will be for encouraging researchers to engage in meaningful data management behaviors beyond the basics required for compliance for grant applications or the degree to which formal data management planning will be adopted for projects without a required DMP.

Current data management practices suggest that many researchers at CMU are managing certain elements of the research data lifecycle well (e.g. a substantial number of survey respondents indicating use of institutional backup solutions), but that institutional support might augment these existing best practices. However, many survey respondents exhibit

questionable data management practices such as poor data backup, inadequate metadata and data documentation, and a lack of overall data management planning for projects. These patterns are not unique to CMU and have been noted many times in similar studies in the US and the UK (Editorial, 2011; Peters & Dryden 2011; Scaramozzino et al., 2012; Steinhart et al., 2012).

Looking at data backup as an example, we see that some of the most frequent data backup solutions reported in the survey align with those recommended for guarding against loss (e.g. backups maintained by departmental IT), though other solutions with high reporting frequency should be discouraged (e.g. external hard drives). Of major concern is the small subset of researchers reporting that they "Don't Know" how their data is backed up, "Do Not" backup their research data at all, or use only a non-optimal backup system (e.g. flash drives). The complexity of responses around data backup indicates that there is an opportunity for targeted education and outreach in this area.

The variety of data formats revealed by the survey results is not surprising, given our previous knowledge of the research communities at CMU. Many of the format types identified in the results may be considered low-difficulty formats for preservation purposes (e.g. text, spreadsheets, documents) since there are clear transformation or migration pathways for these formats, and the University Libraries has extensive experience working with these types of data. On the other hand, two common data categories, executable software applications and computer code, might pose more difficult to preserve, especially when considering the potential for obsolescence of formats and software applications for using these files and the need for an emulation strategy to preserve a functional environment to run them in the future.

The survey results suggest that a nuanced and multi-tiered approach to education and outreach will be necessary for improving research data management at the university. There are many different needs and levels of experience between disciplines and researchers such that one size fits all approach will not be sufficient. We also note that faculty who are managing their research data well may be able to work with data services providers to create resources, including best practices and case studies that can be shared with others.

The primary limitation of the survey results is that, while the response rate was good, it is uncertain how well the results would generalize to the complete research population at CMU. The survey was offered only to a partial sample (approximately 35%) of the university faculty, and response bias may have affected the results, perhaps skewing responses towards more traditionally data rich fields or those with funder data mandates. These sampling concerns should inform future iterations of this survey, but they should be considered caveats rather than flaws in the current results. The response rate we achieved is relatively high for this

type of measure at our institution (Janel Sutkus, Office of Institutional Research, personal communication) and is similar to (Anderson et al., 2007, with ~16%) or higher than (Akers & Doty, 2013, with ~8%; Steinhart et al., 2012, with ~5%) the response rates observed at other universities conducting similar surveys.

Interview Results

There are some salient elements that emerge more deeply from the interview data than from the survey data. First, many interviewees expressed a broad theoretical understanding of, and agreement with, the reasons that data curation and data sharing are important for the research community. These same interviewees, however, often indicated that from a practical standpoint these data sharing and data management initiatives seem arbitrary and ineffective. Faculty also express skepticism about the degree to which new funder mandates for data curation and data sharing (e.g. requirements from NIH and NSF) will truly affect proposal success. At this time, most faculty, even those involved with grant-writing for agencies with data mandates, have not observed any effects of these mandates and thus their skepticism is understandable. We note however that these mandates are quite new and untested, and mostly have not had enough time to become an integral part of the funding cycle.

These same faculty are still deeply interested in research data management, are concerned about their own practices, and are interested in seeking assistance to augment or correct their practices. Many interviewees expressed concern over lack of time and resources for preparing data for sharing, a pattern seen elsewhere in the literature (e.g. McLure et al. 2014, Doty 2012, Akers & Doty 2013). Cragin et al. (2010) also found that researcher attitudes towards research data sharing were colored by the amount of time it would take to prepare data for sharing. Anderson et al. (2007) report that scientific communities are interested in solving these problems themselves, but lack the tools and resources to do so. This duality between skepticism of the regulatory framework for research data management and a strong sense of urgency to safeguard data is a key point in understanding how to approach research data on campus. It suggests strongly that services for research data management should focus on practical services to faculty with the goal of improving the research process, rather than complying with mandates.

As in the survey results we see that faculty express varying levels of interest in services for research data management and that interest is not easily correlated with research domain or school/college affiliation. For example, our two interviewees in the School of Computer Science describe using common data management patterns (though these still vary), but disagree over what types of services the university could usefully provide to researchers. Assistance creating comprehensive data management plans was ranked as a high priority for

one researcher (though only for portions of a project) but was not a priority for the other who stated, "there is no way [you could] do that without making it worse for [researchers]. "

Likewise, attempting to categorize faculty engagement across broad domains can also be misleading. Our interviews with faculty in the humanities (English, music) demonstrate significant differences in the types of research data management concerns that arise and the desire for assistance. Not surprisingly, though more difficult to quantify, is the finding that some seemingly divergent domains of research share strong similarities in research data management needs. Our interviews with faculty in biology and music theory show strong similarities due to the fact that the research methods shared a level of granularity, methodological approach, and analytical methodology. All of this suggests that it may be very difficult to generalize the types of assistance researchers at CMU will require or what services researchers be willing to participate in. As much as possible, it will be necessary for administrative units supporting research data services to address the needs of researchers on a case-by-case basis or with a suite of services that caters to researcher needs without presupposing those needs based on research domain. Cragin et al. (2010) also note that a broad, one size fits all approach to research data services may sacrifice the nuance required to help researchers most in need.

In discussions with faculty around the practicalities of research data management, we found that there were relatively few cases in which the faculty member themselves was the one directly responsible for managing research data (e.g. backups, storage, access controls, documentation etc.). This fits our anecdotal understanding of the research atmosphere in academia and at CMU and is a pattern we see elsewhere in the literature (cf. Peters and Dryden 2011, Scaramozzino et al. 2012). This seemingly obvious result, however, may help us better understand where education and outreach programs for research data management should be focused. In almost all of our data interviews, graduate students (and in some cases postdocs) were identified as the individuals who were actually responsible for day to day decisions around research data management. Our interview data suggests that strict data management guidelines and policies are only set by faculty when projects are very large. While we would not want to focus data management outreach and education solely on graduate students, it seems clear that effective graduate-level education in this area has the potential to have major impacts on research data management practices university-wide.

Lastly, it is of concern that for some faculty we interviewed, the administrative processes in place for dealing with legal, intellectual property, and privacy issues, were seen as burdensome—with at least four interviewees expressing concern over the burdens of research administration. It is unclear the extent to which assistance with these issues should be part of a suite of research data management services, but there does appear to be a

skepticism that these administrative functions will prove beneficial to the researcher, rather than purely burdensome. There may be a role for research data services providers to initiate and facilitate discussion, at a general level, about the importance and utility of interacting with these units of the university.

Data Production and Sharing Estimates

One of the most compelling, though not surprising, patterns in our estimates of data production and sharing by faculty at CMU is that the largest amount of data produced and shared is created by a relatively small number of projects. We estimate that roughly 35% of the total working research data is held by 5% of the projects. This highlights a central set of questions when considering providing research data management services to the university community: How do we allocate resources when demand exceeds supply? One solution could be to prioritize projects or researchers with outside funding, or even more specifically with funding from agencies with data mandates.

However, since we are still in the beginning of this initiative and the community is just becoming aware of the services we offer we could simply prioritize those who seek out our assistance first. In the long-term infrastructure there may be a tradeoff between trying to assist with as many projects as possible, many of which will have small quantities of data, versus trying to assist with the largest volume of data, which may come from a small subset of the total projects on campus.

Decisions about how to fund and direct service is never this black-and-white. However, to focus services on one extreme or another of the data production/sharing spectrum would be a mistake. In one case, the small but not insignificant number of projects producing very large datasets would be left without recourse, while in the other case, the vast majority of research projects would be similarly treated. It is entirely possible that the amount of data produced by a project is not a helpful indicator of what and to whom services should be rendered, and that prioritizing services to those asking for assistance would be a more logical approach. Prioritizing service on an "as needed" basis would make it difficult to predict the size of the infrastructure (specifically storage, but also other infrastructures) required to sustain the services.

Overall Impressions

When results of the survey, data inventory, and interviews are taken together, two findings of our study were surprising: the sheer volume of research data and researchers' admission that their IT operations are not fully professional. The volume of data and the relatively

poor quality of management appear to pose a significant risk to the institution as a whole. Perhaps the simplest way to address this would be to provide backup capabilities to researchers; however, even using Pittsburgh Super Computer's low-cost Data Supercell service, storage costs alone for this volume of data could cost approximately $2.5M/year. This does not include personnel costs (consulting, system setup, access management/other general support, etc), or any additional technical work that might be required (an example might be more finely-grained quota management than the PSC service provides).

Study Limitations

As mentioned previously, the survey data we have gathered most certainly contain biases. With respect to data production and sharing, though, these biases do not affect the outcomes equally. That is to say, biases that would put more researchers into large data production categories (> 1TB) would have much larger impacts on infrastructure needs than biases that put more researchers into small data categories (< 1GB). While it is impossible to know what biases exist and in which direction they skew our data, one might reasonably and conservatively suggest that there are, indeed, more large data producers than suggested by our results. We also note that our survey responses represent less than 10% of the faculty population at CMU and that our interviews represent an even smaller portion of the faculty. Boosting the survey response rate in future assessments may provide a clearer picture of the state of research data management at CMU.

A second limitation is the extent to which our estimates of the number of projects per faculty member can rightly be extrapolated across the survey results (as we have done). Our treatment of this data is almost certainly oversimplified, but more granular data are not available. That said, it would be helpful to better understand how the distribution of the number of projects per faculty member changes across classes of data production and how understanding these patterns might alter our estimates of data production and sharing. As with the potential for biases in the survey data, though, realistic impacts of changes to the number of projects per faculty distributions would be large only if those changes affected estimates of the number of very large data projects.

We find that the approach used in this project, a broad survey of researchers coupled with a sample of in-depth researcher interviews, proves to be effective at painting a complete picture of research data management practices and attitudes at the institutional level. Previous studies that have focused only on surveys (e.g., Doty, 2012; Parham et al. 2012) or on interviews (e.g. Carlson & Stowell-Bracke 2013; McLure et al. 2014; Peters & Dryden, 2011) seem to lack the depth (in the former cases) and breadth (in the latter cases) of information that can be captured by combining both methods. Others who have applied

this mixed model (e.g. Alexogiannopoulos, McKenney, & Pickton, 2010; Beagrie et al., 2009; Scaramozzino et al. 2012) also appear to have benefited from its application. That said, the in-depth interview element of this model is time intensive. Our approach to determining how many in-depth interviews to conduct was somewhat unsystematic, the results provide insight into some of the nuances of data practices at our institution that were not otherwise revealed by the survey.

It is clear that we may need better methods for estimating the amount of data produced and shared, the number of projects or some way to categorize data production into smaller than university wide estimates, data production by domain (and how to define domain in a useful way in this space, and resources by domain external to the university (e.g. domain repositories). The methods we use in this study for estimating the amount of data that might be shared at CMU are undoubtedly flawed, though it is unclear to us in what direction(s) the error lies. That said, it is clear, given the amount of data being produced at CMU that if even a small fraction of that data needed to be shared due to funder mandates, the amount of data to be shared could be very large. We stress that while our estimates may be biased, the methods we use and the resultant estimates provide a useful starting point for conversations about the potential quantity of data to be shared and the subsequent technical and service infrastructures required to support that sharing.

CONCLUSIONS

Overall, we find that there is still much work that needs to be done to help researchers improve data management practices and comply with funder mandates. Survey and interview results presented a preliminary picture of data production, management, use, and sharing practices at the university. This information is currently being used to guide the development of the infrastructure required to support data management services that researchers need.

This investigation revealed a need for expansion of data management services with an emphasis on assisting with data management plans, providing support for long-term data preservation, discoverability, citation and metrics of reuse, and improving general data hygiene (backup, storage, documentation, and preservation of data). The investigation also yielded an estimate of the volume of data being generated and shared across the university. Most projects produce very little data. A few produce enormous data sets. But the volume of data being actively shared is quite small. These findings have important implications for the design of platforms that must not only store and preserve data, but provide the functionality and benefits that will motivate researchers to use them. As we move forward,

what we learned about researcher activities and attitudes must inform the selection and design of data management services that will serve their needs for many years to come.

Based on the results of our investigation we have determined some next steps to expand and enhance the suite of services offered to researchers at Carnegie Mellon. University administration has a greater recognition of the need and commitment to support training and outreach to the primary constituencies supporting data management practices throughout the data lifecycle. Liaison librarians, working with the Data Services Librarian, have received intensive training in the form of in-house workshops, weekly meetings and support for data management conference attendance in order to hone their data management skills and to serve as first and second tier service providers. Researchers can now request a data consultation with a liaison librarian and receive advice and support for better data management practices. The University Libraries website now includes more substantive information concerning research data management resources. To facilitate the proper backup and storage of operational research data, the University Libraries negotiated an agreement with the Pittsburgh Supercomputer Center (PSC) for a PSC Data Supercell, which allows for the secure backup of data sets larger than 10TB. Future plans include the University Libraries licensing space from PSC to store smaller datasets. Last, information sharing with the university research offices (at the university and college levels) has increased, offering entry points into the research lifecycle via univeristy grant management infrastructures.

Three other important initiatives are underway in response to what we learned in our investigation. Given that our current infrastructure for data storage, preservation, and access is insufficient to handle the volume of data generated by Carnegie Mellon researchers, in February 2015 the Libraries began developing a Hydra repository to meet the university's long term needs for data curation, management, preservation, discovery, and access. To help motivate data sharing and enable data citation, the Libraries licensed EZID from the California Digital Library. EZID will support the creation of DOI's for datasets, enabling them to be managed with the same consideration as research publications. In March 2015 the Libraries launched a ORCID @ CMU, a university-wide project encouraging CMU researchers to use a locally developed web application to ensure that they have an ORCID ID linked to their authenticated Carnegie Mellon identity in the University's identity management system. The goal is for researchers to associate their ORCID ID with all their scholarly output—publications, datasets, code, peer reviews, scholarly blogs, etc.—to facilitate discovery, access, and curation of all their contributions in the future.

ACKNOWLEDGMENTS

We thank the members of the Data Management Services committee and liaison librarians at Carnegie Mellon University for help in designing and carrying out this work and two reviewers, the journal editor, and Ana Van Gulick for constructive comments on the manuscript.

REFERENCES

Akers, K. G., & Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Digital Curation*, 8(2), 5–26. http://dx.doi.org/10.2218/ijdc.v8i2.263

Alexogiannopoulos, E., McKenney, S., & Pickton, M. (2010). *Research data management project: A DAF investigation of research data management practices at The University of Northampton*. Northampton: University of Northampton. Retrieved from http://nectar.northampton.ac.uk/2736/

Anderson, N. R., Lee, E. S., Brockenbrough, J. S., Minie, M. E., Fuller, S., Brinkley, J., …Ornoch, P. E. T. A. (2007). Issues in biomedical research data management and analysis: Needs and barriers. *Journal of the American Medical Informatics Association*, *14*(4), 478–489. http://dx.doi.org/10.1197/jamia.M2114

Beagrie, N., Beagrie, R., & Rowlands, I. (2009). Research data preservation and access: The views of researchers. *Ariadne*, *60*. Retrieved from http://www.ariadne.ac.uk/issue60/beagrie-et-al

Carlson, J. (2012). Demystifying the data interview: Developing a foundation for reference librarians to talk with researchers about their data. *Reference Services Review*, *40*(1), 7–23. http://dx.doi.org/10.1108/00907321211203603

Carlson, J., & Stowell-Bracke, M. (2013). Data management and sharing from the perspective of graduate students: An examination of the culture and practice at the water quality field station. *portal: Libraries and the Academy*, *13*(4), 343–361. http://dx.doi.org/10.1353/pla.2013.0034

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A-Mathematical Physics and Engineering Science*, *368*(1926). http://dx.doi.org/10.1098/rsta.2010.0165

D'Ignazio, J., & Qin, J. (2008). Faculty data management practices: A campus-wide census of STEM departments. In *Proceedings of the American Society for Information Science and Technology*, *45*(1), 1-6. http://dx.doi.org/10.1002/meet.2008.14504503139

Doty, J. (2012). *Survey of faculty practices and perspectives on research data management*. Retrieved from http://guides.main.library.emory.edu/datamgmt/survey

Editorial. (2011). Challenges and opportunities. *Science*, *331*(February).

Henty, M., Weaver, B., Bradbury, S. J., & Porter, S. (2008). Investigating data management practices in Australian universities. *Australian Partnership for Sustainable Repositories*. Retrieved from http://eprints.qut.edu.au/14549/1/14549.pdf

Holdren, J. P. (2013). *Increasing access to the results of federally funded scientific research*. Washington, D.C.: Office of Science and Technology Policy. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

McCLure, M., Level, A. V., Cranston, C. L., Oehlerts, B., & Culbertson, M. (2014). Data curation: A study of researcher practices and needs. *portal: Libraries and the Academy*, *14*(2), 139–164. http://dx.doi.org/10.1353/pla.2014.0009

Parham, S. W., Bodnar, J., & Fuchs, S. (2012). Supporting tomorrow's research: Assessing faculty data curation needs at Georgia Tech. *College & Research Libraries News*, *73*(1), 10–13. Retrieved from http://crln.acrl.org/cgi/content/abstract/73/1/10

Peters, C., & Dryden, A. R. (2011). Assessing the academic library's role in campus-wide research data management: A first sep at the University of Houston. *Science & Technology Libraries*, *30*(4), 387–403. http://dx.doi.org/10.1080/0194262X.2011.626340

Scaramozzino, J. M., Ramirez, M. L., McGaughey, K. J., Ramírez, M. L., & McGaughey, K. J. (2012). A study of faculty data curation behaviors and attitudes at a teaching-centered university. *College & Research Libraries*, *73*(4), 349–365. http://dx.doi.org/10.5860/crl-255

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

Van Tuyl, S., Michalek, G. (2014). *2014 faculty data survey instrument* [survey instrument]. Retrieved from Carnegie Mellon University Digital Collections. http://dx.doi.org/10.1184/RDL/01092015

Whitmire, A. L., Boock, M., Sutton, S. C. (in press). Research data stewardship at Oregon State University: Findings and recommendations from a campus-wide survey.

Wright, S. J., Kozlowski, W. a., Dietrich, D., Khan, H. J., Steinhart, G. S., & McIntosh, L. (2013). Using data curation profiles to design the Datastar Dataset Registry. *D-Lib Magazine*, *19*(7/8), 1–14. http://dx.doi.org/10.1045/july2013-wright

Zilinski, L. D., & Lorenz, S. W. (2012). Using data profiles to select digital asset management systems (DAMS). *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–3. http://dx.doi.org/10.1002/meet.14504901306