12-2022

Research Data Management Policy & Organizational Compliance: An Exploratory Study in the Academic Context

Monica Inez Ihli
University of Tennessee, Knoxville, mihli1@utk.edu

Recommended Citation

Ihli, Monica Inez, "Research Data Management Policy & Organizational Compliance: An Exploratory Study in the Academic Context. " PhD diss., University of Tennessee, 2022.
https://trace.tennessee.edu/utk_graddiss/7726

To the Graduate Council:

I am submitting herewith a dissertation written by Monica Inez Ihli entitled "Research Data Management Policy & Organizational Compliance: An Exploratory Study in the Academic Context." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Communication and Information.

Suzanne Allard, Major Professor

We have read this dissertation and recommend its acceptance:

Carol Tenopir, Bradley Wade Bishop, Alex Bentley

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Research Data Management Policy & Organizational Compliance:

An Exploratory Study in the Academic Context

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Monica Ihli

December 2022

Abstract

Research data management (RDM) describes a broad array of processes and activities aimed at ensuring that data are documented, organized, findable, and preserved for future access. In January 2023, the National Institutes of Health will begin enforcing the strictest data management requirements of a U.S. federal agency to date, including potential consequences for organizations whose researchers fail to demonstrate compliance with commitments to data management and sharing. This dissertation makes two major assessment-based contributions in support of organizational preparedness for policy compliance. First, it reports the results of a pilot study at a high research institution for a survey instrument, which has been designed to extend current known RDM assessment practices. More specifically, the survey addresses new factors such as the sensitivity of the data, items that expose knowledge gaps about specific actionable tasks that comprise RDM, and items that reveal organizational communication challenges in terms of researcher uncertainty about institutional support. Additionally, a second pilot study is conducted that demonstrates the assessment for data availability messages tool (ADAM). ADAM has been developed as an analytical workflow and measurement system for extracting messages from publications about what researchers communicate in regards to data sharing. Finally, an open-source library of all data visualization and processing scripts needed to interpret the results has been provided. For a test institution, results from the two pilot studies identify current trends and knowledge gaps in RDM, which should be addressed to ensure the organization's ability to comply with relevant data policies.

Table of Contents

1. Introduction....................1

1.1. Problem Statement & Motivation....................6

1.2. Research Objective....................7

1.3. Dissertation Organization....................7

2. Review....................8

2.1. Data Policy....................8

2.1.1. Federal- and Agency-Level Policies....................8

2.1.2. Data Management Planning Requirements for Funders....................13

2.1.3. Commercial vs. Non-Commercial Funder Data Policies....................15

2.1.4. Journal and Publisher Level Policies....................16

2.1.5. A Note on Data Sharing Strategies....................17

2.2. Studies on Behaviors, Beliefs, and Needs....................20

2.3. Data Services....................23

2.4. Research Problem....................24

3. Data....................28

3.1. Research Questions....................29

3.2. Survey Instrument....................30

3.2.1. Survey Items....................31

3.2.2. Survey Data Preparation....................34

3.3. Researcher Publications....................34

3.4. Data & Code Availability Statement....................38

4. Methods..39

4.1. Methods of Analysis for Survey Data..39

4.2. Pilot Study for Publications-based Assessment of RDM.....................................43

4.2.1. Data Availability Messages...43

4.2.2. Publisher Support for Structured Data Messages...45

4.2.3. Data Citations...48

4.2.4. Regarding Secondary Data Analysis...50

4.2.5. Coding Process..51

5. Results..54

5.1. Characteristics of the Researchers..54

5.2. Characteristics of the Data...57

5.3. Understanding of RDM Tasks..61

5.4. RQ2: Researcher Characteristics and RDM Knowledge......................................66

5.5. RQ3: Other RDM Barriers...69

5.6. RQ4: Institutional Support...69

5.7. RQ5: Current Practices..69

5.8. Results of ADAM Workflow for Publications-based Assessment.......................72

6. Broader Impacts...78

6.1. Broader Impacts for Organizational Planning..78

6.2. Broader Impacts for Benchmarking Progress..84

6.3. Broader Impacts for Understanding Communication Patterns of Researchers..................87

7. Conclusions...90

7.1. Discussion...90

7.2. Limitations...94

7.3. Conclusions and Future Directions...96

List of References..100

Appendix..111

Vita..118

List of Tables

Table 3.1: Articles and proceedings by category in 2021 for pilot institution..............................37

Table 4.1: Contingency table, human subjects data status versus understanding of metadata.….40

Table 4.2: Examples of language used by journals to support data availability statements.........49

Table 5.1: p-values for researcher characteristics and understanding of RDM topics..................68

Table 6.1: Open source software files published for processing data...86

List of Figures

Figure 4.1: An example of a structured data availability message found in a publication..........46

Figure 4.2: A different example of a structured data availability message..................................46

Figure 4.3: Example of an unstructured data availability message...47

Figure 4.4: Assessment for Data Availability Messages (ADAM) workflow...............................53

Figure 5.1: Primary research domain reported by n=54 survey respondents...............................55

Figure 5.2: Amount of research which is funded, for n=54 survey respondents..........................55

Figure 5.3: Distribution of position with university, for n=54 respondents.................................56

Figure 5.4: Distribution of tenure status for n=43 faculty who responded to survey...................56

Figure 5.5: Respondents who have served has principle investigator of a recent project............58

Figure 5.6: Respondents who have served in role of making data management decisions..........58

Figure 5.7: Types of data and other digital artifacts created by n=54 survey respondents...........59

Figure 5.8: Per project digital storage requirements, on average, for n=54 respondents.............59

Figure 5.9: Frequency of working with different sensitivity levels of data, for n=54 responses.. 60

Figure 5.10: Percentage of n=54 respondents who report working with human subjects data.…62

Figure 5.11: Responses to the question "How well has this been explained to you?" for different RDM topics, as a measure of personal understanding. There are n=54 responses represented.…63

Figure 5.12: Responses to the question "How well has this been explained to you?" 65

Figure 5.13: Understanding of data management expectations of funding agencies...................67

Figure 5.14: Researcher perceptions of barriers to data policy compliance.................................70

Figure 5.15: Institutional support units for different RDM tasks...71

Figure 5.16: Researchers who engage in various data management practices..............................73

Figure 5.17: Distribution of unstructured versus structured data availability messages...............74

Figure 5.18: Distribution of assertions about data availability across 42 messages......................76

Figure 6.1: An example model of research data management services...82

1. Introduction

The journal has long been considered a primary channel of formal communication, serving as a mechanism of information dissemination and as a public record of scholarship. Over the years, the size of the average research article has grown in tandem with the development of increasingly sophisticated research methods that continue to produce ever greater amounts of data (Subramanyam, 1981). In some fields, the volumes of data produced have increased by entire orders of magnitude, contributing to what has come to be termed the “fourth paradigm” of science—characterized by the “techniques and technologies of data-intensive science” (Bell et al., 2009).

What counts as "research data" can vary based on context. For example, the U.S. National Institutes of Health define research data in the context of award funding as “The recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications". Specifically excluded are "laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens” (U.S. National Institutes of Health, 2020). The scholarly publisher, Elsevier, defines research data in saying that "research data refers to the results of observations or experiments that validate your research findings. These span a range of useful materials associated with your research project, including: Raw or processed data files, software, code, models, algorithms, protocols, and methods" (Elsevier, 2022). Another publisher, PLOS, chooses to focus more narrowly on what research data means in the context of publications. The PLOS data policy describes the "minimal dataset", defined as "the data required to replicate all study findings reported in the article, as well as related metadata and methods" (PLOS, 2022).

The drive to consider research data as not only a distinct scholarly output but also an object to be disseminated has represented a cultural shift. Some see widespread access to supporting data as a critical step toward accelerating and preserving the integrity of science. For example, a call for open data in the ecological sciences argued that the lack of accessibility to research data after publication constitutes a serious impediment to the advancement of science—especially in cases where the capture of data is tied to events in a time and place (e.g., the ecological impacts of an oil spill or other major events) (Reichman et al., 2011). Open data policies are argued to be critical to addressing the so-called “reproducibility crisis”, a phrase coined to describe the inability of many reported findings to be reproduced—a fundamental tenet of the scientific process (Baker & Penny, 2016). High-profile replication studies in such areas as clinical research (Begley & Ellis, 2012) and psychology (Open Science Collaboration, 2015) have been major drivers in the push for making data open. Similar studies in the physical sciences, such as an assessment of materials science experiments (Park et al., 2017), have reported comparable results.

Note that the phrase "open data" in this context refers to the expectation that research data, much like research findings, should be disseminated and available to the public. What it means to make data "open" is a nuanced topic, and disagreement arises as to how far this concept should go. Cost barriers, requesting or negotiation for access, and merit reviews are examples of how people may be prevented from accessing data, beyond only the simple question of whether research data should be disseminated. These nuances will be discussed in greater detail throughout the review of the literature.

Despite the strong case for normalizing shared access to the data that underlies published findings—meaning that the data are available to others without cost or other obstacles—adoption of open data practices has not been universally accepted. At the level of individual researchers, resistance has been grounded in such factors as technological obstacles, concerns surrounding the systems of reward, financial and personnel limitations, and disagreement or confusion about the ownership of the data (Berman, 2017; Melero & Navarro-Molina, 2020a; Tenopir et al., 2011). More recent work in the literature indicates an increasing willingness of researchers to engage in practices consistent with data sharing and reuse, but the extent of this increase can vary along disciplinary lines (Tenopir et al., 2020). Even where individual support for open data and data management practices exists, pragmatic difficulties remain that must be addressed (Bishop et al., 2020; Tenopir et al., 2020).

Distinct from internal drivers of change, such as personal values and motivations, external factors such as the funding agency and journal policies also influence changes in research data practices. A researcher might become engaged in policies regarding data at multiple levels. Funding agencies (especially taxpayer financed) may be required to expand access to research, support reproducibility, and avoid redundant work by ensuring that the outputs of research investments can be fully leveraged. Institutional data policies are often more concerned with the issues of who owns the data, rights of access, and stewardship. Journal and/or publisher data policies are generally interested in ensuring that the specific data needed to validate the scientific findings communicated in a publication should be available. However, there is wide variation in whether journals and publishers encourage data sharing, versus requiring it as a condition of publication that the data that support the findings also be made available (Briney et al., 2017).

The uptake of journal data policies has become a subject of interest in the Information Sciences literature. Various studies, which will be discussed in greater detail throughout the Literature Review, have sought to characterize the impact of such policies on researchers. Domain-specific investigations of journal and publisher data policy trends have been conducted for titles in numerous areas, such as Business (Dosch & Martindale, 2020), Phonetics (Garellek et al., 2020), Medicine (Huh, 2019), and Ecology (Sholler et al., 2019). Evidence suggests that, where data sharing occurs, citations are positively impacted (Christensen et al., 2019) and that journals with high impact factors are more likely to have a data sharing policy (Resnik et al., 2019).

At the funding level, a requirement for data management plans (DMPs) for grant proposals has become the primary policy vehicle for addressing research data. A 2003 Notice published by the U.S. National Institutes of Health (NIH) is widely considered the starting point for the modern DMP. This notice formalized the requirement of a data sharing plan, or explanation for the absence thereof, within the award-seeking process for certain categories of the largest proposals (U.S. National Institutes of Health, 2003). The U.S. National Science Foundation (NSF) followed suit in 2011 by requiring that every proposal for research funding be accompanied by plans for the management and sharing of data or at least a justification for withholding data (U.S. National Science Foundation, 2011). The requirement of data management planning in some form for proposals has since become widely adopted across government and private research funding sources, although funders can vary in terms of the exact language and requirements for planning (Williams et al., 2017).

At present, funder-level research data management policy is once again poised to transform at least one part of how research institutions engage in the business of science. Effective January 2023, the U.S. National Institutes of Health are revising the agency's data policies. Two of the most significant changes include that the new data policy broadens who is affected (everyone, regardless of funding mechanism) and the addition of a compliance-based component as a consideration for future funding. More specifically, the policy states that "After the end of the funding period, non-compliance with the DMS plan may be considered by the NIH for future funding decisions for the recipient institution." Failure to follow a DMP will henceforth be treated as a failure to comply with the Terms & Conditions of the Award, with all the potential consequences that accompany it (U.S. National Institutes of Health, 2020).

The implications of this change are extraordinary given that, at approximately 41.7 billion USD, the NIH has the largest R & D budget of any single agency in the country (U.S. National Science Foundation, 2022). Of particular importance is that the language of the policy specifically jeopardizes continued access to research funding for the institutions with whom offending researchers are affiliated. Numerous works in the literature have already explored what the nature of services supporting research data management could be, should be, or presently are (Cox et al., 2019; Koltay, 2017; Pinfield et al., 2014; Semeler et al., 2019; Tenopir et al., 2019). These upcoming policy changes are certain to stimulate current data service support efforts.

1.1. Problem Statement & Motivation

To summarize, organizations such as laboratories and academic research institutions continue to grapple with a rapidly changing environment regarding data management and defining the role of an institution in supporting RDM. Beyond a potential for direct benefit in terms of quantifiable impact metrics, the ability to conduct or discuss RDM activities is fast becoming a requisite for different stages of the research life cycle. Organizational motivations for RDM support are as follows:

- Institutions have an immediate interest in stewarding and protecting research data assets produced by their affiliated researchers, which may continue to have scientific or commercial value long after the life of the project that produced them.
- Publication is a job requirement for many research staffs. Researchers now face an environment in which many highly reputable journals have begun to require at least some degree of data sharing, stewardship, or communication as a condition of publication.
- Academic research institutions seek external funding to conduct much of their scientific and research missions. Data management planning has become a requisite component of the proposal stage for many of the biggest sources of external funding.
- DMP compliance is now staged to become a determining factor in continued access to external funding from the NIH, one of the biggest funding agencies in the United States.

The current literature is rich in studies that have explored the beliefs, attitudes, perceived needs, and current data practices of researchers. Several landmark studies in particular (Tenopir et al., 2011, 2020) have thus far been a key source of broad and generalizable insights into how researchers overall may feel, think, and act in regard to research data. However, there remains an absence of instruments or methodologies that institutions may use to assess their specific and internal levels of competencies for RDM activities. Given that policy compliance and RDM knowledge have significant potential to impact the security of external funding, choice of publication venue, or the protection of institutional data assets, research organizations would greatly benefit from the ability to internally assess organizational compliance with research data policies.

1.2. Research Objective

The purpose of this project was to develop approaches that research organizations could use to gauge the internal alignment of knowledge, skills, and practices with those activities commonly necessary for compliance with research data policies. Organizational readiness for research data management compliance is itself a sociotechnical challenge that involves people and technology infrastructure. However, the scope of assessment tools in this project focuses on the social component—understanding the relationship between knowledge gaps or environmental drivers and practices of individual researchers.

1.3. Dissertation Organization

Following the Introduction, Chapter 2 presents a literature review which concludes with an analysis of the gaps in the existing literature, as well as the requirements for translating existing work into tools for organizational assessment. Chapter 3 introduces two pilot studies, including a survey and the Assessment for Data Availability Messages tool for evaluation of publications, as well as the data collected for both. Chapter 4 reviews the methods of analysis, while Chapter 5 reports the results of both research approaches. Chapter 6 discusses the broader impacts of the research approaches and their findings. Chapter 7 offers the concluding remarks.

2. Review

2.1. Data Policy

This review begins with an overview of policy change drivers that have influenced the movement toward data management and sharing. It begins with an overview of major events at the federal level, followed by a discussion of how these events have trickled down to funding agencies. A review of data policies from the perspectives of journals and publishers is also included. For the most part, the surveyed literature draws upon secondary sources that have analyzed policies and their implications at scale, except for the National Institutes of Health (NIH) policies that are discussed in more detail as primary sources. The survey of RDM literature also lays the groundwork for the position that existing studies and institutional assessments have largely focused on the role that personal motivations and beliefs play in researcher data management choices, and this trend has continued in efforts to develop support services. Thus, the role of information and understanding in contributing to RDM outcomes remains understudied.

2.1.1. Federal- and Agency-Level Policies

A comparative review of research data policies for different world regions and nations is beyond the scope of the current project, which primarily focuses on the test case of high research academic institutions within the United States. Therefore, only the U.S. data policy background is discussed at length. The historical context for federal regulation of research data access through early 2013 is comprehensively summarized in a Congressional Report on the topic (Fischer, 2013). The essential facts are as follows: Driven by the argument that “Americans have a right to know how their tax dollars are spent and whether they are spent wisely, as well as the underlying scientific basis for many of our federal policies and rules,” (Shelby, 2000) a 1998 rider to the next fiscal year’s budgetary legislation (*P.L. 105-277, 112th Stat. 2681-495*, 1998) directed the Office of Management and Budget (OMB) to amend existing regulations for the administration of research grants. This amendment was to require all Federal awarding agencies to ensure that research data was publicly available, using the processes established in the Freedom of Information Act and with allowance for reasonable costs incurred.

The passing this legislation resulted in push-back from certain members of the scientific community. Letters and editorials voiced concerns such as a lack of a clear scope, disruption to the scientific process, and the potential for privacy concerns to have a stifling effect on research. For example, the American Political Science Association published a scathing opinion of the OMB's efforts as poorly thought out in stating that the draft efforts constituted a "serious threat to the integrity of federally funded academic research" (Rudder, 1999). Additionally, it was argued that FOIA requests were an improper vehicle through which to request data. After all, in many cases, the researchers themselves had to pass ethical reviews to collect data on human subjects, but FOIA requests required no such vetting (Hollingsworth, 1999; Lutz, 1999; Miller & Baldwin, 2001; Rudder, 1999; Walter & Richards, 2000).

In terms of implementation, the ambiguous nature of the legislation’s language left considerable room for discretion. After a period of public comment and review, the final amendment draft was intended to address many of these concerns by providing clear definitions of important concepts and clarifying the scope of content covered. The final product of the comment and review process is recorded in Section 36 of the amended circular (Office of

Management & Budget, 1999). Previously, no overall federal regulation existed for research data, and U.S. agencies had been left to determine their own policies (or lack thereof) (Fischer, 2013).

Certain details of the amendment were of particular importance in defining the future of data management and sharing at this point. For example, the federal government adopted the following definition of research data: “Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues.“ Explicitly excluded from this definition were physical samples, unpublished research, and the private or identifying information of personnel and participants. Moreover, the final draft significantly narrowed the scope. Whereas the original legislation ordered that agencies should “ensure that all data produced under an award will be made available to the public” (emphasis added), the language of the OMB’s amendment merely obligated agencies to honor FOIA requests pertaining to “data relating to published research findings produced under an award that were used by the Federal Government in developing an agency action that has the force and effect of law.” In other words, the public was only guaranteed access to research data by FOIA request if they had been used as the basis of policy.

While members of some research communities were opposed to changes in data sharing practices, others were vocal in their support. For example, the advent of human genome sequencing led to a pivotal moment in 1996, when genomics researchers called for the expedient and open dissemination of these data (Marshall, 2001). The American Association for the

Advancement of Science formally supported the dissemination of data in addition to published findings as early as 1999 (Fischer, 2013).

Several later events secured the importance of research data access at the Federal level. Critical among these would be the 2013 Executive Memo that expressed the executive branch's clear and direct support for increasing access to the results of federally funded research in the form of published literature and scientific data. Through this action, all Federal agencies with budgets over $100 million were expected to create a plan for facilitating access to scientific publications and research data (Holdren, 2013). An analysis of 19 federal agencies found that the comprehensiveness of data policies developed in response to this memo was varied (Kriesberg et al., 2017). Several years later, key aspects of data policies first laid out in the earlier executive memo were expanded upon and signed into law (Open, Public, Electronic and Necessary (OPEN) Government Data Act, 2019).

As previously observed, certain federal agencies have acted as leaders in terms of adopting policies to promote data stewardship and data sharing. The 2003 NIH Notice (U.S. National Institutes of Health, 2003) formalized the data management planning component of the external funding proposal process for the agency and began the practice of DMP requirements among other agencies. The key contributions of the notice included the definition of the timely release and the sharing of supporting data as, "no later than the acceptance for publication of the main findings from the final data set," and guidance was provided for the sharing of human subject datasets. However, only the largest NIH proposals (over $500,000) were impacted. This result is in contrast to the subsequent NSF policy in 2011 and would require data management plans for every proposal (U.S. National Science Foundation, 2011).

The updated version of the NIH policy, which will be effective January 2023 (U.S. National Institutes of Health, 2020), includes two critical changes: First, a review of compliance will be conducted with the data management plan that was submitted with the funding proposal. Second, the scope of proposals now covered includes all proposals submitted to the NIH, while still accounting for the flexibility and guidance needed in cases of sensitive research data. A few key definitions from the plan that will be useful to share in their entirety and will inform the remainder of the dissertation are as follows:

- **Data Management:** "The process of validating, organizing, protecting, maintaining, and processing scientific data to ensure the accessibility, reliability, and quality of the scientific data for its users."
- **Data Sharing**: "The act of making scientific data available for use by others (e.g., the larger research community, institutions, and the broader public), for example, via an established repository."
- **Data Management and Sharing Plan**: "A plan describing the data management, preservation, and sharing of scientific data and accompanying metadata."

Notably, the NIH chooses the phrase "data management and sharing" plan whereas the NSF referred to this concept as simply a "data management plan". One reason for the NIH's distinction may be that these are arguably two different concepts: Good data management might include practices that protect data assets, such as using storage backups, observing good file organization practices, and creating metadata—whereas data sharing emphasizes the distribution and dissemination of the data. It could be said that even highly classified and sensitive data should still be subject to good data management practices, regardless of whether the intention is to share the results after the fact. Nonetheless, for practical purposes, the phrases "data management plan" or "DMP" and "data management and sharing plan" will be used interchangeably throughout this text and should be assumed to refer to both types of practices.

Section VIII of the new NIH final policy on data management and sharing enumerates how compliance and enforcement of the policy will be addressed. To date, this policy is the most specific statement about enforcing compliance that has been written by a federal agency. Compliance evaluation may occur at the usual annual reporting intervals, as part of the usual practice of assuring that awardees are acting under the terms, conditions, and regulations of the award. In this case, the same risk is present that multi-year funding will not be continued should the report not be satisfactory. If compliance is evaluated after the award performance period, the institution risks a lack of continued funding as a consequence of unsatisfactory outcomes (U.S. National Institutes of Health, 2020).

2.1.2. Data Management Planning Requirements for Funders

It remains to be seen if the NIH's decision to implement a compliance process will prompt other agencies to reevaluate if and how their policies should be enforced. In the case of data management plans, other U.S. agencies followed the NIH and NSF in requiring DMPs in some form (although we have seen that events within the federal government helped to urge agencies along in this regard). Examples of data management planning requirements are not limited to federal agencies but may also be found among private foundations and other nongovernment sources of research funding (Williams et al., 2017). A survey of the current literature on data management planning is a useful source of context for thinking about the implications of their enforcement.

In one of the earliest examples of analyzing data management plans as artifacts, a pool of 1,260 DMPs submitted to the NSF from the University of Illinois from 2011–2013 were evaluated. The authors could not discern differences in the data management strategies between funded and unfunded proposals but did notice an overall increase in the use of institutional and disciplinary repositories as the NSF policy of requiring DMPs matured (Mischo et al., 2014). A multi-institutional study of DMPs, limited this time to only winning NSF proposals, observed substantial differences across domains in terms of the level of detail and quality of information provided in DMPs. They also noted a wide variety in the strategies for data sharing that the investigators behind this pool of winning proposals had included (Parham et al., 2016).

Inconsistency in DMP expectations, from one funder to the next and even from one agency directory to the next, is one source of difficulty for researchers. To illustrate, a topical content analysis was performed on sixty-six data management or sharing plan requirement documents for an assortment of federal, industry, foundation, and other miscellaneous funders. Forty-three subtopics were identified within seven broader themes (Williams et al., 2017). These themes, with some example subtopics, are as follows:

- Aspects of data values (accuracy, attribution, integrity, traceability)
- Aspects of data elements (definition, instruments, calibration, standards)
- Aspects of the dataset (ownership, rights, organization, relationships)
- Data values, elements, or sets (access, workflows, data flows, quality assurance)
- Systems used to manage data (storage, backup, security)

- Data disposition (preservation, formatting, curation, disposal)
- Project management (responsibilities, milestones, timing, status reports)

Products such as the DMP Tool (Swauger, 2015) and an assortment of services have been developed to aid researchers and organizations in coping with these complexities.

2.1.3. Commercial vs. Non-Commercial Funder Data Policies

Thus far, the review of literature has primarily addressed the policies of a few major, federally supported funding agencies. Although not the focus of this dissertation, commercially funded research is given a few considerations as well. Literature on the availability of industry-funded research data is almost, if not entirely, exclusive to the pharmaceutical industry—where human subject protections supersede the rights of industry to protect the commercial value of data. Transparency in clinical trials is perceived to be particularly important to a wide variety of audiences, including participants, regulators, and publishers. In investigating the availability of clinical trial data to independent researchers, one study found slightly over half (52%) of 61 pharmaceutical industry-sponsored trials to have a data sharing policy in place for results. Some trials were ineligible for data sharing, at the time, because of continuing follow-up studies or the medicines lacking regulatory approval (Hopkins et al., 2018). Furthermore, large industry sponsors were more likely to have formal data sharing policies in place (Hopkins et al., 2018) and more likely to have made trial results openly available (Axson et al., 2021). However, it has otherwise been suggested that the language of such policies is too often ambiguous and contradictory (Goldacre et al., 2017).

2.1.4. Journal and Publisher Level Policies

On the basis of articles from a small sample of high-impact journals with data policies, it has been suggested that data sharing statistically coincides with a higher rate of citations (Christensen et al., 2019). However, the proportion of peer-reviewed journals that adopt data policies still varies greatly, both within and across domains. A cross-disciplinary study of 150 journal policies, based on the journal websites, determined that biomedical sciences lead the way in adopting data policies, while social sciences and physical sciences are comparable in their adoption rates (Tal-Socher & Ziderman, 2020). An examination of 146 business journals found policies that require data to be uncommon, although many encouraged as much (Dosch & Martindale, 2020). A similar study of 447 titles in the sciences likewise found that data policies were far more likely to encourage than require data sharing, while also noting that high impact factor journals were more likely to at least have some form of data policy (Resnik et al., 2019). A larger study evaluated 880 journals for data policies from the physical and applied sciences, also using an impact-factor-based sampling approach. The assessment was completed once in 2014 and again in 2016, coding each title as "required, recommended, or no policy". The results indicated that the percentage of journals with a data policy was growing within the sciences. They also analyzed a subset of articles published in these journals in which the corresponding author was from the University of Toronto, finding their Chemistry department to be the most affected by the uptake of policies by science journals (Dearborn et al., 2017).

Notably, however, providing a data policy and enforcing that policy are not one and the same. Overall, actual instances of data sharing within a publication, in terms of statements of the availability, are often found to be low (Jeong, 2020). This finding may be due, in part, to differences in how authors and editors in many cases interpret the meaning of the policies (Christensen et al., 2019). Such confusion is unsurprising, given that journals can vary widely in terms of whether policies exist, their strength, and their levels of specificity (Rousi & Laakso, 2020). Regarding certain types of data, such as human subject data, it has been observed that, "researchers who are not 100% confident that it is legally and ethically fine to allow access to data that they collected may go for the cautious option of keeping data private" (Garellek et al., 2020). To illustrate, a study of 130 papers sourced from substance abuse journals found that data sharing was extremely rare (approximately 6%) (Gorman, 2020). A study of biomedical journals aimed to test not only the availability of data but also the reproducibility of the results for articles reporting randomized controlled trial outcomes. This analysis found 46% (17/37) of the cases to actually engage in data sharing practices, despite being published in biomedical journals requiring data availability. The numbers dropped to 38% (14/37) when considering only those trials whose results could be fully reproduced based on the data (Naudet et al., 2018).

2.1.5. A Note on Data Sharing Strategies

To achieve clarity for the discussion of data policies, it is useful to consider more thoroughly the meaning of data sharing and what kinds of actions satisfy this meaning. How researchers might do so includes the following examples:

- The authors claim that research data will be made available if it is requested from them.
- Supporting data is included with the article at the time of publication, as some form of supplementary information (supplementary files or supplementary information is a form of digital artifact usually made available in the same place as the article).

- The data artifacts are archived separately in a place where others can access them, such as in a repository or a website.

Noting once again the nuances of data sharing, these outcomes are not equal in terms of openness or risk. From a user's perspective, data that are available upon request are differentiated from deposits in a public repository in that requesting data constitutes an extra step of work and an opportunity for authors to later ignore or refuse the request. On the other hand, the contents of data repositories are not always, by default, free of access restrictions. For example, many human subject datasets are found in the repository managed by the Inter-university Consortium for Political and Social Research, and these datasets may have restrictions that include a review process for releasing sensitive data (Inter-university Consortium for Political & Social Research, 2222). Overall, even well-intentioned efforts to share data, such as archival on a personal website, may not qualify as good data practices depending on the level of risk introduced (Tenopir et al., 2020).

For publications and data management plans, statements of data availability found in publications are a straightforward source of evidence about the types of messages that researchers use to communicate about data sharing. A sample of articles from the journal Nature found that 43% (190/441) of the time, authors claimed that their data are "available upon request" to address data sharing policies. Only 15% of cases featured authors who advertised their data to be in a repository. Wide differences were found in the proportion of these behaviors when broken down into broad domains. For example, within life sciences specifically, "upon request" was used only 30% of the time, whereas that number jumps to 65% for physics. The discussion speculates that differences in the availability of domain-specific repositories could explain some of these findings (Grant & Hrynaszkiewicz, 2018). A similar study of articles published in the journal PLOS One, which used a slightly more granular coding scheme, reported "upon request" only 1.4% of the time, while data sharing through supplementary materials was more common. Repositories in this case were mentioned approximately 15% of the time (Federer et al., 2018). A study conducted on a much larger scale using an assortment of titles published by Wiley (n= 124,000) used machine learning classification methods rather than manual coding. The results found that making data available by request to authors was still the most common outcome of data availability statements (Graf et al., 2020).

As mentioned, a statement of data availability is useful as a communication tool, but it cannot be assumed that messages are synonymous with outcomes. Some argue that including "upon request" as an acceptable form of data sharing simply encourages the evasion of open data principles. One study found only an average of slightly under 40% of authors actually following through on such requests for data (Tedersoo et al., 2021). Yet another investigation tested author compliance for publishing in a journal with strict data sharing requirements. In these cases, as well, the author of a publication had committed to stating that supporting data would be available "upon request" to fulfill the journal's data sharing requirement. Compliance was found to be poor, with only 1 out of 10 authors actually providing the requested data (Savage & Vickers, 2009). A separate study in Economics reported a 44% compliance rate for promises of data upon request, further corroborating these conclusions (Krawczyk & Reuben, 2012).

The above examples describe data availability statements in the context of publications. However, this type of data-sharing commitment is also found in the DMP portions of grant proposals. For example, an investigation found that large portions of investigators indicated within their NSF proposals that they would only share research data upon request (although in this case, the study did not test how many authors followed through on such a request) (Parham et al., 2016).

Part of the challenge of compliance enforcement from the publisher's perspective may be disagreement as to whose responsibility this task should be and where these steps should occur in the publication process (Savage & Vickers, 2009). Furthermore, the additional processing time for manuscripts resulting from a review of data availability statements must also be considered (Grant & Hrynaszkiewicz, 2018). Interestingly, however, some of these investigations into the distributions of types of data availability statements varied widely in their results—suggesting that journal publishers may indeed be able to promote compliance.

2.2. Studies on Behaviors, Beliefs, and Needs

With so much happening in the world of research data, it is no surprise that institutions have been interested in what research data management means for their organizations. Several approaches are found in the literature for arriving at this answer. One aforementioned study, which included a dataset of corresponding authors from the University of Toronto, examined how often the researchers published in certain high-impact science journals with data policies (Dearborn et al., 2017). It would indeed be useful to know, at scale, the extent to which journal policies affect an institution's researchers. However, a deviation exists between the implementation of a policy and the extent to which the policy is complied with. A different and equally useful metric, then, would be to ask how often the researchers of an institution actually engage in data sharing.

Some efforts have been made to obtain this kind of information through self-reported data. Often, the studies focus on developing tools and services to support RDM. For example, representatives and scholars funded through the Belmont Forum were interviewed about their practices and training needs as part of a study to inform the development of a toolkit. Among other findings, the study highlights the very real uncertainties experienced by scholars surrounding how data is stored, how to build costs into research budgets, and how to cope with matters of intellectual property, ownership, and responsibilities (Bishop et al., 2020). Investigations of attitudes and data sharing and reuse practices have also been conducted in the context of specific disciplinary fields. Researchers in food science and technology were found to share many of the same concerns about open data reported by previous studies: fear of being scooped, confusion about data rights and legal constraints, concerns about misuse, and other reasons were cited. Common incentives to share included funder compliance, support of reproducibility, and hope for citations (Melero & Navarro-Molina, 2020b). Geophysicists were likewise found to see the merit of sharing data, while simultaneously harboring concerns about potential misuse or failure to cite if their data were shared (Tenopir et al., 2018).

A global and cross-disciplinary study found that, when data practices were scored according to FAIR data principles, most researchers still engage in moderate to high-risk data practices—such as only storing data on their personal computers, departmental drives, or USB drives. At the same time, many researchers indicated their dissatisfaction with these practices. Data sharing and reuse were overall viewed positively. The lack of open data was generally seen as an impediment to the progress of science, regardless of whether an individual could see the use of secondary data sources in their work (Tenopir et al., 2020). The earlier, 2011 version of this study of researcher practices and perceptions (Tenopir et al., 2011) was highly influential, setting off a wave of similar studies seeking to translate these findings to a more localized context.

Some studies are largely dominated by the use of surveys, focus groups, interviews, and other standard research methodologies to investigate the behaviors, choices, and attitudes of researchers surrounding research data management. Usually, they are case studies that describe specific institutions, although a few cut across institutional and disciplinary boundaries. For example, a brief 2014 assessment of faculty from high-research/very high-research universities (per the Carnegie classification) asked where they put their data, how they used data, where they look for data if data were incorporated into teaching and their opinions on who should be in charge of managing data. The authors positioned their study as an inquiry into whether faculty have the skills to comply with NIH or NSF DMP requirements and posited their results to indicate that "while faculty desire to share their data, they often lack the skills to do so" (Diekema et al., 2014). It is debatable, however, if the data produced by the study's six-item survey instrument could support such findings. For example, questions about data discovery, teaching practices, or opinions about who should be in charge of data management are not relevant to funder policy compliance and do not provide insight into preparedness to engage in data management planning. Other examples of institutional assessment are found in the literature of investigative RDM as case studies for specific academic research populations. These include faculties at California Polytechnic State University (CPSU) (Scaramozzino et al., 2012), the University of Vermont (UV) (Berman, 2017), and Emory University (Akers & Doty, 2013). While the Emory University survey instrument asked some useful "data demographic" questions such as the size of data and current backup methods (Akers & Doty, 2013), a continued emphasis was placed on questions about attitudes about data sharing and interest in services. For example, "With whom are you willing to share your data?" and "What services would you be interested in?" are survey items that measure researcher opinions and interests. Likewise, the CPSU (Scaramozzino et al., 2012) and UV (Berman, 2017) studies also measured beliefs, attitudes, and/or interests.

Attitudes and perceptions to our understanding of research data management outcomes. From an assessment standpoint, institutional value may also be found in gauging interests in data services. However, many of these studies seem to implicitly assume that the participants are knowledgeable of information science and research data management and that lack of participation in RDM is a deliberate choice of the researcher. While attitudes may truly predict RDM behaviors in some cases, the role of personal knowledge and an understanding of such concepts as metadata, persistent identifiers, and storage infrastructure should also be explored.

2.3. Data Services

Institutional studies of researcher practices and perceptions for data management are often motivated to inform the development of research data services. In this context, research data services (RDS) refer to formal support for researchers engaging in data management activities. The nature of this support may be broad and can span the boundaries of multiple administrative and functional areas traditionally found within an academic institution, such as Libraries, Information Technology, and Research support offices. Many of these activities have often been seen as a natural extension of support traditionally offered through academic libraries, including advisory and support, information literacy (data literacy), and repository management (Cox et al., 2019). Tenopir et al. (2014) proposed an approach to conceptualizing RDS using two categories of support: informational, consulting-type services, which are more learning and knowledge oriented (such as finding DMP examples or locating a repository), versus technical, hands-on services, which are more about doing, such as running a repository or writing the DMP). Current research suggests that the provision of research data services is largely still a work in progress (Cox et al., 2019; Tenopir et al., 2014). Academic libraries specifically have tended to more commonly offer informational services than technical services, which may be due in part to a staff's perceived need for more opportunities to train the skills needed to provide RDS (Tenopir et al., 2014).

An international study reported that compliance with funder policy was the most widely acknowledged driver behind the development of research data services. At the same time, the study observed that compliance with publisher policies was less often cited as a motivation for providing RDS (Cox et al., 2019). This difference in urgency may be understandable, given that funding agencies have more direct relationships and greater leverage with institutions. The relationship between institutions and publishers, however, is mediated by the authors.

2.4. Research Problem

Some of the major considerations may be summarized as follows: funder compliance has been cited as a compelling reason for researchers to engage in research data management and sharing (Melero & Navarro-Molina, 2020b) and for research institutions to develop support services for these researchers (Cox et al., 2019). The urgency for ensuring that investigators can meet the requirements of data policies has never been greater, with the impending onset of the NIH's latest data policy at the start of 2023. These new rules set the grounds for loss of funding should an institution's researchers fail to comply with NIH data policies (U.S. National Institutes of Health, 2020). Given the history of the NIH as an early adopter of data policies that are later taken up by other agencies, this shift could eventually also reach beyond a single agency. Beyond the potential for lost funding, research organizations have other reasons for assessing the ability of their employees to engage in data management. Research data management is an important consideration for the assurance of research integrity (Bishop et al., 2021). Additionally, as publishers become stricter in their enforcement of data-related policies, these skills could limit the ability of researchers to disseminate their findings in the most desirable journals.

The literature examples cited in this review have shown that internal assessment is one strategy used within academic research institutions to try to understand research data training and service needs. In most cases, the survey instruments do make practical inquiries into information such as the type and size of data produced. While nothing novel is contributed by such survey items, localized data would certainly drive the point home, more so than generic results, for organizational decision-makers who may be on the fence about investing in research data cyberinfrastructure or services.

Attitudes and motivations are an important consideration for understanding the uptake of RDM and data sharing principles, given that they may impact implementation and adoption. Many institutional assessments that have published their survey instruments often prioritize the motivational aspects of why researchers make the data management choices they do, such as how researchers feel about data sharing, what motivates them, or under what circumstances they would engage in data sharing (Tenopir et al., 2020). An example of an opinion-focused survey item would include, "Do you think it is important to share your data with others" (Scaramozzino et al., 2012).

Other examples of existing survey questions ask researchers what kind of help they think they need. While all of these answers contribute to our understanding, they do presuppose that researchers have sufficient knowledge of RDM to make informed choices or to effectively communicate their needs. For example, the survey items "I follow criteria for preserving my data" and "I create descriptive information" are found in the literature (Scaramozzino et al., 2012). However, the validity of these items as data collection instruments rests upon an assumption that researchers are knowledgeable enough to answer. What if they do not fully understand what those criteria should be, and what if they are not trained to recognize what sufficiently counts as descriptive information to support data sharing, discovery, and reuse? More evidence is needed to understand the relationship between knowledge and actions.

Many of the examples of studies cited in this review appear to be inspired by seminal works (Tenopir et al., 2011). This detail demonstrates that field practitioners often apply the methodologies and investigative techniques that are first developed in Information Sciences or Social Sciences to guide their local assessment practices. The overarching objective of the current project has been to further develop the methods and instruments that are available to practitioners. Moreover, the need for such assessment tools is greater than ever, given the impending possibility of serious consequences for failure to comply with NIH data management and sharing policies. The assessment survey that has been developed, demonstrated in Chapter 3, fills this need and employs questions designed to reveal trends in a community's understanding of important RDM concepts needed to comply with data policies. Chapter 4 demonstrates a process for using the research outputs of that community to benchmark their current data practices and potential readiness to comply with data sharing policies.

3. Data

Data management and sharing (DMS) behaviors have, up until now, largely been investigated through individual drivers such as personal motivations and interests. The newest NIH data management and sharing policy introduces a new source of external influence to engage in ethical sharing and management behaviors. More specifically, the final DMS Policy includes, "an expectation that researchers will maximize appropriate data sharing," while still accounting for any ethical, legal, or technical factors that may require alterations. The existence of these unique factors does not negate the responsibility of the researcher to maximize data sharing where possible. For example, researchers working with human subjects data are urged to consider how to modify the informed consent process so that participants understand what will happen to their data. The final draft policy also warns that individual divisions within the NIH should be expected to promote more specific expectations within funding announcements, such as the example of particular metadata and data standards with the intent of maximizing interoperability (U.S. National Institutes of Health, 2020).

Standards for details within the data management and sharing plan have been raised. Researchers are expected to communicate any limiting factors in the context of data management planning. "To be determined" is no longer an acceptable response to a request for information as part of the data management and sharing plan. Researchers are explicitly asked to describe how scientific data sets will be findable and identifiable, such as via a persistent identifier or similar means (U.S. National Institutes of Health, 2020).

In summary, the NIH policy change strengthens expectations for support of data management and sharing through the lens of FAIR data principles. These policy expectations are not unique to the NIH; data management and sharing policies have become common across funding agencies and publishers. What has changed, however, is both the scope of NIH data covered by the policy and the consequences of failing to carry out activities as described within data management and sharing plans. Given the new situation, academic research institutions will benefit from assessment strategies that can inform their efforts to prepare for data policy compliance.

3.1. Research Questions

The research questions for this dissertation are intended to represent the information needs of an organization preparing to ensure compliance with research data management and sharing policies. This includes not only researchers' knowledge of how to carry out the kinds of specific RDM activities described in the NIH and other data policies, but also researchers' expectations of the relationship between themselves and their institution. As researchers are faced with changing demands, they must know how to seek out support for meeting those demands through formal institutional channels. It is also crucial to know to what extent researchers' current practices are consistent with the data management and sharing principles that will be necessary for compliance with policies that are expected to play a larger role than ever before.

- **RQ1:** How well do researchers understand how to perform the kinds of data management tasks associated with data policy compliance?

- **RQ2:** Does level of understanding of RDM tasks differ based on characteristics of the researcher?
- **RQ3:** Which RDM obstacles are the most significant barriers to data policy compliance?
- **RQ4:** Where do researchers expect to find support for RDM tasks within their home institution?
- **RQ5:** What can different tools tell us about how well data management practices align with data management and sharing principles, at the time of assessment?

3.2. Survey Instrument

The primary instrument for collecting data to address the research questions was a survey, found in full within as an Appendix. After obtaining IRB Approval (UTK IRB-21-06400-XM), the survey was distributed using a combination of purposive and snowball sampling to an existing mailing list of researchers. The University of Tennessee at Knoxville was chosen as the pilot institutional location. The emailed recruitment letter contained a link to the online survey, administered through QuestionPro. Within the recruitment letter, email recipients were asked to help distribute the survey among their respective groups and departments to anyone engaging in research at the University of Tennessee, Knoxville. Response rates are not reportable due to a non-targeted distribution approach, but n=54 completed responses were obtained. Before beginning the survey, participants were required to provide informed consent. The IRB-approved consent document included a statement explicitly letting the participant know that their anonymous responses would be archived in a publicly accessible repository.

3.2.1. Survey Items

Survey items included questions designed to help an institution characterize the types of digital artifacts created in the course of research, including types of files and storage requirements for data and related digital artifacts. Participants were asked to provide demographic data such as domain and position within the university. They were also asked to indicate if they worked with human subjects data, as well as the frequency with which their projects involved classified or sensitive data of any form.

To explore RQ1, participants were asked questions designed to measure their level of understanding of seven common data management and sharing tasks that may be referenced in data policies, such as including a data availability statement, archiving data in a repository, or creating metadata and other descriptive documents. For researchers who had indicated their use of human subjects data, several additionally relevant RDM activities were presented, such as how to draft informed consent statements that are compatible with data sharing.

The potential for social desirability bias was a concern in the wording of these questions. Bias is one example of possible measurement problems in which observed differences among individuals are due to errors in the measurement process rather than true individual differences. Social desirability bias refers to a tendency of individuals to "deny socially undesirable traits or qualities and to admit socially desirable ones." More specifically, social bias introduces threat to measurement validity in that participants will be biased towards responses that portray them favorably based on their perceptions of desirable traits (Phillips & Clancy, 1972). For example, survey respondents are known to under-report socially undesirable activities and over-report socially desirable ones. Careful survey design can reduce participant embarrassment and improve validity of data (Krumpal, 2013). In the context of the present study, admission of personal ignorance about any aspect of the research process could result in personal embarrassment or professional consequences for participants.

The language of specific survey items sacrificed a degree of face validity in exchange for mitigating the threats to measurement validity that would have otherwise been introduced by social desirability bias. For RDM topics, the objective was to measure individual levels of understanding. For example, the survey could have asked, “If a data policy required you to create descriptive metadata for your research data, how well would you understand this concept?” Regardless of one’s actual knowledge about scientific metadata standards, research data expertise is a desirable trait for the population targeted by the current study, and ignorance is an undesirable trait. Alternatively, the survey design chose to situate the onus of responsibility for participants' knowledge externally by asking, "How well do you feel the task of creating metadata and documentation for research data has been explained to you?" Participants were provided a four-item Likert-type scale to record their responses, with possible answers including "Poorly Explained,” "Modestly Well Explained,” "Well Explained,” and "Very Well Explained.”

To address RQ2, data about the researcher was collected, including primary domain of research, position at the university, and the role of the participant on projects. Whether or not the participant had worked with human subjects data, or any other form of sensitive or classified research, also provided additional dimensions of researcher characteristics. This emphasis on distinguishing the sensitivity of the data is unique to this survey, and these questions were designed with the intent of better understanding the relationship between data constraints and other outcomes.

To investigate RQ3, participants were asked the extent to which several kinds of broad obstacles found in the literature would pose a barrier to data management and sharing. While it would be difficult to acknowledge every possible barrier at a highly granular level, four very broad examples of obstacles to good RDM were presented within the survey. These included finding an appropriate long-term archival solution, the time to prepare for preserving and sharing data, costs to archive, and security considerations. Participants were asked to what extent each example would be a barrier to data archival, with possible answers from a four-item, Likert-type scale. Possible answers included "Not a barrier,” "Somewhat of a barrier,” "Moderate barrier,” and "Extreme barrier.”

For RQ4, and to better understand researcher expectations for sources of institutional support, participants were presented with the same list of seven RDM tasks previously presented, then asked which of several common administrative units they would be most likely to seek help from. These units included the Office of Technology, Libraries, Office of Research, and the Office of the Provost. These options were chosen with the intent to be as institution-agnostic as possible so that the survey instrument could be used beyond the pilot institution. This decision would also support the ability to compare across institutions in future studies.

To answer RQ5, participants were asked to indicate their use of nine different examples of long-term data storage options, ranging from localized solutions such as storing post-project data on a personal computer, to cloud-based storage or departmental servers.

3.2.2. Survey Data Preparation

In anticipation of computational analysis using R statistical computing language, data preprocessing steps were taken to restructure the online survey software's export format to maximize machine readability. Results were exported from the QuestionPro online platform in .csv format, using the "Display Answer Codes" and "Single Header Row" options. These options exported responses in numeric code format instead of verbose labels and condensed the column headers into single-row values. A preprocessing file, *preprocesing.R*, has been archived with the data, and contains the script used to prepare the data for analysis. Preparation steps for restructuring the survey platform's export included filtering out incomplete answers, stripping lines that did not contain data, and discarding unpopulated or irrelevant columns. Additional steps included renaming columns to meaningful, machine-friendly labels and recoding data values. The output of the preprocessing script is stored as *IHLI-2022-surveydata.csv*, which has been reviewed to ensure it contains no identifying information,r and archived with the project files in figshare as the data version of record. All subsequent scripts for analyzing data (further detailed in Chapter 4) have been written to ingest *IHLI-2022-surveydata.csv* as input.

3.3. Researcher Publications

RQ5 asked what we can learn about current RDM practices using different assessment tools. Therefore, survey data is supplemented by an additional data source with a new workflow.. Publications are a primary channel through which researchers communicate about the availability of data that supports the findings. However, the types of behaviors that can be observed are limited. For example, if a paper indicates the data is available upon request, we have no way of knowing if and what steps were taken to preserve the data after the life of the project, only that the data are not likely to be in a place where others can find it. However, if the paper notes that supporting data are available in a repository, this would be evidence of data management and sharing behavior consistent with policies such as the NIH's.

Despite the limitations to what can be learned from full-texts, there are also benefits. Publications may already be accessible to an institution without additional intervention by the researcher, through existing library subscriptions or policies for deposit into institutional repositories. Given the risk of low response rates that commonly plague surveys (the present work being no exception), it can be desirable to consider data that can be independently collected.

For this study, published works from researchers within the target population (The University of Tennessee) were examined to explore their value as a source of information regarding current data management practices. The selected sample is limited to journal articles and conference papers because they are generally accessible as full-texts in electronic format. The sample scope is limited to 2021 to coincide with the approximate time period in which the survey was administered. Selecting from among papers where the lead author is from the target institution provides a meaningful approach to discriminating between authors who are likely to have played a leading role in how the project was carried out, versus those who may have only played a minor role. While some exceptions are expected, such as Economics, the position of first authorship is generally expected to coincide with the greatest effort towards the project for many fields (Weber, 2018).

Due to the large number of publications produced by any high-research institution in a given year, some method was needed for reducing the number of publications to be manually evaluated to a feasible number. Given that NIH-sponsored research specifically is slated to experience direct and meaningful changes to data policy in the near future, NIH-sponsored publications were an ideal subset of documents to examine. Table 3.1 was generated using the Incites research evaluation platform and summarizes the number of publications produced by this institution in 2021. Values are organized across the broad categories of the GIPP schema (help.prod-incites.com/inCites2Live/ filterValuesGroup/researchAreaSchema/gippDetail /version/2). Note that there is some degree of overlap in the category classifications of the journals in which these publications appear. The OVERALL category is included to provide the number of unique documents in each column. The first column represents all documents published by the University of Tennessee, Knoxville, in 2021. The second column shows the number of these which acknowledged funding, in part or in whole, from any part of the NIH during this same time period. The third column expresses the percentage of total UTK publications that attribute support to the NIH. In the next part, the fourth column shows only the subset of all UTK publications in which the first author was UTK affiliated. The fifth column reduces the NIH sponsored count of those documents in the same way. Finally, the NIH-sponsored percentage of all documents with UTK-affiliated first authors is reported in the sixth column. From this data, we can see that 94 publications meet the requirements of this sample definition. It is also interesting to note that NIH-sponsored publications are distributed across the categories, suggesting that the implications of these policy changes may be far reaching in terms

Table 3.1: Articles and proceedings by category in 2021 for pilot institution. Total documents versus NIH-supported are shown on the left, and affiliated first-author only are shown on right.

GIPP Category	**UTK Authored Documents**	**UTK Authored, + NIH Funded**	**% UTK Authored, + NIH Funded**	**UTK First Authored Documents**	**UTK First Author, NIH Funded**	**% UTK First Author, NIH Funded**
Eng & Tech	1,477	22	1.49%	641	12	1.87%
Phys. Sciences	1,223	28	2.29%	405	16	3.95%
Life Sciences	1,106	111	10.04%	488	51	10.45%
Social Sciences	581	28	4.82%	316	15	4.75%
Clinical Health	329	72	21.88%	150	34	22.67%
Arts & Humanities	49	0	0.00%	46	0	0.00%
OVERALL	3,744	195	5.21%	1,624	94	5.79%

of domain. Given that over 20% of Clinical, Pre-Clinical, and Health publications attribute support to the NIH, it is clear that researchers in this area will be most strongly impacted.

Web of Science metadata records were retrieved for each of the 94 publications meeting sample criteria. Each paper was downloaded for further analysis, with the exception of 2 documents which were not available through available subscriptions or other means. Upon examination, 5 documents were classified as review papers and discarded from further analysis. The remaining 87 documents consisted of 2 conference proceedings and 85 journal articles.

3.4. Data & Code Availability Statement

All supporting data, as well as R scripts for processing and visualizing data, are archived in a public repository at figshare under DOI 10.6084/m9.figshare.20280150.

4. Methods

4.1. Methods of Analysis for Survey Data

All data processing and visualization for survey responses was completed using the R statistical computing language (R Core Team, 2022) (version 4.2.1), with heavy reliance on *ggplot* and other *tidyverse* modules. Most analysis consisted of descriptive statistical visualizations such as bar charts, stacked bar charts, and pie charts to summarize characteristics of the data and demographic information for the participants.

Researcher characteristics were furthermore tested for independence from understanding of of the seven data management tasks measured. These characteristics included researchers' domain, whether or not they worked with human subjects data, their position at the university, and whether or not they worked with data which was sensitive or classified in any form. Fishers exact test was used as a test of independence, due to the small size of the sample.

In order to perform Fisher's exact test, the data for each pair of independent variables was restructured as appropriate to achieve a contingency table. Participants had been asked how well various RDM topics had been explained to them, as a measure of personal understanding. For each of the 7 RDM topics, possible responses included "Poorly Explained", "Modestly Well Explained", "Well Explained", and "Very Well Explained". These were collapsed into broader categories: "Poorly/Modestly" and "Well / Very Well".

Table 4.1 demonstrates the format of resulting contingency tables. This example shows the responses for how well the concept of metadata had been explained, along the dimensions of whether or not an individual worked with human subjects data. This process was repeated

Table 4.1: Contingency table for human subjects data status versus understanding of metadata creation.

	Metadata: Poorly / Modestly Well Explained	**Metadata: Well/Very Well Explained**
Works with Human Subjects Data	20	9
Does Not Work with Human Subjects Data	19	4

for research domain, funding status, human subjects status, principle investigator status, employment position, and data sensitivity status. Certain other researcher variables with non-binary responses were collapsed for contingency tables, including Funding status and Sensitivity of data. Possible funding status responses (“None”, “Some”, “Most”, “All”) were collapsed into the broader categories of "Not funded" and "At least some research funded". In the case of data Sensitivity, if the participant had indicated they worked with sensitive or classified data at all, their responses were collapsed into “Works with Sensitive or Classified Data” or “Never Works with Sensitive/Classified Data”. Full details for how the contingency tables were formed are available by examining the corresponding processing script for each researcher variable. The six researcher characteristic variables and the seven RDM topics resulted in forty-two separate contingency tables and computed p-values.

Survey items inquiring about current research data management practices were organized using faceted bar plots which reflect a categorization approach inspired by Tenopir et al.'s (2020) interpretation of FAIR data principles. The FAIR data principles describe an approach to characterizing the quality of RDM practices. Observing that data management is not desirable simply for the sake of itself, FAIR principles offer guidance on how to prioritize data stewardship principles which lead to favorable outcomes: knowledge discovery, innovation, and integration. FAIR data are that which meet the principles of findability, accessibility, interoperability, and reusability (Wilkinson et al., 2016) . Tenopir et al.'s (2020) approach was to characterize research data management practices as *good*, *mediocre*, or *bad*, dependent on their alignment with FAIR data principles. This study defined good data practices, such as archival in a repository, as those which which both protected data, by guarding against loss and facilitating curation, and also supported data sharing. Mediocre practices, such as storing on a personal cloud account, take some steps to protect data but do not support FAIR principles. Bad data practices both put data at risk and make it difficult to find or share the data. Examples of bad practices include storing data on a personal drive or computer, or printing data for storage (Tenopir et al., 2018).

The present dissertation further develops this classification scheme by also characterizing practices in terms of two dimensions: Risk and Openness. *Risk* refers to the extent to which data management practices protect the data itself, such as protection against data loss. *Openness* refers to supporting the ability of others to find and use the data. The survey had presented nine data management practices to participants, and asked how often each of those were used for long-term data storage. Possible responses for the five-item scale ranged from "Never" to "Always". These nine data storage options were organized into 3 groups which represent good, mediocre, and bad data management practices. Good practices were low in risk and high in openness, and included archival with a publisher, archival within the institutional repository, or archival in another public repository. Mediocre practices were low in risk and low in openness, meaning that the data archival options were choices that protected against personal data loss, but did not support data sharing for reasons such as limited access and lack of discovery features. These included cloud storage, archival with departmental servers, or with Office of Technology servers. Bad data practices were both high in risk and low in openness, and included personal storage media, personal computers, and data stored as printed records. The data visualization strategy for this analysis used stacked bar plots faceted along the classification of good/bad/mediocre practices. The stacked bars within each plot represented distribution among the range of possible responses.

All R scripts used to perform the analysis have been archived with the data, and are freely available. Those who may wish to repeat this study within their own organizations need only to restructure the data export format from whatever survey software is used to match the structure of the universal ingest file referenced in each R script (*IHLI-2022-surveydata.csv*).

4.2. Pilot Study for Publications-based Assessment of RDM

This part of the dissertation develops a workflow for incorporating publications as part of an organizational assessment strategy for researcher data management practices. In order to be realistically considered as an institutional assessment tool, the workflow must balance labor costs with the payoff of actionable information. Reliability is also a concern given the variations in researcher language used to describe data. This workflow addresses reliability by prioritizing a well-defined scope for what is being measured and how it is to be coded.

4.2.1. Data Availability Messages

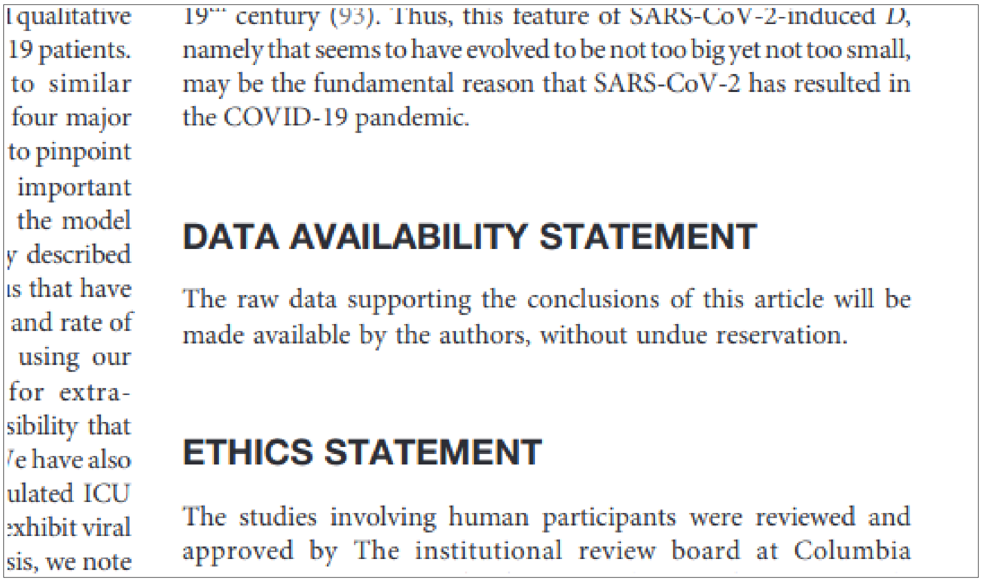
The Assessment for Data Availability Messages (ADAM) workflow is an organizational assessment tool for using publications to measure research data management outcomes. The unit of analysis for this workflow is a publication, and the units of measurement are messages about research data availability. The presence or absence of data availability messages were coded at the document-level. These were not weighted observations, meaning that even if multiple data sources were used and multiple statements about data availability observed, the document would simply be coded as containing such messages. A segment of text is considered a message about data availability if it guides the reader in locating the data which supports the findings presented in the publication. Thus, the statement "We studied the samples using data collected from a rheometer," would not be considered a message about data availability. It does not communicate information that would help the reader locate the data that was created, only how the data was collected. Examples of data availability messages would include "The data are availability as supplementary materials," or "The data used in this analysis are available on GitHub," or "The data are available upon reasonable request to the author." This range of examples illustrates that designation of a statement as a message about the availability of research data is free from any assertion as to the appropriateness, openness, or risks associated with the choices expressed in the content of the message. Nor does the observation of a data availability message make assertions as to the quality of the message. "The data are archived in Dryad," may be, in reality, wholly inadequate for the reader to succeed in locating the data within said repository, but it counts as a message about research data availability none the less. Further qualifying the determination of a data availability message, text segments could not be coded as communicating about data availability if the word "data", or some variant, did not explicitly occur in the context of the statement. For example, the phrase "The Supporting Information is available free of charge at the publisher website," would not qualify as a relevant message, due to ambiguity about what is included in the supporting information and whether or not it includes data to support the findings. This methodological approach serves two practical purposes: first, it avoids inconsistency in interpreting statements. Second, it provides a practical, keyword-based mechanism for quickly searching even long documents for potentially relevant messages.

While all of these decisions limit the richness of information that can be acquired from the analytical process, they provide a simplified decision-making architecture for quantitatively assessing the content of publications. It should also be noted that verifying the truth of messages about data availability was also outside the scope of this analysis. For example, the process would not involve looking up data in a repository to verify the credibility of a statement that research data had been archived.

To better understand the communication patterns of authors with regard to research data, messages were further classified based on the form in which they appeared in the publication. A *structured message* of data availability was defined as a statement under a section or header within the manuscript which is designated specifically for that purpose, such as but not limited to an explicit "Data Availability Statement". Structured statements can be located at various parts of the manuscript, but are commonly found at either the start or the end of the paper. An *unstructured message* would be a reference to the availability of the data which is not distinct from the surrounding text. For example, if the statement, "The data for this analysis were obtained from Protein Data Bank" were to appear within the the publication lacking any visible distinction from the surrounding text, it would have been classified as unstructured message. Examples of structured messages about data availability are shown in Figures 4.1 and 4.2. an example of an unstructured message about data availability is shown in Figure 4.3.

4.2.2.   Publisher Support for Structured Data Messages

Formal support for structured data messages have become increasingly common at the journal level. Several studies were cited in the literature review which address how authors



l qualitative
19 patients.
to similar
four major
to pinpoint
important
the model
y described
ıs that have
and rate of
using our
for extra-
sibility that
/e have also
ulated ICU
:xhibit viral
sis, we note

19th century (93). Thus, this feature of SARS-CoV-2-induced *D*, namely that seems to have evolved to be not too big yet not too small, may be the fundamental reason that SARS-CoV-2 has resulted in the COVID-19 pandemic.

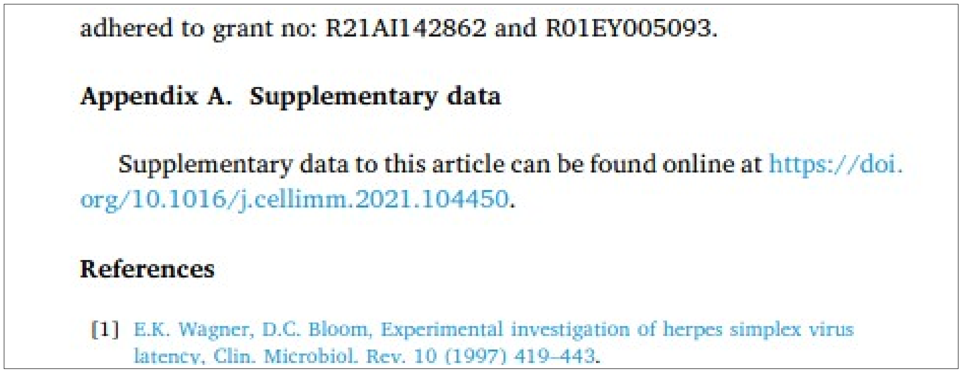
DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The institutional review board at Columbia

Figure 4.1: An example of a structured data availability message found in a publication.



adhered to grant no: R21AI142862 and R01EY005093.

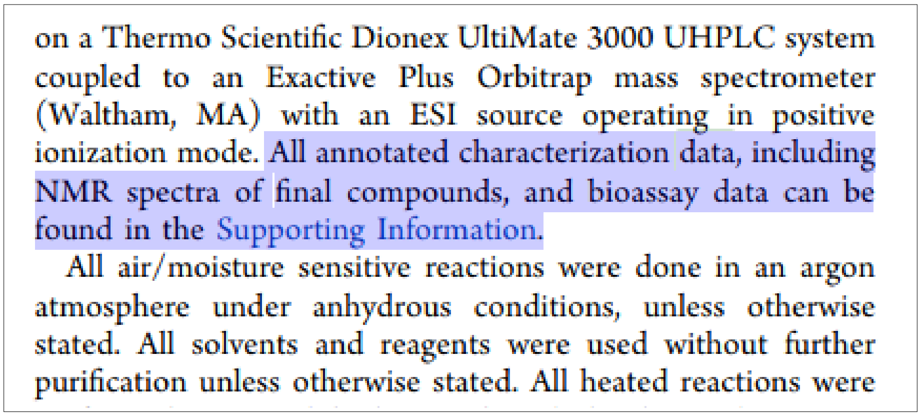
Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cellimm.2021.104450.

References

[1] E.K. Wagner, D.C. Bloom, Experimental investigation of herpes simplex virus latency, Clin. Microbiol. Rev. 10 (1997) 419–443.

Figure 4.2: A different example of a structured data availability message found in a paper from this set of articles.



on a Thermo Scientific Dionex UltiMate 3000 UHPLC system coupled to an Exactive Plus Orbitrap mass spectrometer (Waltham, MA) with an ESI source operating in positive ionization mode. All annotated characterization data, including NMR spectra of final compounds, and bioassay data can be found in the Supporting Information.

All air/moisture sensitive reactions were done in an argon atmosphere under anhydrous conditions, unless otherwise stated. All solvents and reagents were used without further purification unless otherwise stated. All heated reactions were

Figure 4.3: Example of an unstructured data availability message found in the text of a publication.

communicate through such data availability statements. However, the present work is the first study to apply the contents of data availability messages towards organizational assessment. Including both structured and unstructured messages strengthens the assessment by ensuring that researchers' data sharing messages are accounted for, regardless of whether or not they publish in a journal which formally supports formal communication in this way.

It is expected that authors will provide data availability statements when required to by journals with policies to that effect. Journals may express support or requirements for formal data availability messaging as part of their policies and other information presented to prospective authors. Table 4.2 shows different kinds of language used by journals to indicate support of formal statements for the availability or transparency of data. This language can be both varied and ambiguous. Journal-specific information for authors was looked up for the 85 journal articles in the publications analysis. Only journal-level information was considered, unless the journal explicitly pointed to publisher policies. The web addresses for each set of policies and information were recorded, and the contents reviewed. Any references to formal statements of data availability or transparency were recorded. Journal support for data availability statements was coded as a 1 if this language existed, regardless of if language indicated such statements were encouraged, required, or if it simply provided an explanation of their purpose. The value of 0 was coded if there was no observable mention for any kind of DAS

4.2.3. Data Citations

In this analysis, formal citations to the data used in a study were treated as a special property of the data availability message. Publications were identified containing a data citation

Table 4.2: Examples of language used by journals to support formal, structured data availability statements.

Journal	**URL**	**Language Supporting Data Availability Statements**
Annals of Behavioral Medicine	https://academic.oup.com/abm/pages/general_instructions	"During submission, authors are **required** to provide a transparency statement about data availability and how to access their data, analytic code, and research materials, which will be published with the journal article."
Biophysical Chemistry	https://www.elsevier.com/journals/biophysical-chemistry/0301-4622/guide-for-authors	"To foster transparency, we **encourage** you to state the availability of your data in your submission."
Biological Conservation	https://www.elsevier.com/journals/biological-conservation/0006-3207/guide-for-authors	"To foster transparency, we require you to state the availability of your data in your submission **if** your data is unavailable to access or unsuitable to post."
Journal of Cell Science	https://journals.biologists.com/jcs/pages/manuscript-prep	"Data availability All publicly available datasets supporting your work **should** be included in the Data availability section."
International Journal of Molecular Sciences	https://www.mdpi.com/journal/ijms/instructions	"Data Availability Statements provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study."
Journal of Immunology	https://www.jimmunol.org/info/authors#SupplementalData	Data availability statements not mentioned at all. But other author instructions state: "Any paper submitted… that contains new high-resolution structural data requires an accession number from the Protein Data Bank and assurance that unrestricted release will occur at or before the time of publication…. Complete microarray data must be deposited in the appropriate public database (e.g., GEO or ArrayExpress), and must be accessible without restriction from the date of publication."

only if a data availability message included a citation referring to the data, including a corresponding entry in the references list. No effort was made to otherwise evaluate the contents of references lists. As a consequence, if a dataset were to be included in the references, but not cited in context of a data availability message, that citation would not have been observed.

4.2.4. Regarding Secondary Data Analysis

One of the more distinct threats to validity arose in situations where the findings were based in whole or in part from secondary data. More specifically, this refers to cases in which a secondary data source is described, but availability of any new data produced in the current study were not acknowledged. Among this sample of publications, secondary data sources such as U.S. Census data or U.S. Medicare data were frequently relied upon to generate new statistical findings. An additional way that existing data might be used is to test the development of new models, algorithms, or other computational approaches. In such cases, the outside data serves the purpose of validating the primary innovation being communicated in the paper. Presumably, secondary data analysis for any reason still produces new results and data. By extension, benefit would still be derived from making said derivatives available in a raw and machine actionable form. These examples describe circumstances in which the description of the secondary data source qualifies as a message about data *used* by the study, not the data *produced* by the study.

To address this distinction, data availability messages were differentiated as referring to *original* or *secondary* data. For example, if a paper stated that, "The primary source of data for this analysis is 2010 U.S. Census data," the text would be classified as a data availability message about secondary data sources. It still helps the reader to locate the data which supports the study's findings, but the data is not original to the paper. If the same study later states, "All data and results derived from the analysis 2010 U.S. Census data are archived with the figshare repository at the following url,", this would be coded as a message about (original) research data availability. These rules were applied even when a secondary data source was included in a structured heading, such as a formal Data Availability statement. Citations to secondary data sources were likewise recorded separately.

4.2.5. Coding Process

While it is conceded that many cases will be varied and ambiguous, the effort has been made to define a consistent coding process which can be reliably applied and with reasonable burden. Full texts were retrieved for publications from the sample identified in Chapter 3. For purposes of this analysis, the PDF version was considered the version of the record, if available in lieu of an HTML presentation on the publisher's website. The first step was to scan the abstract to determine if the article constituted a review document or original research. As noted in Chapter 3, review papers were discarded from further analysis. The ADAM workflow is not designed for this type of document.

The PDF reading tool’s search functionality was used to scan every instance of the word "data" in the paper, to determine if it occurred in the context of a statement about the availability of data. While this automated some of the process, the keyword-based searching could be admittedly burdensome at times. For example, in one paper the word "data" appeared 156 times, whereas in other papers it only appears a few times. Text segments were evaluated as structured or unstructured messages using the described evaluation criteria, and further classified as referencing original or secondary data. The location of the document in which the messages were found was also recorded. The data availability messages were furthermore scanned for the presence of formal citations describing the data.

Once messages had been extracted, the message contents for original research data were next categorized. Content categories described the specific strategies for making data available. These coding options included: “With the Publisher”, “With a Repository”, “Upon Author Request”, “Author Website”, “Not Available”, and “Other”. Figure 4.4 summarizes the overall data collection and coding process for ADAM.

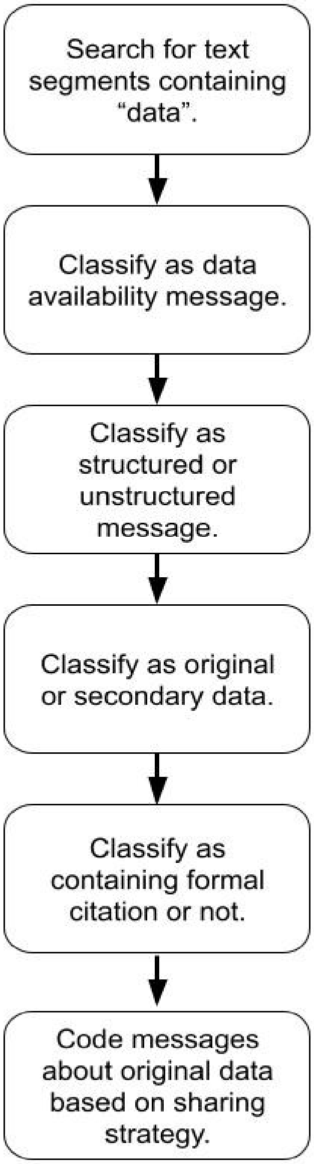

Search for text segments containing "data".
Classify as data availability message.
Classify as structured or unstructured message.
Classify as original or secondary data.
Classify as containing formal citation or not.
Code messages about original data based on sharing strategy.

Figure 4.4: Assessment for Data Availability Messages (ADAM) workflow, for evaluating full texts of publications.

5. Results

5.1. Characteristics of the Researchers

This section describes additional demographic information about the researchers who took part in the survey.   To simplify analysis and to ensure participant confidentiality, respondents were asked to identify their primary area of research using a very broad set of categories. A more detailed explanation with examples from within these categories is outlined within the survey instrument, in Appendix A. Figure 5.1 shows that the highest participant rate came from the Natural Sciences, followed by Social & Behavioral Sciences, then Health Sciences, and Veterinary Medicine and Agriculture. Figure 5.2 illustrates that only 9.3% of survey participants did not work with externally funded research at all, suggesting that funder-level data policies can be expected to be wide-reaching in their implications for this community.

Although several other types of positions were available, such as staff or undergraduate students, the survey responses were limited to 80% faculty and 20% graduate students, as shown in Figure 5.3. It is unclear if this outcome was an unintended consequence of the recruitment method, or if the data reflects that there genuinely are not many non-faculty participating in research at this institution. This result will be further addressed in the limitations. Drilling down further into the makeup of faculty in Figure 5.4, the majority of respondents to this survey were tenured faculty (65%) whereas those who were on tenure track but not yet tenured made up 26%. Only 9% of respondents were non-tenure track faculty.

Additional questions intended to further explore the significance of job role in research

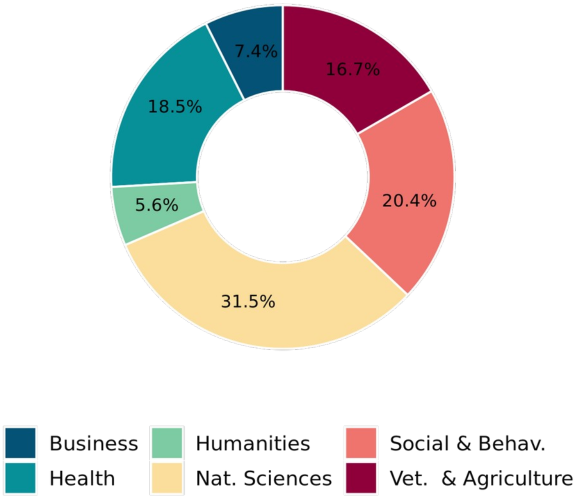


Figure 5.1: Primary research domain reported by n=54 survey respondents.

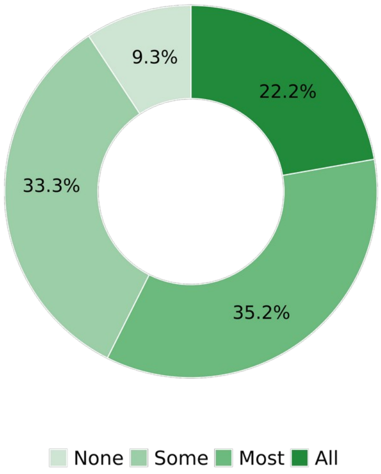


Figure 5.2: Amount of research which is funded, for n=54 survey respondents.

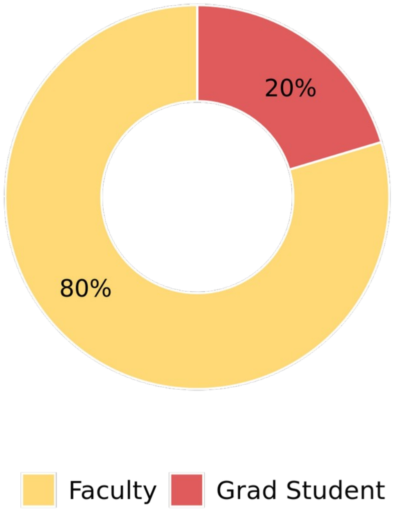


Figure 5.3: Distribution of position with university, for n=54 respondents

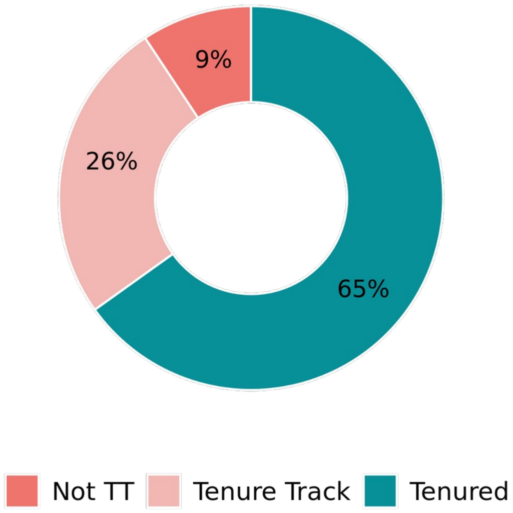


Figure 5.4: Distribution of tenure status for n=43 faculty who responded to survey.

data management included asking whether the participant had served as a principle or co-investigator in recent projects, and whether or not they felt they had decision-making responsibilities for RDM in recent projects. These results are reported in Figures 5.5 and 5.6, respectively. The majority (82%) of respondents had indeed served recently in a principle investigator role, meaning they had primary responsibility for the execution and fulfillment of a project. Comparably, 79% felt they had been in a role of decision-making for RDM, in the context of recent projects.

5.2. Characteristics of the Data

The distribution of types of digital artifacts created in the course of research is shown in Figure 5.7. Spreadsheets and generic text data are shown to be most common to the majority of researchers, followed by images and various proprietary formats. The archival storage requirements for data produced by research projects is reported in Figure 5.8. Of particular interest, 17% of the respondents reported data storage requirements on the order of terabytes or beyond, suggesting that the size scientific data continues to grow.

Figure 5.9 reports the percentage of respondents who indicated each possible answer for the question, "How often do you work with the following categories of data?". The three possible categories presented included Sensitive (medical or participant privacy, trade secrets, export control, etc.), Classified (by government designation), or Open (no external constraints on sharing and reuse). Through the visualization, it is clear that the distribution lean towards open data. However, a sizable contingent does work with data that is sensitive for one reason or another at least some of the time. Classified data represents only a small portion of the work that

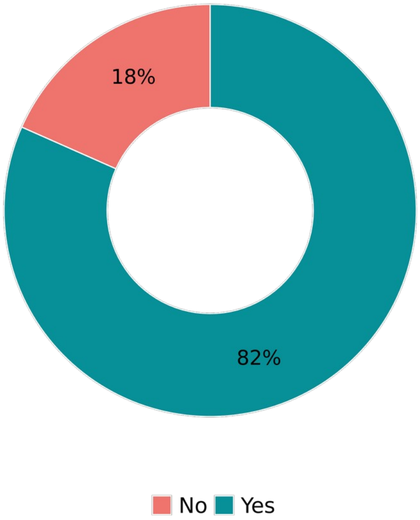


Figure 5.5: Percentage of n=54 respondents who have served has principle investigator of a project in the last 3 years.

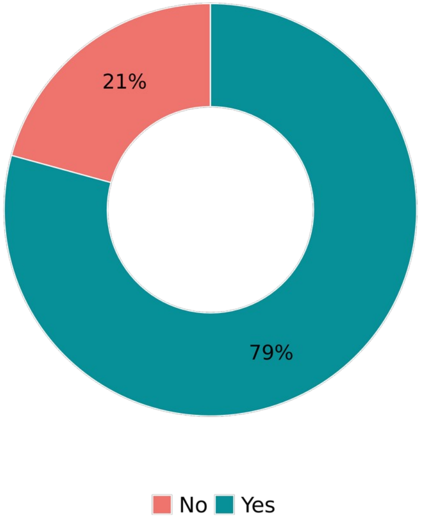


Figure 5.6: Percentage of n=54 respondents who have served in role of making data management decisions in the last 3 years.

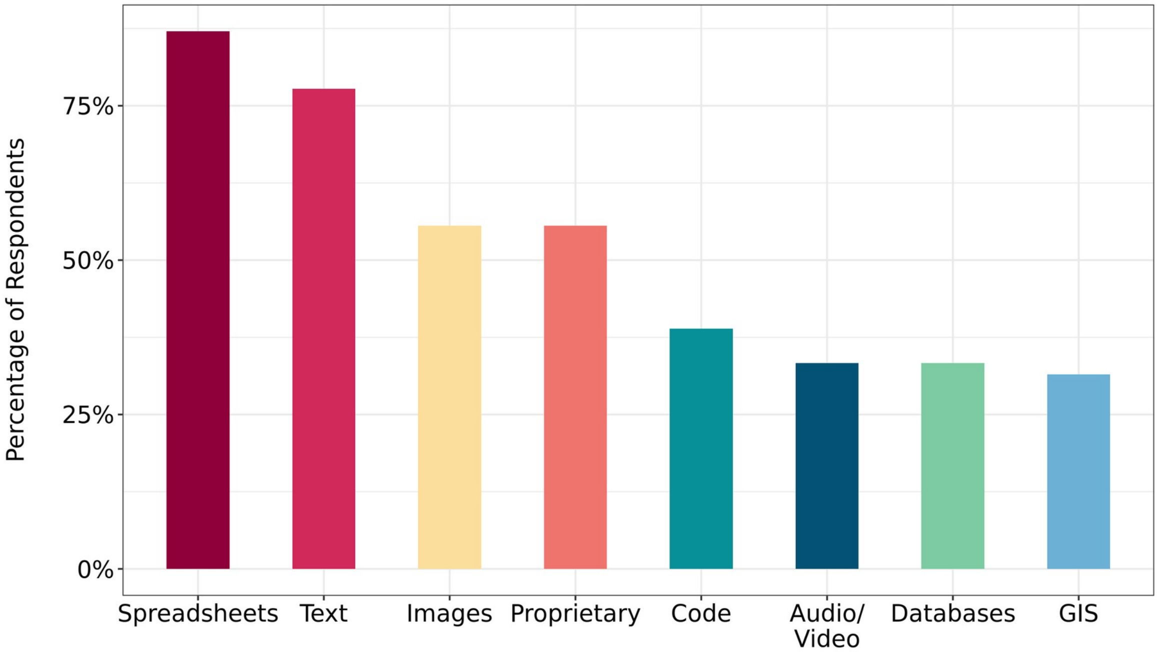


Figure 5.7: Types of data and other digital artifacts created by n=54 survey respondents.

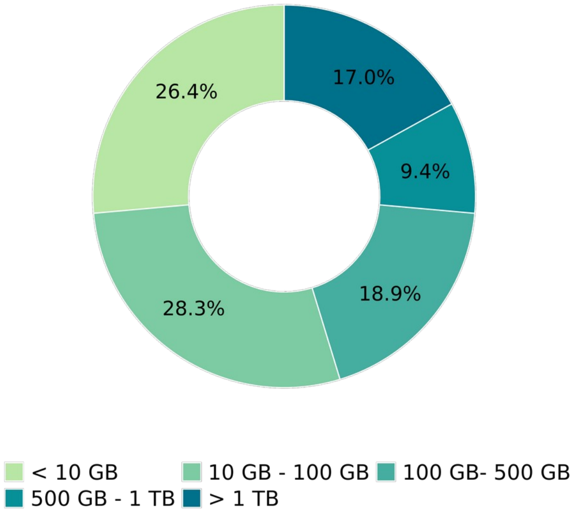


Figure 5.8: Per project digital storage requirements, on average, for n=54 respondents.

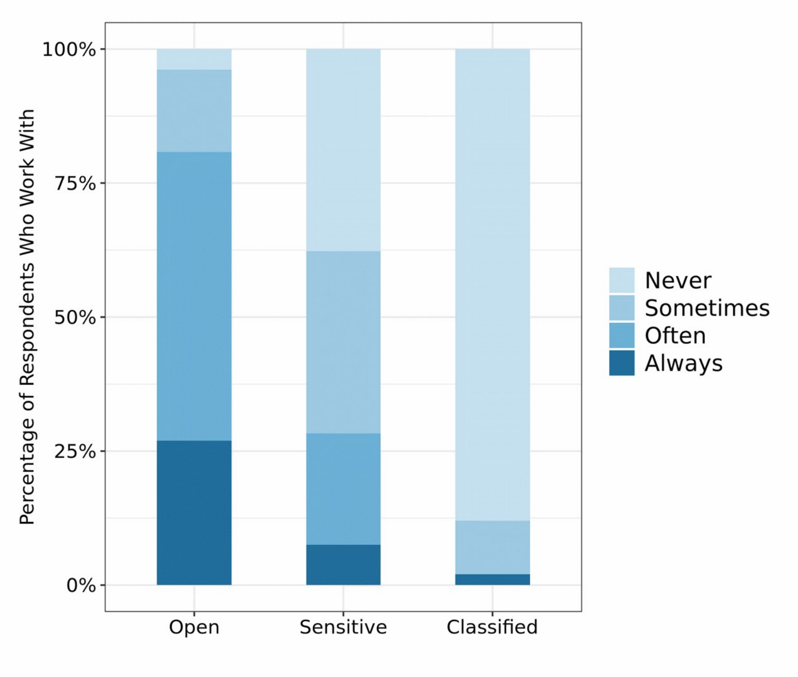


Figure 5.9: Frequency of working with different sensitivity levels of data, for n=54 responses.

these researchers perform. Due to its intensely regulated nature, researchers were also asked whether or not they worked with human subjects data specifically. Figure 5.10 shows that over 45% of researchers reported working with human subjects data.

5.3. Understanding of RDM Tasks

RQ1 asked, "How well do researchers understand how to perform common data management tasks?" Survey participants were presented with the following seven different activities associated with research data management, and asked how well these had been explained to them. These results are reported in Figure 5.11, with the bar plot headers corresponding to the following activities:

- **Create Metadata** - Creating metadata and documentation to describe research data.
- **Data Archival** - Archiving data in a repository for long-term storage.
- **Data Statements** - Including a Data Availability Statement in a manuscript.
- **Data Licenses** - Creating a license to describe how data may be used or reused.
- **Assign Identifier** - Assigning a permanent identifier (such as DOI) to datasets.
- **Data Mgt. Plans** - Creating a data management plan.
- **Data Citation** - Citing the use of data that you did not personally create

The data visualization makes it clear that there were a sizable proportion of respondents who felt the nature of activities had not been well-explained to them, across all data management tasks. Data citation and data management plans seemed to have a more balanced distribution

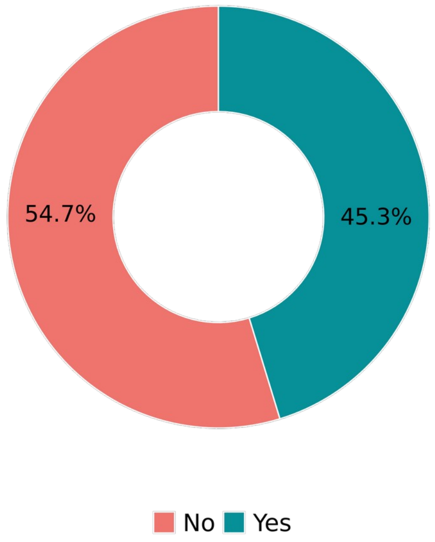


Figure 5.10: Percentage of n=54 respondents who report working with human subjects data.

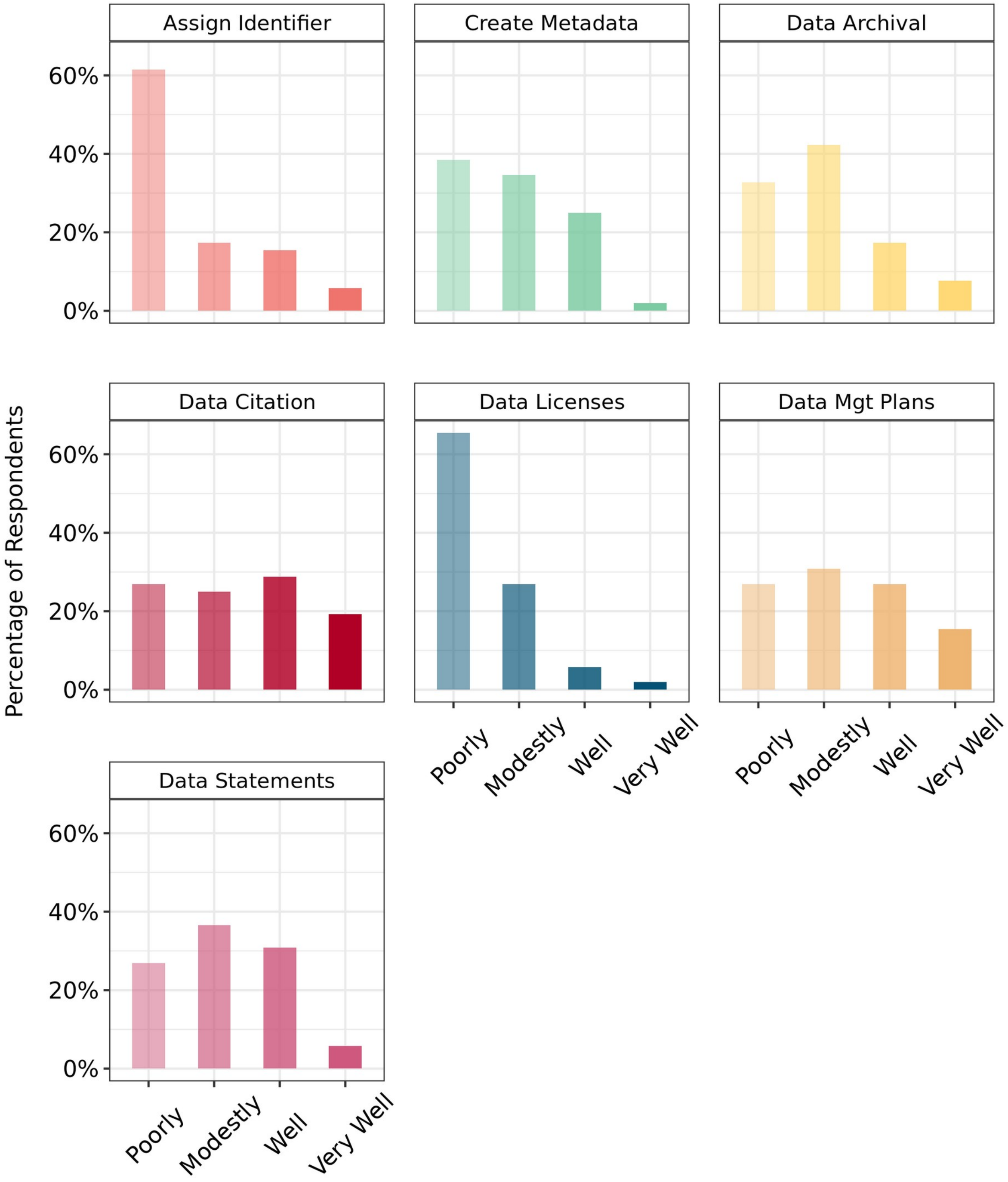


Figure 5.11: Responses to the question "How well has this been explained to you?" for different RDM topics, as a measure of personal understanding. There are n=54 responses represented.

among those who felt that the nature of these tasks had been communicated effectively versus those who did not. The concepts of data licenses and the assignment of permanent identifiers were particularly poorly understood. The data suggests that there are substantial knowledge gaps across the board for much of this community of researchers.

Researchers who indicated that they work with human subjects data were additionally presented with the following options, the results of which are presented in Figure 5.12. Facet headers for each bar plot correspond to survey items as follows:

- **Consent -** How to write an informed consent statement using language that that does not conflict with data sharing.
- **IRB -** Understanding of how secondary use of existing human subjects data affects the IRB review process.
- **Privacy -** Understanding how privacy concerns for human subjects data affects where data can be stored and how it should be protected.
- **Data Prep -** Knowledge of how to sanitize and otherwise prepare data from human subjects research for post-project archival.

For the human-subjects specific RDM activities, we see the first case in which the majority of participants felt they had been given a strong understanding of how to accomplish the task. Most respondents felt that how to satisfy the expectations of protecting human subjects data had been explained either well or very well to them. Implications of use of secondary data for

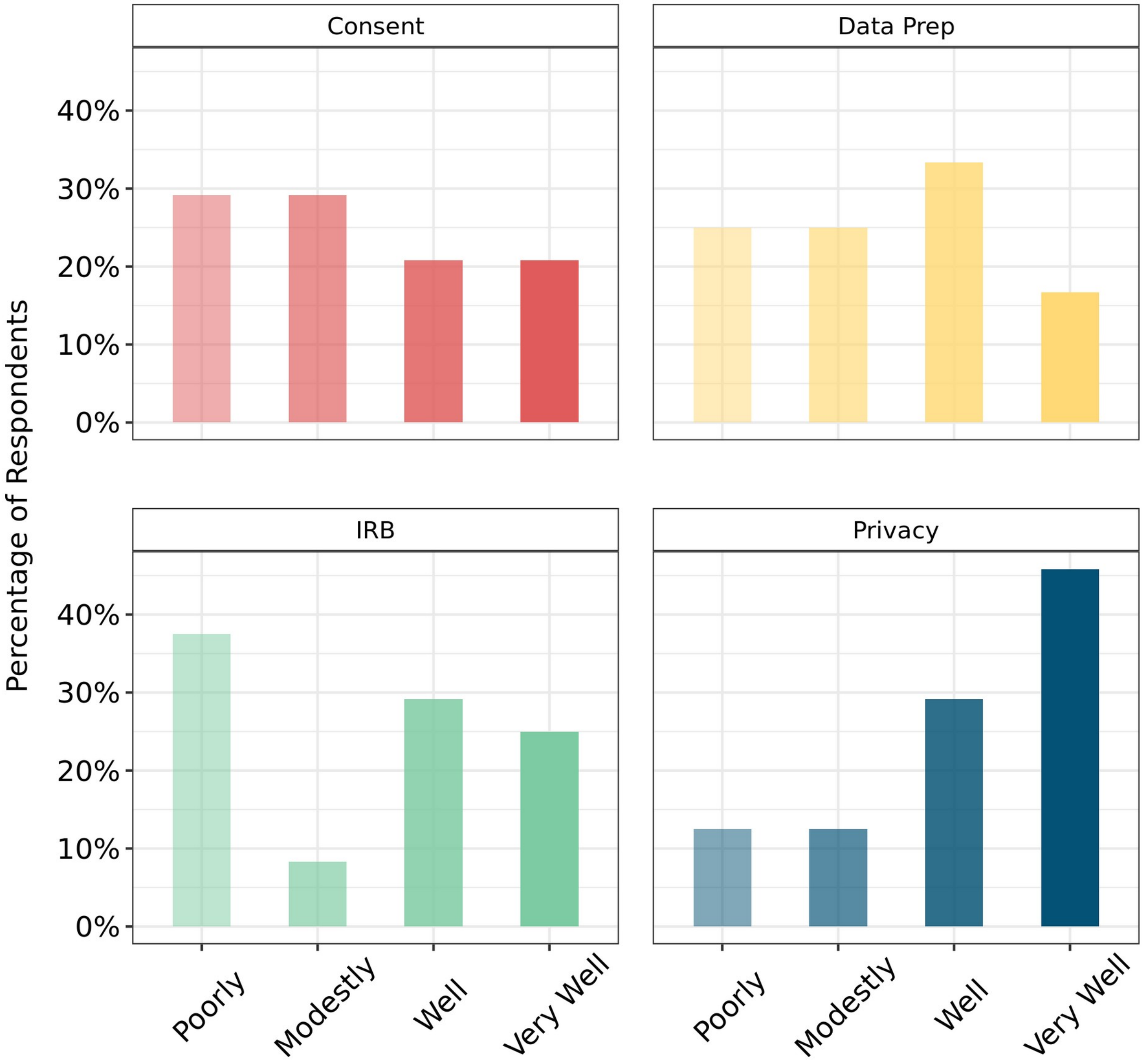


Figure 5.12: Responses to the question "How well has this been explained to you?" for RDM topics specific to human subjects research, as a measure of personal understanding. There are n=24 responses represented, limited to those who indicated they work with human subjects data.

human subjects protections tended to fall to either extreme of the scale. However, topics of consent and how to sanitize data for archival still have substantial room for improvement.

Finally, participants were also directly asked about how well they felt they understood the research data management policies and expectations of the funding agencies who support their work, if applicable. Figure 5.13 reports these results, in which approximately 41% of respondents indicated they felt they had a Poor or Modest understanding, whereas 59% felt they had a Good or Excellent understanding. Discrepancies between responses when the question is presented this way, versus when they are asked about specific kinds of activities that contribute to data policies, are discussed in the Discussion of results.

5.4. RQ2: Researcher Characteristics and RDM Knowledge

RQ2 asked, "Does level of understanding for RDM tasks differ, based on characteristics of the researcher?" The results for computing Fisher's exact test for the forty-two combinations of variables representing understanding of RDM activities and researcher characteristics are reported in Table 5.1. These characteristics include Funding status, research Domain, if they work with Human subjects or not, whether or not they reported being a Principle/Co-Investigator, their Position at the university, and whether or not they work with Sensitive data of any form. The null hypothesis for this test is that the relative proportions of one variable are independent of the other variable. Given the results that p $> .05$ in all cases, we fail to reject the null hypothesis for every test performed. The results support that, for this specific institution's community of researchers, none of the observed characteristics are associated with differences in understanding of RDM tasks.

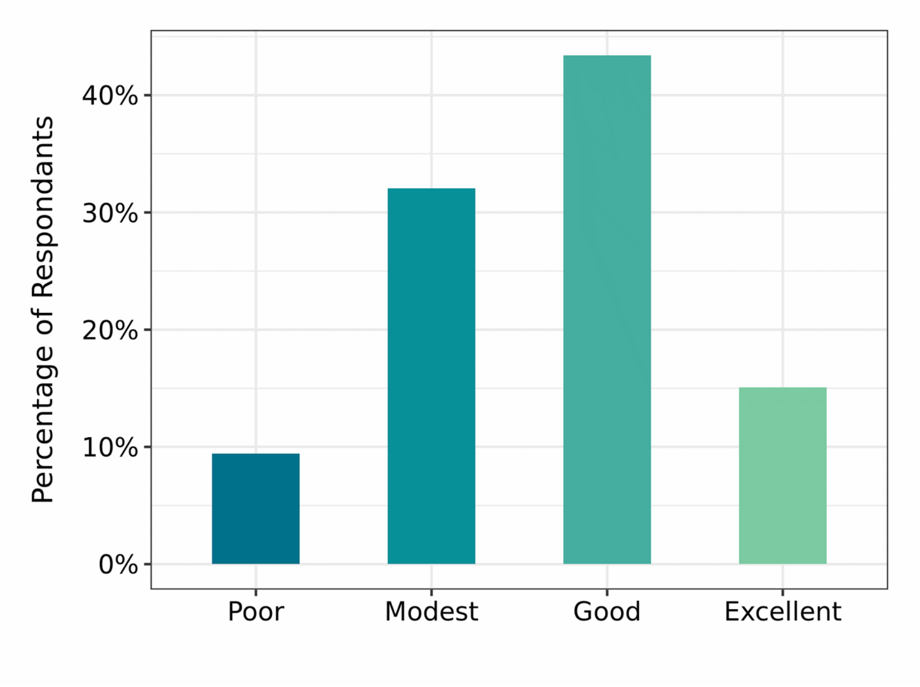


Figure 5.13: How well researchers felt they understood the data management expectations of relevant funding agencies (n=54).

Table 5.1: Computed p-values for Fishers exact test of independence between researcher characteristics and understanding of RDM topics (n=54).

Topic	**Funding**	**Domain**	**Human Subjects**	**Investigator**	**Position**	**Sensitivity**
Archival	0.317	0.778	0.513	1.000	0.692	1.000
Citation	0.662	0.077	1.000	0.228	0.492	0.776
DAS	0.643	0.347	1.000	0.692	1.000	1.000
Data Statements	1.000	0.245	0.779	1.000	1.000	0.576
Identifiers	0.571	0.479	0.155	1.000	0.187	1.000
License	1.000	0.948	1.000	0.124	0.157	1.000
Metadata	1.000	0.490	0.341	1.000	0.106	0.746

5.5. RQ3: Other RDM Barriers

RQ3 asked, "Which RDM obstacles are the most significant barriers to data policy compliance?" Figure 5.14 summarizes participant responses along a four-item, Likert-type scale for each of several overarching barriers. Although there is some contingent for which every obstacle is at least a moderate barrier, the time required to prepare data for archival appears to be the most significant issue.

5.6. RQ4: Institutional Support

RQ4 asked, "Where do researchers expect to find support for RDM tasks within their home institution?" To address this question, participants were provided the same list of seven RDM activities, and asked where they would be most likely to seek help for these tasks amount four standard administrative support units found in universities. Results are reported in Figure 5.15, and show a lack of consensus for where participants would expect to find support across all tasks. Libraries and Office of Research are the most common choices, though the Office of Technology is the first place that many would look for support when it comes to tasks such as data archival and the creation of descriptive metadata.

5.7. RQ5: Current Practices

Ideal data preservation outcomes are low in risk and high in openness. Data management options such as personal computers and external media tend to place university data assets at high risk for loss in the event of accidents, disasters, or equipment failures. Cloud-based storage, or servers paid for by departments and managed through the Office of Information Technology,

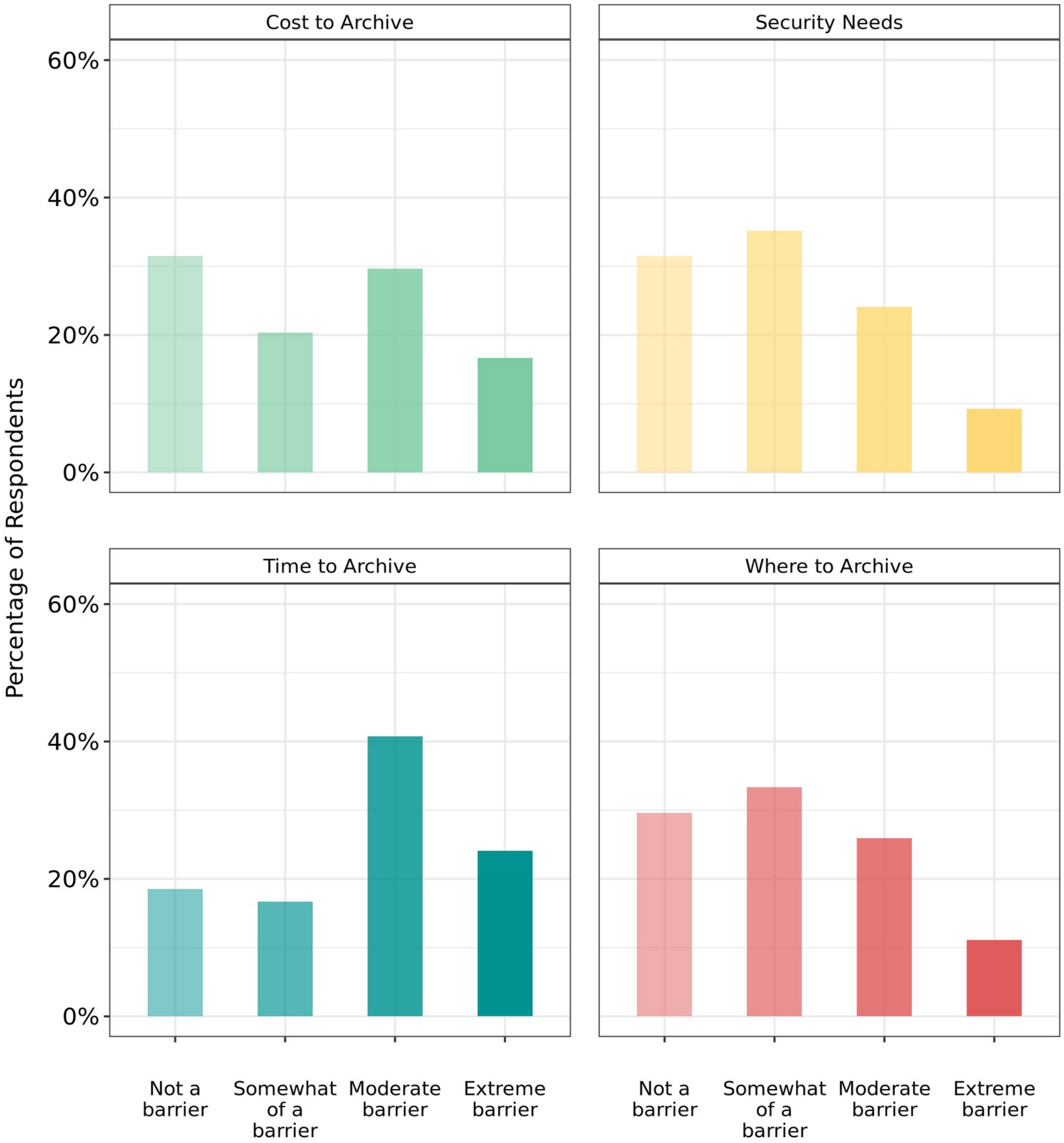


Figure 5.14: Researcher perceptions of the extent to which different obstacles are a barrier to data policy compliance. (n=54)

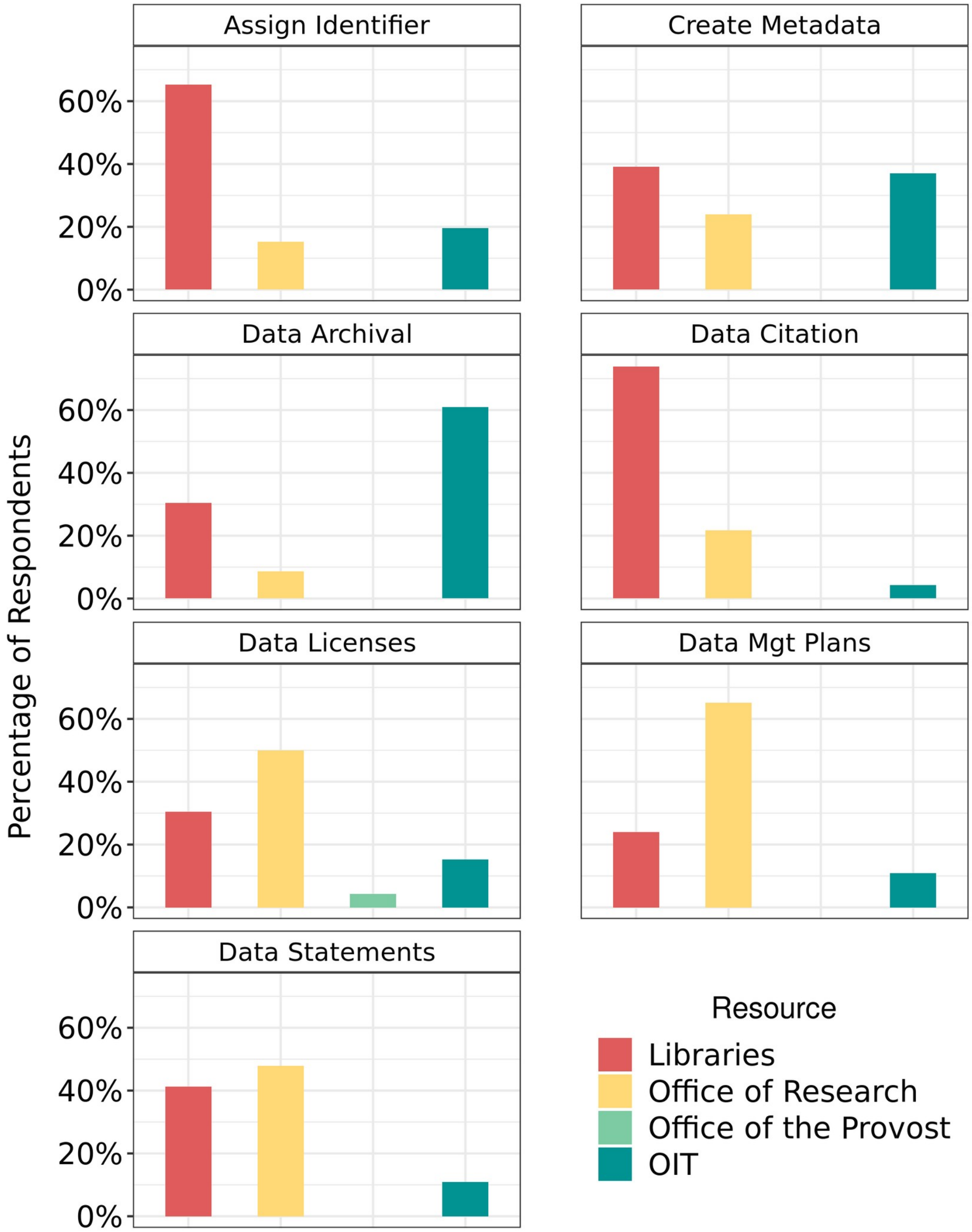


Figure 5.15: Institutional support units that researchers would be most likely to turn to for support regarding different RDM tasks. (n=54)

reduce risks to data assets through built-in systems for backup and redundancy. However, lack of access to the public, as well as absence of search and discovery features, mean these solutions fail the test of openness which has become a priority for funding agencies and many journal publishers.

As shown in Figure 5.16, most researchers who responded to this survey did not report using the types of low-risk and open data preservation strategies that would both protect university data assets and meet the data policy requirements of funding agencies or many publishers. There can be a variety of reasons for this, including a lack of options, time, funding, or understanding.

5.8. Results of ADAM Workflow for Publications-based Assessment

The pilot study consisted of n=87 NIH-sponsored original research publications whose first authors were affiliated with this same university. Only approximately 48% of these papers contained at least one statement speaking towards the availability of original research data whether structured or unstructured. Figure 5.17 describes the communication patterns about data sharing in more detail, for this subset of data. We see that formal, structured communication patterns are the more common approach, although there is overlap in that some publications feature both kinds of messages. Recall that structured messages are those which are visibly distinct, usually in the form of headers, and include labeled Data Availability Statements, Appendices, and references to data in the Supplementary Materials sections. Approximately 12% of data availability messages were unstructured, which is an important observation for characterizing the information that would be lost if unstructured messages were excluded.

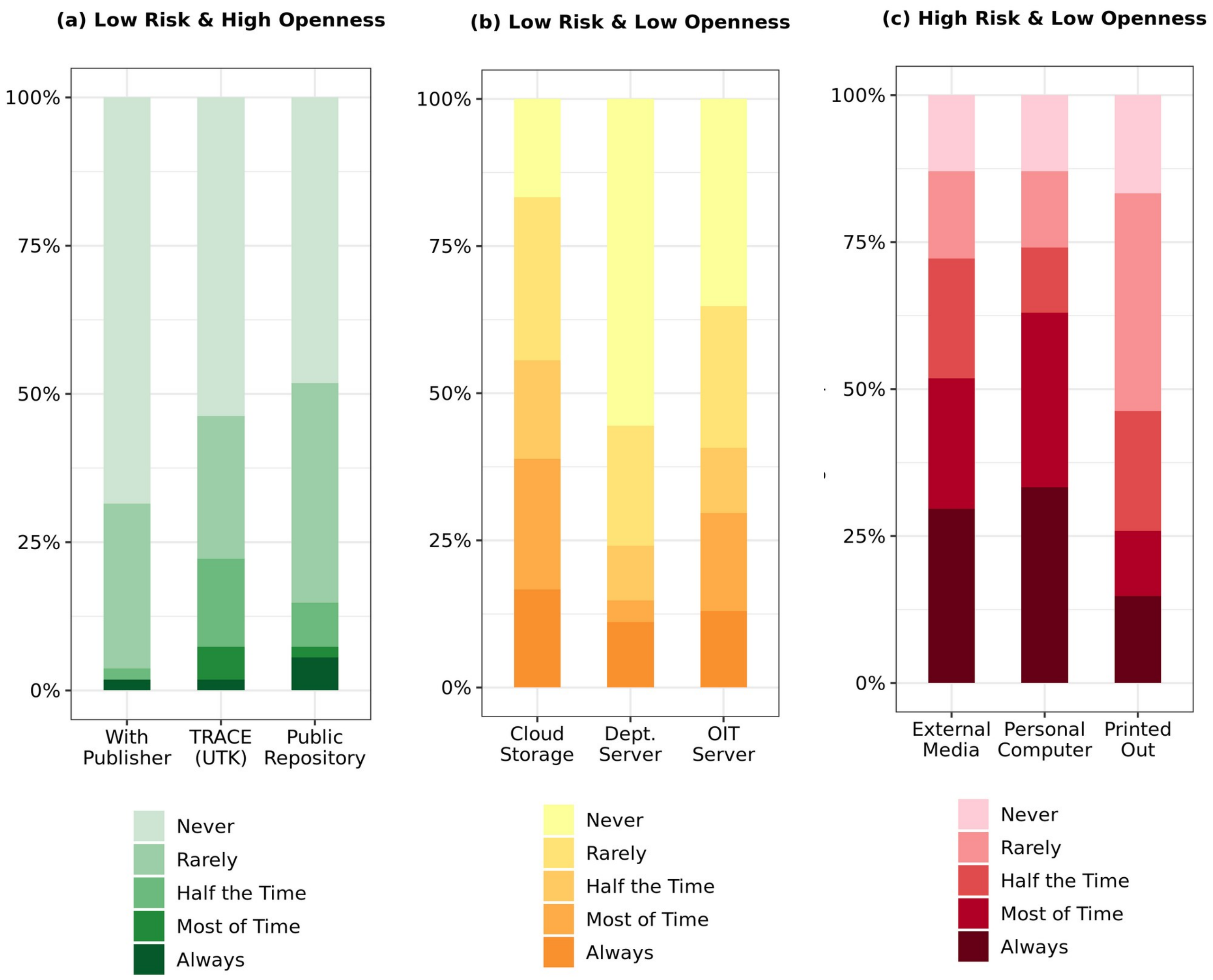


Figure 5.16: Distribution of researchers who engage in various data management practices, and the frequency with which they do so. (n=54)

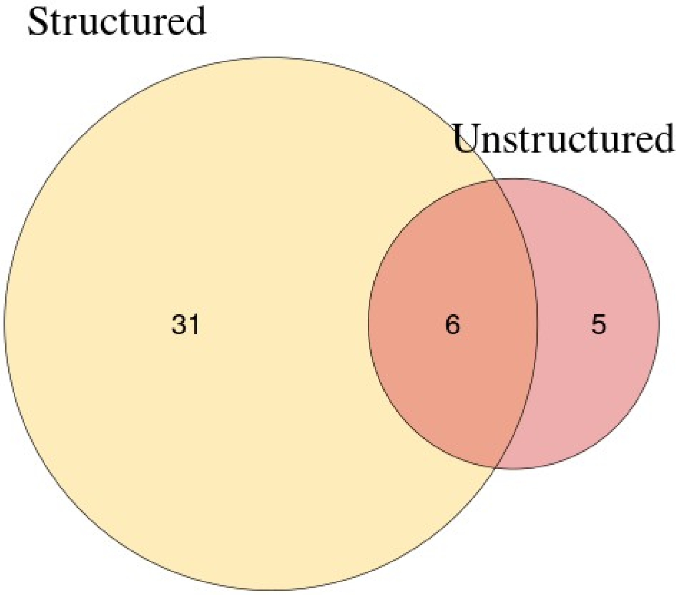


Figure 5.17: Distribution of unstructured versus structured data availability messages found among n=87 journal articles and conference papers.

Furthermore, formal citations to ones own original data were rare, with only 1 case observed.

Approximately 65% of the 85 journal articles came from journals that either described, encouraged the use of, or required the use of data availability statements in their information for authors. Fishers exact test was performed comparing the journal's status for communicating about data statements with whether or not a structured data availability message was found for original research data, with the results that $p = 0.0397$. Given that $p \leq .05$, the results are significant, suggesting that journals' choices to support the use of formal data messaging channels impacts the communication choices of authors.

Data citations were more common for papers using secondary data. Messages about where to find supporting secondary data were found in approximately 37% of the 87 publications, all of which used in-text (unstructured) messages to communicate about where they had gotten the data for the analysis. Only 4 papers also included a reference to the secondary data in a structured message about data availability, suggesting that authors may not feel data availability statements are as relevant for use of secondary data. Approximately 75% of messages about the origins of secondary data also included a corresponding citation in the references list, although these citations were most often to literature related to the data in some way rather than to the data as an artifact itself.

The distribution of data sharing strategies described in the contents of both structured and unstructured messages are described in Figure 5.18. Only original research data was coded for data management strategies. This data provides a complementary approach to answering RQ5, which seeks to describe the current data management practices of researchers. In interpreting

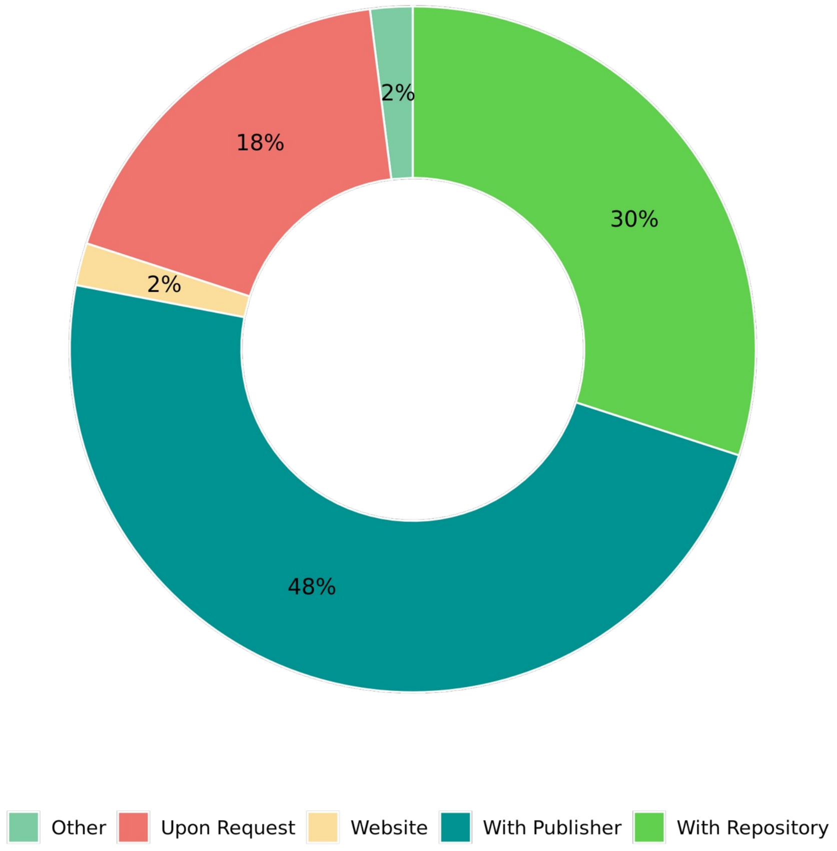


Figure 5.18: Distribution of 50 types of assertions about data availability across 42 messages (both structured and unstructured). Some messages contained multiple assertions.

this figure, note that multiple assertions about data availability may be represented for each message. For example, a statement may indicate that part of the data is available in the supplementary materials and that the rest of the data is available upon request. Both of these types of strategies would be represented for the paper in Figure 5.18. The majority of original data availability messages communicated that data were available without having to request it from the authors, such as through the publisher or with a repository. GitHub was the most commonly noted repository.

6. Broader Impacts

Triangulation between self-reported and publication data as an organizational assessment strategy is a new approach that has not been demonstrated by any other example of RDM assessment in the literature. The results from the Assessment for Data Availability Messages (ADAM) tool and the survey demonstrate two different approaches that organizations can use to understand the knowledge and behaviors of their researchers, as well as the extent to which these align with relevant data policies.

Variations will exist in how institutions adapt and respond to evolving data management and sharing requirements. However, the development of best practices for assessing and improving RDM efforts will help to inform and guide these efforts. The current studies report outcomes in the context of a pilot institution, but their broader impacts include how they advance the ability of research organizations to proactively measure and act on disparities between policies and practices. This section reviews the broader impacts of this study and their implications for organizational RDM planning.

6.1. Broader Impacts for Organizational Planning

Research data management support at an organizational level has not been well-explored in the literature. One of the most significant distinctions between the previous works and the present pilot study is that the current survey is the first to measure the information-seeking behavior of researchers in relation to formal support channels in academic organizations.

For a comparison, one assessment study probed the interest in RDM services from the specific perspective of the libraries (Whitmire et al., 2015). Another used a survey designed by a working group of campus-wide stakeholders with items asking about current data management practices and data sharing intentions in response to the NSF requirements for data management plans (Steinhart et al., 2012). However, the study did not discuss what the relationship between those campus partners looked like, and it did not inquire as to how the participants expected to engage with different stakeholders.

A survey of administrators from 209 research libraries across eight countries asked about the institutional dynamics of RDM. This was limited, however, to the perspectives of library decision-makers as to which administrative units, including the office of research, the office of technology, or the libraries, had participated in or guided RDM services and policy development at their respective organizations. There was no further inquiry into what aspects of RDM were supported by these units. Explorations of the relationship between organizational structure and specific RDM services were limited to the context of libraries (Cox et al., 2019). These studies do not adequately recognize the boundary-spanning nature of RDM support.

Changes in data policy represent a new opportunity for information needs to arise. Work-related seeking of information may be modeled as a situation in which information needs occur in response to specific work-related tasks that must be fulfilled in relation to one of the roles assumed by a professional. Complexity, degree of importance, and urgency are all examples of factors that may be relevant to how information is sought, and information seeking can occur through a variety of channels such as formal or informal and those internal or external to an organization (Leckie et al., 1996). Academic institutions customarily spread research support across multiple formal internal channels—each with its own budget, mission, and priorities.

The results of the present pilot survey revealed the uncertainty faced by researchers about seeking support for RDM topics, and they justified the inclusion of information-seeking behavior as a concept to be measured in organizational assessments. Other organizations can carry out this survey by using the open-source analysis software to process their results in order to better understand the researchers' uncertainty about seeking information for RDM support within their organizations.

Requirements such as those from the NIH put organizations in the position of needing to better understand the larger picture of how their researchers manage and share data. The results of the pilot studies show the value of these tools for providing this insight. The assessment outcomes should furthermore not be interpreted as a simple reflection of where and how RDM support is offered or even as a simple reflection of the researchers' expectations. Rather, the survey results will also be influenced by how well the organization has communicated about the availability of RDM support, and how RDM fits into the existing missions of these functional areas.

Examples of the boundary-spanning investigations of RDM support are rare in the literature. One exception is the qualitative, comparative analysis of role-based responsibilities for the research integrity of officers and libraries that explored in-depth both the overlap and distinctions between these two functional areas (Bishop et al., 2021). This study positioned both as they related to stakeholders in research data management. The broader impact of the current study, which factors organizational channels into information-seeking behaviors, will be that organizations have tools to better understand the need to develop and communicate clear RDM support strategies beyond a single functional area. The most appropriate model of support may vary from one institution to the next, but the survey piloted in this study makes a significant contribution to testing the alignment between actual service models and community understanding.

Consistency between existing institutional support priorities and those related to RDM would provide some advantages for organizational messaging. In practice, the actual data-related work performed by individuals with comparable job titles and responsibilities in different support units can be as varied as the institutions they come from, given the examples of librarians (Bishop et al., 2022) and the research integrity officers (Bishop et al., 2021). The establishment of how these roles and responsibilities should fall within an organization, whatever that may be, is the necessary prerequisite for both developing and communicating about support for research data.

Figure 6.1 illustrates a potential strategy for RDM support across typical administrative units. Data discovery and data preservation activities are natural extensions of libraries' traditional roles as the finders and custodians of information. Data preservation activities include such tasks as assigning digital object identifiers, creating metadata, and evaluating archival options. While many libraries retain sufficient technical staff to manage the institution's repositories for publications, the complexity and scope of research data management may be better placed within the actual underlying technological infrastructure management in campus technology departments. While some libraries also retain a number of staff with data science and analytical expertise, information technology departments typically have dedicated staff for research computing and data analysis. Given the expertise that technology departments maintain

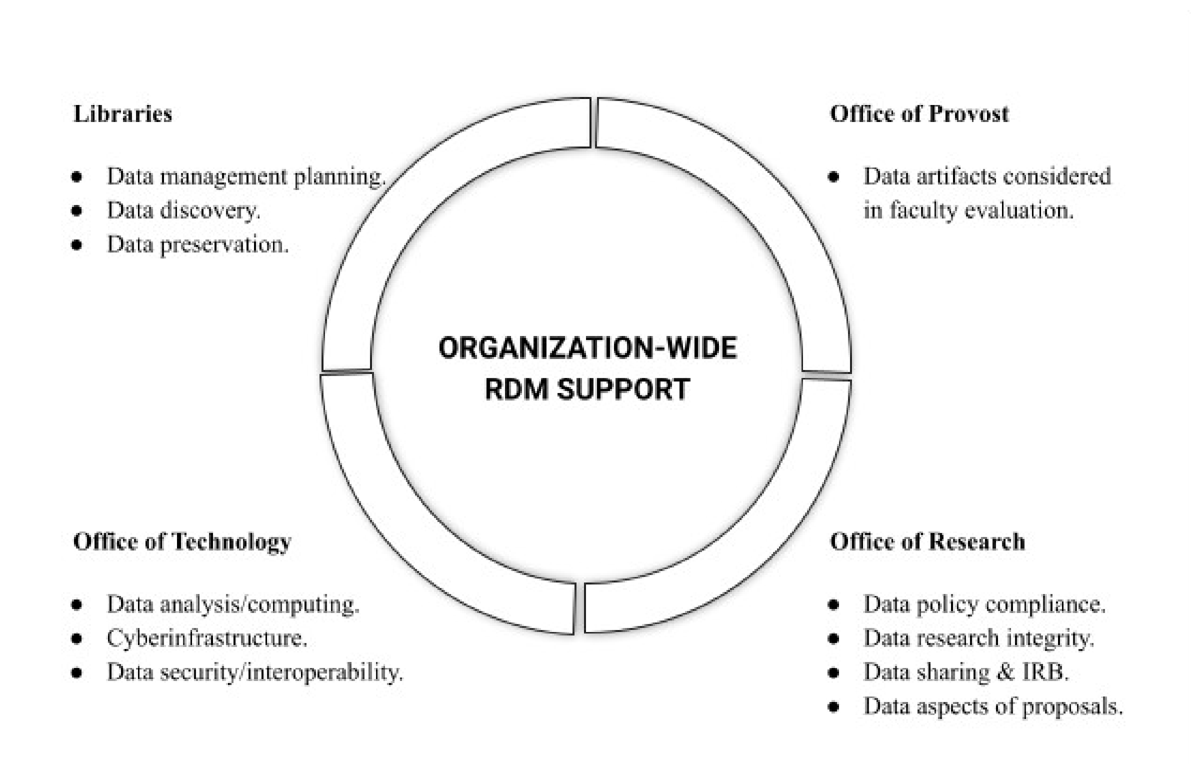


Figure 6.1: An example model of research data management services, organized across functional support units which are common to academic research institutions.

with systems' interoperability and security standards, they also would be valued partners in evaluating technical specifications and the appropriateness of various external archival options.

Data management and sharing is also a consideration for roles within the offices of research. As the front-line of communications with funding agencies, there will be an immediate need for understanding the requirements of data policy compliance. Many cases of data-related research misconduct have been identified, making knowledge of data management a priority for Research Integrity Officers (Bishop et al., 2021). Institutional Review Boards' (IRBs) responsibilities include protecting human participants. Therefore, IRBs will deal with competing priorities when maximizing data sharing for human subjects. For example, IRBs may need to help researchers understand how to modify their informed consent templates and data management processes to avoid conflicts with data policies.

Technology departments may provide guidance on how to ensure sensitive data are secured, and that appropriate licensing and access restrictions are in place. Data management plans supporting proposals are often handled within the libraries, but the offices of research should ensure that all proposals appropriately budget for any costs for preservation and sharing, when such costs are allowable. Even the Provost's Office has a role to play and is represented in this model, given that archived data products should be considered as research outputs in faculty evaluations.

In short, data management and sharing cut across every part of the research lifecycle and all areas of research support in an academic institution. A clearly defined organization-wide strategy will help institutions prepare for data-policy compliance and facilitate their ability to communicate to their internal stakeholders.

6.2. Broader Impacts for Benchmarking Progress

For both methodological approaches, a significant contribution to the field is that they have been developed in such a way as to support benchmarking for research data management practices over time. Benchmarking is defined as, “A continuous, systematic process for evaluating the products, services, and work processes of organizations that are recognized as representing best practices for the purpose of organizational improvement,” (Spendolini, 1992, p. 9). More specifically, internal benchmarking describes activities designed to measure the internal operational performance of organizations (Spendolini, 1992, p. 16).

Benchmarking supports the ability of organizations to internally identify performance gaps and determine the best practices for addressing said gaps. External dimensions of benchmarking involve the consideration of metrics and best practices from other organizations, both within the same industry and in other domains (Yasin, 2002). Practices are those repeatable processes and their outputs that may be selected for benchmarking, and metrics operationalize practices so that they are measurable (Camp, 1989, p. 251).

The survey data and the ADAM workflow provide two different approaches to operationalize the measurements of research data management practices. The ADAM workflow is particularly effective as a measure of data practice outcomes, given that research projects culminate in the dissemination of findings. Self-reported data from the survey enriches the organization’s understanding of the dimensions that contributed to RDM practices and outcomes by providing specific targets for improvement. Together, the two-pronged benchmarking approach, using both self-reported data and objective publication metrics, provides tools for research organizations to engage in continuous improvement. An organization can directly measure efforts to shrink the knowledge gap for the variety of RDM topics addressed in the survey. An improved alignment with FAIR data principles could be directly measured from repeated collections of self-reported data and publication-observed data at later points in time.

The open-source publication of an analytical code for translating data into findings is a substantial contribution to the broader impacts of this dissertation. Analyzing data collected from assessment studies is time-consuming, and it is difficult to ensure that a consistent process is used each time the study is repeated. However, the software published with this study facilitates the ability of organizations to repeat both exercises for benchmarking purposes, without inconsistencies or additional labor to translate the data into findings. Table 6.1 summarizes the open-source software that has been developed for translating data into findings for survey and publication metrics. All software has been developed using R, which is an open-source statistical computing language.

The results of the two studies function effectively as benchmarking metrics for measuring internal progress. However, the methods are also transferable across institutions, and they would provide valuable insights into the establishment of external benchmarking practices at the industry level. The ADAM tool is particularly useful as a process for third-party researchers interested in tracking industry-level changes, as it uses widely available inputs through institutional subscriptions or repositories, does not require active participation from the organizations being studied, and does not qualify as research on human subjects. These tools are future-proof, given that history has shown that data policy adoption continues to pick up momentum across funding agencies and publishers as a whole. The ADAM workflow will continue to be effective, independent of the rate of adoption of data policies by publishers. The

Table 6.1: Open source software files published for processing data collected through the dissertation survey assessment instrument and through the ADAM workflow for publications.

Product	File	Description
Fig 5.7	Q1.DataCreated-Bar-Chart.R	Types of data collected by participants.
Fig 5.8	Q2.StorageSizes-Pie-Chart.R	Pie chart for project data storage sizes.
Fig 5.9	Q3.Sensitivity-Bar-Chart.R	Stacked bar plot - data sensitivity levels.
Fig 5.6	Q4.Decisions.Pie.R	Data decision-making status.
	Q5.Long-Term-Storage-Bars.R	Good, bad, & mediocre data practices.
Fig 5.11	Q7.Skills-Facet-Bar-Plot.R	7 plots for understanding of RDM skills.
	Q8.Resource-Bar-Chart.R	4 plots for seeking support from org.
	Q9.Barriers-Facet-Bar.R	4 plots describing perceived barriers.
Fig 5.10	Q10.HumanSubjects.Pie.R	Proportion working with human subjects.
Fig 5.12	Q11.HumanSubjects.Faceted.R	4 plots – human subjects RDM topics.
Fig 5.2	Q13.Funded.Pie-Chart.R	Proportion of research which is funded.
Fig 5.5	Q14.PI.R	Distribution for status as PI.
Fig 5.13	Q15.Funder.Policies.R	Knowledge of funder policies.
Fig 5.3	Q16.Position.Pie-Chart.R	Distributions of participant positions.
Fig 5.4	Q17.Tenure.R	Tenure status for faculty participants.
Fig 5.1	Q18.Domain.Pie-Chart.R	Distribution of domains.
	custom_styles.R	Uniform styles imported into all plots.
	Preprocessing.R	Clean and prep survey data from online software format into standardized format.
Table 5.1	Fisher_Sensitivity.R	Fisher test: sensitivity & 7 RDM topics.
Table 5.1	Fisher_Position.R	Fisher test: position & 7 RDM topics.
Table 5.1	Fisher_Investigator.R	Fisher: Investigator status & 7 RDM topics.
Table 5.1	Fisher_HumanSubjects.R	Human subjects status & 7 RDM topics.
Table 5.1	Fisher_Funded.R	Fisher test: Funding status & 7 RDM topics.
Table 5.1	Fisher_Domain.R	Fisher test results: Domain & 7 RDM topics.
Table 5.1	Merge_p-values.R	Combine all Fisher tests to one table.
All ADAM Results, Fig 5.17, 5.18	Publications.R	Statistics, Venn diagrams, and plots summarizing ADAM data collection outcomes.

assessment tools demonstrated in this dissertation give organizations the opportunities to respond proactively instead of re-actively to the changing environment of research data management and sharing.

6.3. Broader Impacts for Understanding Communication Patterns of Researchers

The ADAM tool has demonstrated its value for organizations in assessments and benchmarkings. However, the broader impacts of a publication-based evaluation metric include its potential to better inform our understanding of how researchers communicate about data availability as a whole. It was determined for the test case that messages about data availability most often took place in the context of structured messaging opportunities, such as data availability statements (DAS) and supplemental data sections. The fact that publishers are giving researchers formal channels through which to communicate about data availability is well established by existing studies and was not the scope of the current work.

Other studies have thoroughly established the great variety of ways that publishers communicate data policies (Dosch & Martindale, 2020; Grant & Hrynaszkiewicz, 2018; Huh, 2019; Tal-Socher & Ziderman, 2020), the extent to which institutions may be affected (Dearborn et al., 2017), and the confusion they may cause for authors (Christian et al., 2020).

Examples of different publisher language about data availability statements were reported in Table 4.2. Some publishers may indicate that data statements are “required,” while others “recommend,” or “encourage” them. Authors will, in some cases, be tasked with determining if the phrase “should be included” is synonymous with “must be included.” Furthermore, they will have to determine if a description of what a data availability statement (see *International Journal* *of Molecular Sciences*) is intended to implicitly mean that the journal *requires* a statement—or that the journal is letting the author know that data availability statements are an option. In other cases, authors may have to determine if their research data meet any of the criteria for specific conditions in which data sharing or a DAS is required, whereas these are otherwise optional. All of this is to say as the inclusion of data availability statements is supported by a wide variety of journals, but these journals can differ in the strength, scope, and language of their expectations.

The ADAM workflow’s unique contribution to the field of scholarly communication is that, by ensuring even if messages buried in the text of a manuscript are included, ADAM can characterize if, and how, researchers engage in messaging about data availability *independent* of the formal channels provided by journals regarding data availability statements. The second study showed that researchers only communicated about data availability outside of those formal, structured channels about 12% of the time. This provides a serendipitous insight into the importance of journals making these formal messaging opportunities available, despite this not being the objective of the study.

Another implication of characterizing structured versus unstructured messaging is that it gives scholarly communication researchers and organizational assessors a context for their methodological choices. Existing studies that have examined research data availability messages as artifacts of a researcher’s behavior have strictly examined DAS for publishers or journals, and deliberately excluded papers that did not have a structured DAS. This approach is partially a sample of convenience, as structured messages with specific headings can easily be parsed by both humans and machines. For example, the distributions of the types of DAS have been manually analyzed and reported for specific journals, including a selection of nature journals (Grant & Hrynaszkiewicz, 2018) and PLOS ONE (Federer et al., 2018). A machine-automated analysis and classification of 124,000 DAS from various journals of the publisher Wiley was made possible by the structured and parsing-friendly nature of formal DAS (Graf et al., 2020). In all of these cases, the studies describe data sharing outcomes specific to journals or publishers.

However, those are not relevant contexts for organizational assessment applications. If an organization chooses to limit its assessment to only journals or articles that contained a formal structured data availability statement, then they should at least understand the extent to which their methodology would introduce a bias into the findings. In this case, approximately 12% of data sharing messages for original research occurred outside of those formal, structured messages. Not only is ADAM a novel approach to using publications as an assessment tool for characterizing research data management practices specific to an institution, it is also unique in that it presents a workflow that does *not* exclude data messaging outside of formal, structured statements. Both of these points make significant contributions to the broader impacts of the study.

7. Conclusions

This dissertation piloted two assessment approaches available to academic research institutions that wish to better understand their current alignment with data management and sharing principles featured in policies such as the NIHs. This understanding can help institutions identify the most-needed services and training.

7.1. Discussion

The pilot study's results demonstrated several opportunities and concerns for guiding RDM strategies moving forward within this institution. For example, over 50% of the participants reported working with proprietary data formats (requiring specific software or versions to access). With proprietary data formats, there is risk of even archived data being locked into unusable formats (Patel, 2016). The scope and size of data produced must also be factored into any institutional research data management strategy. In this case, over 45% of respondents reported research data storage requirements beyond 100 GB. This means a number of the current free or low-cost data archival solutions available to researchers, independent of an institutional subscription, are eliminated. For example, at the time of this dissertation, the figshare repository places a limit of 20 GB for free personal accounts (https://help.figshare.com /article/how-to-upload-and-publish-your-data) and Mendeley's data place the gap at 10 GB (https://data.mendeley.com/faq). Another repository, Dryad, charges a $120 data processing fee for submissions up to 50 GB, beyond which additional fees are applied that scale with the size of the data. These repositories will no longer accept sensitive human subject data, which over 45%

of participants indicated was a factor into their work. Furthermore, local institutional repositories that accept data must also consider the size and scope of what they are prepared to manage and for how long among other considerations as part of a digital curation strategy (Higgins, 2008). In short, the data from this survey will help institutions to understand and interpret the options available to their researchers for complying with this data policy.

The results from RQ1 suggested a need for better information and training for many researchers on data management topics relevant to compliance with data policies. It is interesting to observe that, on the one hand, over 90% of the participants who perform funded research felt they had at least a modest understanding of the data management requirements of their funding sources. On the other hand, when the participants were asked how well they felt specific data management and sharing tasks had been explained to them, their responses indicated substantial knowledge gaps across each topic. One explanation may be that the wording of the first question was more vulnerable to a social desirability bias in that it asked about knowledge in terms of the individual’s personal understanding.

However, the concept of data management plans, at a high level, was reported to be one of the better understood RDM activities, even if the survey participants did not feel as strongly about their understanding of the kinds of specific RDM tasks that might be expected to go into a data management plan. An alternative explanation may be that there has historically not been either a reward or a consequence as a result of data management practices beyond personal motivations. Recalling that one study found no differences in the strategies of funded versus unfunded proposals (Mischo et al., 2014), and noting that there have up until now not been any known examples of compliance enforcement in effect by a major U.S. funding source, there may be a simple absence of meaningful feedback.

Still, it is unlikely to be a coincidence that that the participants had a better understanding of DMPs, when the submission of was DMP is already required as part of the data policy of many funding agencies. The lower severity of knowledge gaps for both DMPs and human subjects' privacy through IRBs may be a form of success resulting from institutional factors, such as educational programs, support services, and organizational messaging regarding these topics.

In terms of how to address RDM knowledge gaps, the results organized under RQ2 were informative. No distinguishable relationships were identified between the various characteristics of researchers and their RDM knowledge outcomes. Thus, faculty fared no better or worse than their graduate students. The research domain did not alter the outcomes. In fact, none of the six dimensions of research attributes showed a statistically significant relationship with the understanding of the seven RDM tasks.

Based on these findings, it should be understood that any RDM training and service strategy must be broadly applied and available to all persons performing research. Perhaps as a reflection of these knowledge gaps, the survey results in response to RQ5 suggested that many UTK researchers are not currently engaging in practices that adequately preserve data, ensure the openness of data, or meet the standards of common data policies.

Circling back around, the evidence suggests that much of the community lacks knowledge of how to carry out many of the tasks that contribute to data management and data policy compliance. Furthermore, the results in response to RQ4 suggested that there is an observable uncertainty about where to seek support within an academic research institution for carrying out these activities.

ADAM, or the Assessment for Data Availability Messages, triangulated the survey findings as a tool to reliably extract information about data sharing practices from researchers communications. Using the example of a subset of NIH-supported publications, it was found the researchers frequently did not choose to communicate about their data's availability.

While the exact rubric through which the NIH will judge such efforts is as of yet unknown, we do know that future NIH-sponsored research will entail expectations that data are disseminated with other findings and that the justifications for any actions to the contrary are communicated. In terms of practices within the proportion of researchers who did communicate about data availability, the most common choice for data sharing was with the publisher, such as through supplementary files. 18% of the messages indicated that the data would only be available by request.

This established practice of data gate-keeping by investigators will be, in all likelihood, one of the more interesting disruptions to watch unfold. We do not know why the authors made this choice. It could be they did not feel motivated to engage in data sharing, but it is also possible that the data are of a scale, complexity, or sensitivity that rules out preservation and sharing as supplementary materials with a publisher or through general purpose repositories. But if the reason is based on motivation, personally vetting or denying access to research data appears at face value to be incompatible with upcoming NIH policy changes that state, "shared scientific data should be made accessible as soon as possible, and no later than the time of an associated publication, or the end of the award/support period." The policy further requires that the sharing of scientific data be maximized though "acknowledging certain factors (i.e., legal, ethical, or technical) that may affect the extent to which scientific data are preserved and shared," (U.S. National Institutes of Health, 2020)

However, the language of the policy does not qualify data sensitivity considerations as an exemption to data sharing. Rather, it elaborates, "that limitations on subsequent data use should be communicated to individuals or entities (e.g., data repository managers) that will preserve and share the scientific data" (U.S. National Institutes of Health, 2020). One interpretation of this language could be that sponsored researchers will be expected to take the necessary steps to sanitize and secure sensitive data so that it may be shared, rather than circumventing the sharing of data altogether.

7.2. Limitations

There are several limitations to this survey study that should be addressed. Most notably, the small sample size threatens the power of the study. Though the exact number of research faculty at the time of survey administration is unknown, current estimates per Incites' research evaluation platform put the studied university at 1,654 researchers. The 54 complete responses put this estimate at a little over 3% of the target population. The low number of responses could be a consequence of inadequate distribution channels, but another explanation may be that the numbers reflect survey fatigue among members of the target population. Survey fatigue in this context is defined as a phenomenon in which individuals refuse to complete surveys at all as a consequence of being approached too frequently for survey participation in too short a time (De Koning et al., 2021).

The responses also lacked diversity in terms of position in terms of position within the university. The graduate student response rate was less than optimal. Most responses (80%) came from faculty, and the majority (65%) were tenured. There was considerable overlap between the proportion of responses from principal investigators (82%) and RDM decision-makers (79%). While one may fully expect that principal investigators will be making research data management decisions, it would have been useful to get a better sense of whether or not other types of roles such as graduate students, post-doctoral associates, and other kinds of research staff found themselves making data management decisions. On that note, it was considered in hindsight that there would have been some benefit to offering a greater distinction between the types of roles. Future dissemination of the study should consider adding a postdoctoral research associate as a position type and modifying the processing script accordingly.

Th limitations of the principal ADAM tool were largely by design, trading depth, and principal breadth in exchange for reliability and reduced burdens. By limiting the assessment to text statements that explicitly contain the word “data,” some messages about data availability may potentially be missed. For example, if at some point the researcher used the word “files” instead of data, then the message would not have been flagged for review. While it is possible that the word “files” in some context might be intended to refer to data, it could also refer to any number of other things. The methodological decisions in this case were prioritizing the reliability and avoiding the need for lengthy investigations into the context of an ambiguous word’s intended usage, especially in cases where subject matter expertise may be needed. The potential effect of missing data could be statistically explored in a future study.

Regarding the ADAM workflow, it was observed that 65% of journals for articles in this dataset described data availability or transparency statements in their information for authors. For context, different kinds of policies about stating the availability of data were shared. Classifying whether data statements were required versus encouraged was outside the scope this analysis, and communicating directly with journal editors may be a more reliable and current source of this information than what is presented on their websites. However, an important limitation for interpreting the current results is that some journals do require these statements as a condition of publication. Therefore, the existence of structured, formal data messages should not be construed to represent authors' choices about whether or not to communicate about data. Differentiating between structured and unstructured messages is useful for understanding the distribution of messages about data availability among formal and informal messaging. This, in turn, informs understanding of the bias that would be created by not considering the full texts in the assessment methodology.

Otherwise, the limitations of ADAM are largely as stated in the Methods section of the dissertation. We are only able to discern knowledge of post-project data management practices for certain categories of data sharing messages. For example, if an author states that data are available in a well-known repository, we can make inferences about data risk and openness based on information about the repository. If an author states that the data are available "upon request," this could mean that the data is stored on a cloud service, sitting on a departmental server, or tucked away in an external drive under the principal investigator’s mattress. However, there are many things we can learn from this information, such as trends in the communicative behaviors around data sharing and in the types of sharing strategies used.

A final limitation for both the use of ADAM, and the survey is that the data are only relevant for the institution targeted in the pilot study. It would be expected that results may vary in either case for other institutions, depending on their progress and strategies for research data management. Training, availability of support services, cyberinfrastructure, and organizational communication strategies are all factors that might influence the results for a specific institution.

7.3. Conclusions and Future Directions

With major funding agencies and publishers becoming ever more serious about data management, and with compliance enforcement on the horizon, it becomes increasingly critical for academic research institutions to define their strategy for supporting data management.

This dissertation has made two major contributions in support of that goal. First, a survey instrument has been designed that extends current RDM assessment practices by including items that factor in the sensitivity of the data, items that expose knowledge gaps about specific actionable tasks that comprise RDM, and items that reveal organizational communication challenges in terms of a researcher’s uncertainty about institutional support. Second, the ADAM tool has been developed as an analytical workflow and measurement system for extracting knowledge of how researchers communicate about and share research data in the context of publications. The ability to perform an ADAM-based analysis independent of researchers’ participation is an extremely important benefit. In both these cases, the open-source publication of a library of all data visualization and processing scripts needed to interpret the results is yet another significant contribution.

For the academic research institution examined in this study, the findings can be used to guide the development of training support services and internal communication strategies to address RDM knowledge gaps. In defining research data management competencies, it is recommended that institutions think in terms of the specific tasks that must happen to achieve good data management and sharing principles, and how the execution of those tasks might differ depending on the circumstances of the research. A detailed requirements checklist, for example, may help researchers to plan for how they will address all the expectations enumerated within NIH guidance.

The ADAM workflow and the survey instrument are assessment tools that provide different views of current readiness for RDM compliance. Either analysis can be repeated in the future to benchmark changes in behavior over time. Future research should administer both research instruments at other institutions to learn more about how they may differ along these institutions. Differences in results could help research institutions learn from each other and pinpoint successful strategies.

Independent of the needs of a specific institution, the ADAM tool provides some of the most interesting results to be found in the dissertation. Communication patterns about research data are currently understudied. The quantitatively identified patterns in how researchers communicate about data are broadly useful to the information science community. Further analysis should be conducted across different populations and domains to develop additional useful findings. More research that investigates the motivations behind specific choices would advance our understanding of the relationship between motivations, available resources, personal knowledge, and outcomes.

Other potential opportunities for future research include the analysis of DMPs submitted as part of proposals to the NIH after the new data policies take effect in January 2023. Historically, studies that examined National Science Foundation proposals did not find meaningful differences between the DMPs of funded versus unfunded proposals (Mischo et al., 2014). Wide differences were found in the DMPs of winning NSF proposals (Parham et al., 2016). However, these studies are several years old at the time of this writing, and more current data would be needed to determine if the findings still hold true. Furthermore, the NIH is the first to implement a formal compliance mechanism and this indicates the agency’s intention to be taken seriously.

In summary, the world of research data management is evolving rapidly. With tools such as the survey demonstrated in this dissertation, and the ADAM tool, academic research institutions have multiple information sources at their disposal for better understanding their needs and developing targeted strategies to enhance their research data management.

List of References

Akers, K. G., & Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Digital Curation*.

Axson, S. A., Mello, M. M., Lincow, D., Yang, C., Gross, C. P., Ross, J. S., & Miller, J. (2021). Clinical trial transparency and data sharing among biopharmaceutical companies and the role of company size, location and product type: a cross-sectional descriptive analysis. *BMJ Open*, *11*(7), e053248.

Baker, M., & Penny, D. (2016). Is there a reproducibility crisis? *Nature*. https://doi.org/10.1038/533452A

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533.

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the Data Deluge. *Science*.

Berman, E. (2017). An Exploratory Sequential Mixed Methods Approach to Understanding Researchers' Data Management Practices at UVM: Integrated Findings to Develop Research Data Services. *Journal of EScience Librarianship*. https://doi.org/10.7191/jeslib.2017.1104

Bishop, B., Gunderman, H., Davis, R., Lee, T., Howard, R., Samors, R., Murphy, F., & Ungvari, J. (2020). Data curation profiling to assess data management training needs and practices to inform a toolkit. *Data Science Journal*, *19*(1).

Bishop, B. W., Nobles, R., & Collier, H. (2021). Research Integrity Officers' Responsibilities and Perspectives on Data Management Plan Compliance and Evaluation. *Journal of Research Administration*, *52*(1), 76–101.

Bishop, B. W., Orehek, A. M., Eaker, C., & Smith, P. L. (2022). Data Services Librarians' Responsibilities and Perspectives on Research Data Management. *Journal of EScience Librarianship*, *11*(1), 4. https://doi.org/10.7191/jeslib.2022.1226

Bishop, W., Collier, H., Orehek, A. M., & Ihli, M. (2021). Potential Roles for Science Librarians in Research Data Management: A Gap Analysis. *Issues in Science and Technology Librarianship*, *98*.

Briney, K., Goben, A., & Zilinski, L. (2017). Institutional, Funder, and Journal Data Policies. *Curating Research Data: Practical Strategies for Your Digital Respository*, 61–78.

Camp, R. C. (1989). *Benchmarking: The Search for Industry Best Practices that Lead to Superior Performance*. ASQC Quality Press.

Christensen, G., Dafoe, A., Miguel, E., Moore, D. A., & Rose, A. K. (2019). A study of the impact of data sharing on article citations using journal policies as a natural experiment. *PLoS One*, *14*(12), e0225883.

Christian, T.-M., Gooch, A., Vision, T., & Hull, E. (2020). Journal data policies: Exploring how the understanding of editors and authors corresponds to the policies themselves. *PloS One*, *15*(3), e0230281.

Cox, A. M., Kennan, M. A., Lyon, L., Pinfield, S., & Sbaffi, L. (2019). Maturing research data services and the transformation of academic libraries. *Journal of Documentation*.

De Koning, R., Egiz, A., Kotecha, J., Ciuculete, A. C., Ooi, S. Z. Y., Bankole, N. D. A., Erhabor, J., Higginbotham, G., Khan, M., Dalle, D. U., & others. (2021). Survey fatigue during the COVID-19 pandemic: an analysis of neurosurgery survey response rates. *Frontiers in Surgery*, *8*, 690680.

Dearborn, D., Marks, S., & Trimble, L. (2017). The changing influence of journal data sharing policies on local RDM practices. *International Journal of Digital Curation*.

Diekema, A. R., Wesolek, A., & Walters, C. D. (2014). The NSF/NIH Effect: Surveying the effect of data management requirements on faculty, sponsored programs, and institutional repositories. *The Journal of Academic Librarianship*, *40*(3–4), 322–331.

Dosch, B., & Martindale, T. (2020). Reading the fine print: A review and analysis of business journals' data sharing policies. *Journal of Business & Finance Librarianship*, *25*(3–4), 261–280.

Elsevier. (2022). *Research Data*. https://www.elsevier.com/authors/tools-and-resources/research-data

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y.-L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: an analysis of data availability statements. *PloS One*, *13*(5), e0194768.

Fischer, E. A. (2013). *Public access to data from federally funded research: provisions in OMB Circular A-110*. https://crsreports.congress.gov/search/#/?termsToSearch=R42983&orderBy=Relevance

Garellek, M., Gordon, M., Kirby, J., Lee, W.-S., Michaud, A., Mooshammer, C., Niebuhr, O., Recasens, D., Roettger, T., Simpson, A., & others. (2020). Toward open data policies in phonetics: What we can gain and how we can avoid pitfalls. *Journal of Speech Science, 9*(1), 3–16.

Goldacre, B., Lane, S., Mahtani, K. R., Heneghan, C., Onakpoya, I., Bushfield, I., & Smeeth, L. (2017). Pharmaceutical companies' policies on access to trial data, results, and methods: audit study. *Bmj*, *358*.

Gorman, D. M. (2020). Availability of research data in high-impact addiction journals with data sharing policies. *Science and Engineering Ethics*, *26*(3), 1625–1632.

Graf, C., Flanagan, D., Wylie, L., & Silver, D. (2020). The open data challenge: An analysis of 124,000 data availability statements and an ironic lesson about data management plans. *Data Intelligence*, *2*(4), 554–568.

Grant, R., & Hrynaszkiewicz, I. (2018). The impact on authors and editors of introducing Data Availability Statements at Nature journals. *BioRxiv*, 264929.

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation*.

Holdren, J. P. (2013). *Memorandum for the heads of executive departments and agencies: Increasing access to the results of federally funded scientific research*.

Hollingsworth, P. J. (1999). The research enterprise: accountability and integrity. *American Scientist*, *87*(5), 386.

Hopkins, A. M., Rowland, A., & Sorich, M. J. (2018). Data sharing from pharmaceutical industry sponsored clinical studies: audit of data availability. *BMC Medicine*, *16*(1), 1–6.

Huh, S. (2019). Recent trends in medical journals' data sharing policies and statements of data availability. *Archives of Plastic Surgery*, *46*(06), 493–497.

Inter-university Consortium for Political & Social Research. (2222). *The dataset I wish to analyze is restricted. What do I have to do to get the data?* https://www.icpsr.umich.edu/web/ICPSR/cms/2048

Jeong, G. H. (2020). Status of the data sharing policies of scholarly journals published in Brazil, France, and Korea and listed in both the 2018 Scimago Journal and Country Ranking and the Web of Science. *Sci Ed*, *7*(7), 136–141.

Koltay, T. (2017). Research 2.0 and research data services in academic and research libraries: Priority issues. *Library Management*, *38*(6/7), 345–353.

Krawczyk, M., & Reuben, E. (2012). (Un) available upon request: field experiment on researchers' willingness to share supplementary materials. *Accountability in Research*, *19*(3), 175–186.

Kriesberg, A., Huller, K., Punzalan, R., & Parr, C. (2017). An analysis of federal policy on public access to scientific research data. *Data Science Journal*.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, *47*(4), 2025–2047.

Leckie, G. J., Pettigrew, K. E., & Sylvain, C. (1996). Modeling the information seeking of professionals. *The Library Quarterly*, *66*(2), 161.

Lutz, B. (1999). Will New Federal Provisions Open Laboratory Notebooks to the Public? *Biotechnology Law Report*, *144*(2), 144.

Marshall, E. (2001). Bermuda rules: community spirit, with teeth. *Science*, *291*(5507), 1192. https://www.science.org/doi/full/10.1126/science.291.5507.1192

Melero, R., & Navarro-Molina, C. (2020a). Researchers' attitudes and perceptions towards data sharing and data reuse in the field of food science and technology. *Learned Publishing*, *33*(2), 163–179.

Melero, R., & Navarro-Molina, C. (2020b). Researchers' attitudes and perceptions towards data sharing and data reuse in the field of food science and technology. *Learned Publishing*. https://doi.org/10.1002/leap.1287

Miller, H. G., & Baldwin, W. H. (2001). A terse amendment produces broad change in data access. *American Journal of Public Health*, *91*(5), 824.

Mischo, W. H., Schlembach, M. C., & O'Donnell, M. N. (2014). An analysis of data management plans in University of Illinois National Science Foundation grant proposals. *Journal of EScience Librarianship*, *3*(1), 31–43.

Naudet, F., Sakarovitch, C., Janiaud, P., Cristea, I., Fanelli, D., Moher, D., & Ioannidis, J. P. A. (2018). Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in The BMJ and PLOS Medicine. *Bmj*, *360*.

Office of Management & Budget. (1999). *OMB Circular A-110*.

Open, Public, Electronic and Necessary (OPEN) Government Data Act, Pub. L. No. 5529 (2019).

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

P.L. 105-277, 112th Stat. 2681-495. (1998).

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.

Park, J., Howe, J. D., & Sholl, D. S. (2017). How Reproducible Are Isotherm Measurements in Metal-Organic Frameworks? *Chemistry of Materials*. https://doi.org/10.1021/acs.chemmater.7b04287

Patel, D. (2016). Research data management: a conceptual framework. *Library Review*.

Phillips, D. L., & Clancy, K. J. (1972). Some effects of" social desirability" in survey studies. *American Journal of Sociology*, *77*(5), 921–940.

Pinfield, S., Cox, A. M., & Smith, J. (2014). Research data management and libraries: Relationships, activities, drivers and influences. *PLoS One*, *9*(12), e114734.

PLOS. (2022). *Data Availability*. https://journals.plos.org/plosone/s/data-availability

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. In *Science*. https://doi.org/10.1126/science.1197962

Resnik, D. B., Morales, M., Landrum, R., Shi, M., Minnier, J., Vasilevsky, N. A., & Champieux, R. E. (2019). Effect of impact factor and discipline on journal data sharing policies. *Accountability in Research*, *26*(3), 139–156.

Rousi, A. M., & Laakso, M. (2020). Journal research data sharing policies: a study of highly-cited journals in neuroscience, physics, and operations research. *Scientometrics*, *124*(1), 131–152.

Rudder, C. E. (1999). APSA responds to OMB’s draft regulations. *PS: Political Science & Politics*, *32*(2), 188–190.

Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PloS One*, *4*(9), e7078.

Scaramozzino, J. M., Ram\’\irez, 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.

Semeler, A. R., Pinto, A. L., & Rozados, H. B. F. (2019). Data science in data librarianship: core competencies of a data librarian. *Journal of Librarianship and Information Science*, *51*(3), 771–780.

Shelby, R. (2000). Accountability and transparency: Public access to federally funded research data. *Harv. J. on Legis.*, *37*, 369.

Sholler, D., Ram, K., Boettiger, C., & Katz, D. S. (2019). Enforcing public data archiving policies in academic publishing: A study of ecology journals. *Big Data & Society*, *6*(1), 2053951719836258.

Spendolini, M. J. (1992). *The Benchmarking Book*. Amacom Books.

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), 63–78.

Subramanyam, K. (1981). *Scientific and Technical Information Resources*. Decker, Inc.

Swauger, S. (2015). DMPTool. *The Charleston Advisor*, *16*(3), 12–15.

Tal-Socher, M., & Ziderman, A. (2020). Data sharing policies in scholarly publications: interdisciplinary comparisons. *Prometheus*, *36*(2), 116–134.

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., & others. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, *8*(1), 1–11.

Tenopir, C., Allard, S., Baird, L., Sandusky, R. J., Lundeen, A., Hughes, D., & Pollock, D. (2019). Academic librarians and research data services: Attitudes and practices. *IT Lib: Information Technology and Libraries Journal. Issue 1*.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, *6*(6). https://doi.org/10.1371/journal.pone.0021101

Tenopir, C., Christian, L., Allard, S., & Borycz, J. (2018). Research data sharing: Practices and attitudes of geophysicists. *Earth and Space Science, 5*(12), 891–902.

Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., Grant, B., Olendorf, R., & Sandusky, R. J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLoS ONE*, *15*(3). https://doi.org/10.1371/journal.pone.0229003

Tenopir, C., Sandusky, R. J., Allard, S., & Birch, B. (2014). Research data management services in academic research libraries and perceptions of librarians. *Library and Information Science Research*, *36*(2), 84–90. https://doi.org/10.1016/j.lisr.2013.11.003

U.S. National Institutes of Health. (2003). *NOTICE: NOT-OD-03-032, FINAL NIH STATEMENT ON SHARING RESEARCH DATA*. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

U.S. National Institutes of Health. (2020). *Notice NOT-OD-21-013: Final NIH Policy for Data Management and Sharing*. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

U.S. National Science Foundation. (2011). *NSF 11-1: Grant Proposal Guide*.

U.S. National Science Foundation. (2022). *Preliminary federal obligations for research and experimental development, by agency and type of R&D: FY 2021*. https://ncses.nsf.gov/pubs/nsf22323/assets/data-tables/tables/nsf22323-tab006.xlsx

Walter, C., & Richards, E. P. (2000). When does the Freedom of Information Act apply to privately held data produced under a Federal grant? I. *IEEE Engineering in Medicine and Biology Magazine*, *19*(4), 121–125.

Weber, M. (2018). The effects of listing authors in alphabetical order: a review of the empirical evidence. *Research Evaluation*, *27*(3), 238–245.

Whitmire, A. L., Boock, M., & Sutton, S. C. (2015). Variability in academic research data management practices: implications for data services development from a faculty survey. *Program*, *49*(4), 382–407. https://doi.org/10.1108/PROG-02-2015-0017

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., & others. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9.

Williams, M., Bagwell, J., & Zozus, M. N. (2017). Data management plans: the missing perspective. *Journal of Biomedical Informatics*, *71*, 130–142.

Yasin, M. M. (2002). The theory and practice of benchmarking: then and now. *Benchmarking: An International Journal*.

Appendix

STUDY 1: RESEARCH SURVEY

STATEMENT OF CONSENT
By clicking Next, I indicate that I have read this form and the research study has been explained to me. I have been given the chance to ask questions and my questions have been answered. If I have more questions, I have been told who to contact. Please click Next below to proceed with participating in this study.

On average, how much digital storage space do you think it would take for you to archive the data, code, and supporting files for a single research project? (Select one)

1 Less than 10 GB
2 More than 10 GB but less than 100 GB
3 More than 100 GB but less than 500 GB
4 More than 500 GB but less than a terabyte
5 On the order of one or more terabytes.
6 Other

How often do you work with the following categories of data? (Select for each row)

	Never	Sometimes	Often	Always
Sensitive (medical or participant privacy, trade secrets, export control, etc.)	□	□	□	□
Classified (by government designation)	□	□	□	□
Open (no external constraints on sharing and reuse)	□	□	□	□

How often do you rely on each of the following as a solution for long-term data archival, after the project is complete? (If you have not had official responsibility for data management on a project, choose the answers based on the decisions you think you would be most likely to make.)

	Never	Rarely	About Half the Time	Most of the time	Always
My personal computer or laptop.	□	□	□	□	□
External drive, USB, or other external media.	□	□	□	□	□
On a departmental server.	□	□	□	□	□
On a university-hosted server through our Office of Technology.	□	□	□	□	□
With a publisher / a publisher-related repository / as supplementary paper materials.	□	□	□	□	□
With a data repository outside the university.	□	□	□	□	□
I store my data in my organization’s institutional repository.	□	□	□	□	□
Cloud storage.	□	□	□	□	□
On paper / as physical materials (printed data records, lab notebooks, etc.)	□	□	□	□	□

Is there anything else you'd like to say to describe your data?

The items below describe activities that might occur in the context of Research Data Management. These items use the type of language that might be found in journal, funder, or institutional policies. How well do you feel each of these has been explained to you?

	Poorly Explained	Modestly Well Explained	Well Explained	Very Well Explained
Creating metadata and other documentation to describe my research data.	□	□	□	□
Archiving data in a repository for long-term storage.	□	□	□	□
Including a Data Availability Statement in a manuscript.	□	□	□	□
Creating a license to describe how data may be used or reused.	□	□	□	□
Assigning a permanent identifier (such as DOI or accession number) to my datasets.	□	□	□	□
Creating a data management plan to be submitted to a funding agency, as part of a proposal.	□	□	□	□
Citing the use of data that you did not personally create.	□	□	□	□

If you need help with each of the following factors, which campus-wide support unit would you be most likely to turn to?

	Libraries	Office of Information Technology	Office of Research	Office of the Provost
Creating metadata and other documentation to describe my research data.	□	□	□	□
Archiving data in a repository for long-term storage.	□	□	□	□
Including a Data Availability Statement in a manuscript.	□	□	□	□
Creating a license to describe how data may be used or reused.	□	□	□	□
Assigning a permanent identifier (such as DOI or accession number) to my datasets.	□	□	□	□
Creating a data management plan to be submitted to a funding agency, as part of a proposal.	□	□	□	□
Citing the use of data that you did not personally create.	□	□	□	□

To what extent would you consider each of the following factors to be a barrier to archiving your data and supporting code?

	Not a barrier (Includes not relevant, or if you have already solved this problem)	Somewhat of a barrier	Moderate barrier	Extreme barrier
Finding a long-term archival solution that works for the type, subject, or size of my data.	□	□	□	□
The time it takes to organize and document the data, code, and other supporting files so that it can be understood by others.	□	□	□	□
The fees or costs to submit data for preservation in a repository or other solution.	□	□	□	□
Finding a repository which enforces adequate access restrictions or security protections for sensitive, private, or classified data.	□	□	□	□

Does your work ever include research with human subjects?

1 Yes
2 No

Regarding data for projects involving Human Subjects: How well do you feel each of these has been explained to you?

	Poorly Explained	Modestly Well Explained	Well Explained	Very Well Explained
How to write an informed consent statement using language that that does not conflict with data sharing.	□	□	□	□
Understanding of how secondary use of existing human subjects data affects the IRB review process.	□	□	□	□
Understanding how privacy concerns for human subjects data affects where data can be stored and how it should be protected.	□	□	□	□
Knowledge of how to sanitize and otherwise prepare data from human subjects research for post-project archival.	□	□	□	□

Is there anything else that you would like to say about research data management?

Roughly how much of your research is funded?

1 None
2 Some
3 Most
4 All

Have you served as a Principal or Co-Investigator for funded research within the last 3 years?

1 Yes
2 No

How well do you feel you understand the research data policies and data management expectations of the funding sources you work with (or that you might want to work with in the future)?

1 Poor Understanding
2 Modest Understanding
3 Good Understanding
4 Excellent Understanding

What is the name of the institution which you consider to be your primary affiliation?

NOTE: Your response to this question will be mapped to an arbitrary identifier such as "Institution 12" so that aggregate data can be compared across organizations without identifying the institution. Your original response will then be destroyed. No copy of the mapping will be retained or published.

How long have you been at this institution?

Which best describes your position at the university?

1 Faculty
2 Staff
3 Graduate Student
4 Undergraduate Student
5 Other

Which of the following best describes your primary area of research?

1 Business (Examples: Accounting, economics, finance, management, marketing, hospitality, tourism)

2 Humanities (Examples: Art, history, languages, literature, music, philosophy, religion, theater)

3 Natural Sciences (Examples: Biology, chemistry, computer science, engineering, geology, mathematics, physics)

4 Social & Behavioral Sciences (Examples: Anthropology, education, geography, law, political science, psychology, sociology.)

5 Health (Examples: Medicine, nursing, kinesiology)

6 Veterinary Medicine & Agriculture (Examples: Veterinary science, agronomy, agriculture, forestry)

7 Other __________

Vita

Monica Ihli was born in Sulphur, Louisiana. She graduated with an Associate of Science in Computer Science from Pellissippi State, as well as a Bachelor of Arts in Communication Studies from University of Tennessee, both in Knoxville, Tennessee. Her Master of Science in Information Science is also from the University of Tennessee. Her research background includes research informatics, and scientific data systems as well as research data management policy and adoption.