Policies, Procedures, and Decision-Making: Data Managers and the Research Lifecycle

A.J. Million
Inter-university Consortium for Political and Social Research, University of Michigan, USA
millioaj@umich.edu

Jenny S. Bossaller
School of Information Science & Learning Technologies, University of Missouri, USA
bossallerj@umsystem.edu

ABSTRACT
Research data is an asset. Researchers may be required to provide access to their data by scientific funders or academic journals and deposit their data in archives. Managers of archives are guided by principles, policies, and the law when curating and providing access to data. Practices around data storage and access, however, are not always cut and dry; research data managers sometimes need to interpret policies. This paper presents findings from qualitative interviews with 15 data managers from 8 repositories in the U.S. These repositories were all affiliated with universities but served varied constituents and provided a range of services. Differences revealed opportunities and challenges in managing data repositories regarding, for instance, who can access data and the level of protection data requires. We also found that data-related policy challenges may stem from any stage of the research lifecycle.

KEYWORDS
Research data management; archives; policy; decision-making; research support.

INTRODUCTION
Empirical research is performed through surveys, measurements, observation, and other methods that produce data. Data can take significant time to collect, and thus, it has value. Primary data is often used to answer original research questions, but if it is made accessible after its original use, other researchers can reuse it to answer questions through secondary data analysis (Devine, 2003; Heaton, 2008; Moulaison Sandy et al., 2017). Researchers might share their primary data for altruistic reasons, to support readers' interpretations of their work, or because they are required to do so by a funding agency or academic journal (Stebbins, 2013; Tenopir et al., 2018). Datasets can be stored in a university repository, a data repository associated with a publisher, or a standalone repository. Repositories are often managed by a librarian or data manager who helps researchers deposit their data, curate the data, and provide access to data.

Increasingly, research data is available through archives because of data sharing requirements (e.g., the White House Office of Science and Technology Policy (OSTP) 2013 memorandum entitled *Expanding Public Access to the Results of Federally Funded Research* "directed Federal agencies with more than $100M in R&D expenditures to develop plans to make the results of federally-funded research freely available to the public" (para. 1)). Digital tools make it easy for researchers to collect and share data, and the open science movement encourages the sharing of data to allow for study replication.

Policies, created by repositories, funding agencies, etc., guide decisions about what kinds of data archives can accept, how data will be protected, who will be granted access to data, and how users will access it (e.g., in person, via VPN). Interpreting or implementing policy, though, can be difficult, especially with sensitive or problematic data. Research managers also must comply with legal and policy requirements (e.g., *The Family Rights and Privacy Act*, or FERPA; Institutional Review Board (IRB) requirements; *The Common Rule*) created by other entities governing the research lifecycle.

This study focuses on problems that research data managers encounter and how they interpret and implement policy to meet organizational goals and advance science.

RESEARCH QUESTIONS
This paper presents findings from interviews with researchers in their work as data archivists. What is the nature of problems that arise in research data archives? How did our study participants navigate decisions in their work that challenged or complicated policy adherence and implementation?

METHOD
We completed our study in January 2021 by conducting semi-structured interviews, using Zoom, with 15 participants (7 administrators and 8 staff) in 8 U.S. data archives, including large, internationally known archives, smaller institutional archives affiliated with university libraries, and archives affiliated with research institutes. Our

84th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – Nov. 3, 2021 | Salt Lake City, UT. *Author(s) retain copyright, but ASIS&T receives an exclusive publication license.*

interview protocol and IRB form were distributed to participants prior to our interviews. Each participant was invited to describe their job and then describe a problem they had encountered in the course of their regular work. Our interviews were recorded and transcribed using the online service Rev. The researchers read the transcriptions, identified main policy and decision-making problems that the participants discussed, then created overarching themes that framed the problems discussed by study participants. The researchers identified (1) the "types" of dilemmas during the research lifecycle in the context of data management, and (2) situations where policy was insufficient to avoid dilemmas or guide decision-making.

FINDINGS

The main ideas we identified in our interviews related to *access and sharing data*, *data collection*, and *infrastructure and resources*.

Participant and Role	Case	Main Idea
P-1: Project Manager	Challenges in responding to restricted data access violations and providing computing resources	Access and sharing data; Infrastructure and resources
P-2: Research Professor, Past Director	Obtaining permission from school districts for a large-scale research project that was consistent with FERPA to create a data access plan that could be approved by IRB	Access and sharing data; Data collection
P-3: Project Manager	Conflict between the U.S. executive and judicial branches regarding who provides access to a statistical data series	Access and sharing data
P-4: Research Investigator	Dilemma in determining how to allow sharing of crosswalk file for a legacy longitudinal study	Access and sharing data
P-5: Project Manager	Worries by curator about releasing a high-profile, well-funded research dataset because of questions about accuracy and falsification	Data collection
P-6: Privacy Officer	Questions about how to make uniform decisions about the repercussions of restricted data access violations	Access and sharing data
P-7: Assistant Director	When working with a journal, there is a requirement for computational reproducibility using data; however, journal editors can overrule this, so the archive creates notes if this happens	Access and sharing data; Data collection
P-8: Associate Director	Cases of questions whether qualitative data can be shared	Access and sharing data
P-9: Project Manager	Lack of detail in grant document and policy to determine who to contact when there was a restricted data access violation	Access and sharing data
P-10: Research Data Librarian	Issues with determining ownership of code	Access and sharing data
P-11: Director of Curation	Difficulty in curating data from multiple funding agencies to meet federal data archiving requirements	Infrastructure and resources
P-12: Librarian	Challenges associated with creating a university-wide research data policy	Infrastructure and resources
P-13: Senior Curation Specialist	Difficulty in accommodating data depositors request for their data to be provided as open access	Access and sharing data
P-14: Director of Research Data Services	Balancing the desire to provide open, free access to data by the library with IRB desire to protect study participants, and doing so in a comprehensive manner under a larger policy framework	Access and sharing data; Data collection; Infrastructure and resources
P-15: Director of Research Data Services	Finding how much storage space to provide for users who self-deposit for them to meet data archiving requirements	Infrastructure and resources

Table 1. Participants, dilemmas, and main ideas found in dilemmas

Access and sharing data

The most common problem presented by our study participants, as expected, related to providing access to data after it had been deposited in an archive. This step falls at the end of the traditional research cycle but at the beginning of the secondary research cycle (that is, re-use). Participants discussed reasons that they encountered for depositing data in a repository (e.g., a requirement of a federal grant or journal) as well as secondary users' research intentions (e.g., combining datasets to produce new knowledge or for investigatory reporting).

P-1 reported that although their archive has policies regarding terms-of-use, sometimes researchers do not adhere with these terms. Their dilemma centered on a researcher's publication that included sample sizes that were small enough to identify individuals; they explained that their repository has a responsibility and contractual obligations to

prevent such data from being published. They also had policies to balance providing data access with the potential for harm, which may come from reuse: "there's this grid where it's easy to identify little harm, or hard to identify, but a great deal of harm could happen." P-2 reiterated the point and demonstrated a challenge for the open science movement: "In this space we're talking about confidential data, the dilemma is, how much protection do you put, do you use, versus how easy you make it for researchers? The researchers want everything to be costless from their point-of-view, and they want to get as much data as possible."

Meeting policy or legal requirements

All data repositories have policies to guide decisions, but they are not iron-clad; there are exceptions which arise based on changes in situations, personnel, technology, etc., and differing interpretations of policy at all stages of the research process. For instance, data managers ensured that researchers followed research protocols approved by IRBs prior to data collection, and that archives followed their own policies for data storage and access after deposit. P-7 described a policy conflict about publishing data with a journal editor. The archive is contracted to verify all statistics following peer-review:

> the authors are required to submit their data sets and their code and certain documentation…We curate [data] to make sure all the documentation is complete, and then that dataset goes to our statisticians. These are Ph.D. students who will run the code on the data to make sure the outputs match what's in the manuscripts before the manuscript is published. And then we publish the data.

Disagreement arose when statisticians were unable to verify study results, but the editor wanted to publish the paper anyway. Ultimately, the editor's wishes were granted, but P-7 felt mollifying the author and editor ultimately undermined the journal and archive's reputations.

P-13 discussed a need for consent documents that clearly allowed data-sharing, marking a legal issue at the beginning of the research cycle. P-2 discussed storage and access to student test scores and teacher observation data, which fell at the end of the research cycle. The school system's lawyers focused on FERPA, but the archive staff were concerned about teacher data because it was identifiable. P-6 discussed external regulations regarding the security of information systems. P-10's dilemma concerned the intellectual property of code; policy or legal requirements were not sufficient and muddied by university regulations about intellectual property, as it was the property of a university. P-11 discussed policies surrounding levels of curation and resources that data managers are allowed to devote. Less experienced staff members, complicated data, and unfamiliar technologies can cause processing delays, and archivists must have some flexibility in meeting federal archiving policy requirements when delays arise.

Infrastructure and resources

The third theme in our findings related to policies in place within data archives regarding infrastructure and resources, such as how much data an archive can accept, and resources (human, time, finances) that the archive is able to devote to a project. Data security is an important systems-related consideration; smaller university repositories might not be equipped to provide secure data storage and access, so their policy may be to direct researchers to a more appropriate archive. Time was frequently mentioned as a restraint, especially regarding data curation. P-15 discussed constraints on storage space with researchers allowed two terabytes of storage per year, but they explained, "it's not like there's a hard stop in the system. [At] 10 [GB], we start to get a little bit more anxious. 80 [GB], we can't do. It's not the right system for that, honestly." If researchers required more resources than the archive could provide the archivist directed them to another, more appropriate institution. Resource availability and infrastructure were especially problematic for smaller, general university data archives as well, which tend to have fewer resources than large or topical repositories. Larger, specialized repositories appeared to have more resources to perform tasks associated with data curation and providing secure data storage, as well as other services.

DISCUSSION

One of the purposes of this study was to find if rules and policies effectively guide data managers' decision-making, and whether they have sufficient capacity to navigate conflicts or anomalies that fall outside of policy. The people we interviewed were all professionals and had substantial expertise in their domain. Most worked on a team that can help clarify procedures or directions when no existing policy applied. P-6 explained that they "create policies that we can generally all agree on or at least policies that are 80-20, right? Like, it covers 80% of the use cases, and then 20% we deal with as one-offs." This means policy was generally sufficient to address problems, which we found most often related to our three themes.

Research data managers often work within a multi-layered system, though, and the values inherent in such layers are not always in alignment. This was noted by P-14:

> I think our organizations all have particular ways that we think about this. Like the library is coming at [data policy] from one angle of wanting to make information as open and accessible as possible to satisfy people's

> […] needs. The IRB has the mantra or the rule of protecting human subjects. And sometimes that leads them to think more overly cautious than perhaps might be warranted in my generalist opinion. The university itself, I think they recognize that there is a compliance aspect to this, and they do not want to jeopardize research funding or being able to publish in prestigious journals. But they're also very worried about getting sued or having legal action taking against them.

P-12, however, found that they could rely on campus partners (e.g., IRB or the legal team) to help address dilemmas in research data management.

During the course of our interviews, most participants stated that their repositories exist to support scholarship, and when possible, to make data open and reusable. However, "open" does not always mean freely available to everyone. Determining who is eligible to access data (e.g., credentials needed), and how (in person, via secure VPN, etc.) created intractable problems for study participants. Several discussed occasions when they refused to accept or provide access to data. P-6 said that it was easy to comply with associated federal regulations, but that "individual users who want to access data and there's something weird about it, or a dataset has come to us where we're not sure how best to protect it because it's something we don't often deal with" are problematic. Some datasets are very complex. These data might have been created with specialized software or domain-based knowledge. In such situations, policy was insufficient, and archives turned to partners for help.

Publishers, too, might present an emerging concern in the research data space. P-12 explained that data services located in university libraries focus on open and reproducible data, but that now publishers:

> are really expanding the scope of their services to grab more than just research articles. They're trying to get data, they're trying to move into this open area and enclose them a little bit, and there's a concern. We're very protective of our intellectual property on campus, but we don't really have anything that we say around data.

This quote emphasizes both the value of data and the necessity investment in the infrastructure of data curation and preservation to support scholarship.

CONCLUSION

The research data management space is not new, but it is still evolving. Different models of data archives continue to emerge in response to researchers' needs, technological evolution, and grantors' requirements. Laws and policies exist which pertain to each phase of the research lifecycle, but they cannot anticipate all problems. New technologies, legal or funders' requirements, and competition tax data archives' resources. Data managers still have to make decisions on a case-by-case basis in these instances, shaping policy as they muddle through decisions.

REFERENCES

Devine, P. (2003). Secondary Data Analysis. In Miller, R. & J. D. Brewer, The A-Z of Social Research. London: Sage.

Durant, R. F., & Edwards III, G. C. (Eds.). (2010). *The Oxford handbook of American bureaucracy*. Oxford University Press.

Heaton, J. (2008). Secondary analysis of qualitative data: An overview. *Historical Social Research/Historische Sozialforschung*, 33-45.

Moulaison Sandy, H. M., Mitchell, E., Corrado, E. M., Budd, J., West, J. D., Bossaller, J., & VanScoy, A. (2017). Making a case for open research: Implications for reproducibility and transparency. *Proceedings of the Association for Information Science and Technology*, *54*(1), 583-586.

Stebbins, M. (2013). Expanding public access to the results of federally funded research. *Office of Science and Technology Policy*.

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.