Million Anthony (Orcid ID: 0000-0002-8909-153X)

The Research Data Life Cycle, Legacy Data, and Dilemmas in Research Data Management

Jenny Bossaller

University of Missouri

School of Information Science & Learning Technologies

bossallerj@missouri.edu

A.J. Million

University of Michigan

Inter-university Consortium for Political and Social Research

millioaj@umich.edu

Author Manuscript

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/asi.24645

This article is protected by copyright. All rights reserved.

Abstract

This paper presents findings from an interview study of research data managers in academic data archives. Our study examined policies and professional autonomy with a focus on dilemmas encountered in everyday work by data managers. We found that dilemmas arose at every stage of the research data lifecycle, and legacy data presents particularly vexing challenges. The iFields' emphasis on knowledge organization and representation provides insight into how data, used by scientists, are used to create knowledge. The iFields' disciplinary emphasis also encompasses the sociotechnical complexity of dilemmas that we found arise in research data management. Therefore, we posit that iSchools are positioned to contribute to data science education by teaching about ethics and infrastructure used to collect, organize, and disseminate data through problem-based learning.

Author Manuscript

Keywords

Archives; Data science; Decision-making; Education; Policy; Research data management

Author Manuscript

Introduction

In 2020, we interviewed 15 research data managers from eight data archives in the United States (U.S.). Data archives included large, internationally known archives, research institutes, and small institutional repositories affiliated with large research universities. Our interview questions focused on decision-making, policies and professional autonomy, and asked research data managers to identify cases or situations where they had encountered dilemmas that challenged their professional judgment. Speaking with frontline managers revealed perspectives that are largely absent in prior research, especially regarding data reuse and ethics. When speaking with data managers, we found they act as problem solvers responsible for implementing and interpreting policy while managing data.

Our study found that data managers can generally rely on policies and rules to guide their actions, and that as professionals, they have sufficient experience to navigate ethical and technical challenges faced on-the-job (Author & Author, 2021). We also found that dilemmas arise at every stage of the research data lifecycle, sometimes spilling over into the actual research lifecycle, and legacy data and associated deposit agreements may create difficult challenges. These two findings are the framework for this communique.

To conclude, we tie our findings to the contributions the iFields make to data science, arguing they provide a broad, sociotechnical view of how data are collected, used, and stored to generate knowledge. We suggest that iSchools are positioned to advance data science by integrating applied ethics and problem-based learning into curricula. We make this suggestion, because new technologies and analytical methods often create tension between broader efforts to promote data reuse, and the ethical conduct of research, which elude bureaucratic and policy-

bound institutional ethics boards (among other governing bodies). We present an evolving cadre of possible cases for inclusion in an information science curriculum.

Dilemmas at all Stages of the Research-data Lifecycle

Unlike domain-specific fields that leverage data to answer research questions, the iFields focus on knowledge organization and representation (Bates, 1999). As such, iSchools approach research data management (RDM) as phenomena relating to how individuals collect, curate, disseminate, and preserve information, and train students how to be managers in research data environments. Because the iFields, especially library and information science (LIS), conceptualize data as an asset to manage from creation to destruction, they treat it as information which lends itself to the creation of knowledge. Data science is frequently understood as a "concept to unify statistics, data analysis, informatics, and their related methods" to "understand and analyze actual phenomena" by using data (Hayashi, 1998), but this view neglects other phases of the research data lifecycle.

Our study examining research data managers suggests the need for a broader curriculum for data science education that moves beyond technical skills (which must be included) encompassing the complexity of activities that might occur throughout the complete research data lifecycle, which is the domain of RDM librarians and archivists. Our interviews with professionals who managed data to facilitate academic inquiry revealed dilemmas throughout all phases of the research data lifecycle. We focused on the infrastructure that supports researchers (Star & Ruhleder, 1996) and houses data for primary use by original researchers, as well as secondary use by subsequent researchers. All data must come from somewhere; the systems used by academic researchers in our findings demonstrate the interdependence of primary and secondary data users.

Author Manuscript

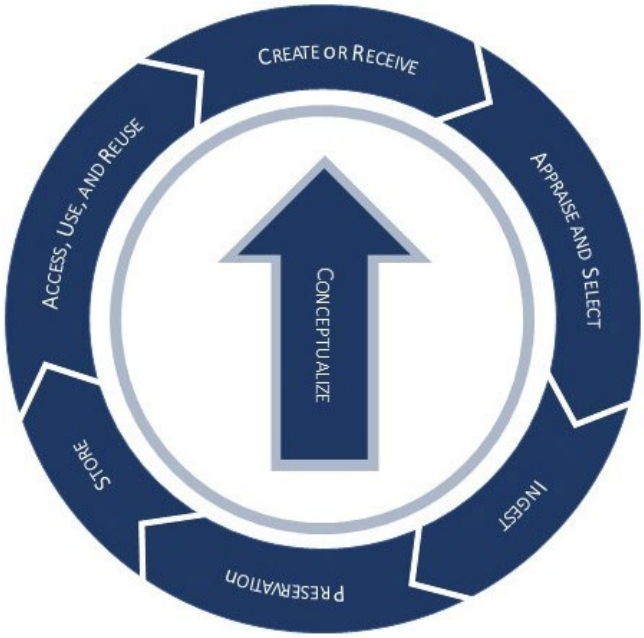
Author Manuscript

Findings

Research Data Management

Best practice dictates that RDM should begin during the initial stages of research, before data collection takes place. Funding agencies increasingly require researchers to create data management plans (National Science Foundation, 2021), and researchers often work with RDM archivists or librarians to write these plans.[1] Likewise, data archives create policies they use to make decisions about ingesting, organizing, and disseminating data. Despite careful planning and precautions taken by researchers and archivists, however, our study participants reported they encountered dilemmas throughout every stage of the research data lifecycle (see *Figure 1*), which we define as similar to Higgins' (2008) model. Several dilemmas spanned across multiple lifecycle stages, even spilling over into the research process itself.

Figure 1. The Research Data Lifecycle



[1] The Research Data Management Librarian Academy (RDMLA) encourages librarians to work with researchers at this stage of the process.

Author Manuscript

Conceptualization.

Conceptualization, or identifying key concepts and/or variables in an investigation, is the first step in a research project. In our study, we found most dilemmas relating to conceptualization centered on what data can be legally and ethically collected and stored. P2 described a conflict at this early stage where researchers wanted to archive data they collected in public school classrooms. Data reuse was not initially a study goal, though, and teachers were not asked to give their consent because they were not the study's focus. Once the study was complete, archivists realized that teachers were especially at-risk for reidentification. Examining a deposit agreement in depth, legal counsel and an institutional research board (IRB) said archiving the data might also violate *Family Education Rights and Privacy Act* regulations.[2] The dataset was ultimately archived, but only after extensive negotiations with multiple stakeholders to ensure data could properly be reused and teachers' identities protected. Moreover, this dilemma related to the *creation* of data and its *receipt* by archives.

Appraise, Select, and Ingest.

The Digital Curation Centre (Whye & Wilson, 2010) explains that, as with other types of archives, data archives must be selective in what they ingest. It is easy to select a particular dataset, but it can be difficult for archivists to appraise their value. Selecting data for inclusion in an archive may be guided by a combination of legal requirements, organizational policies, and professional practices. Two cases from our study exemplify dilemmas found in the *data appraisal and selection* and *ingestion* lifecycle stages. P5 spoke about deciding whether the archive could accept potentially falsified study data. Their challenge was an ethical one, because organizational policy did not provide clear guidance on the matter. P10 discussed software code

[2] A *deposit agreement* is a legal document giving an archive "permission to deposit datasets and carry out activities [to ...] facilitate the long-term preservation and sharing of datasets" (Imperial College London, n.d.).

ownership. Determining who owns code can be challenging, and in this case, library employees were unable to determine who had authority to deposit the code and if a research team was required to archive it.

Preservation and Storage.

The next phases of the research data lifecycle are associated with curation, which includes both *preservation* and *storage*. One interviewee (P11) working for a standalone research institute described a case of where her staff were paid to curate data for grant awardees. Insufficient staff slowed curation, however, and as a consequence the funding agency could not adequately meet statutory requirements. P15, who worked in a major university library, explained that their institution's archive was constrained by file sizes. Their self-archiving system helped sidestep a staff shortage, but even minimally curated data requires space. These cases suggest a basic truism—that research data archives typically require significant investments to support the needs of the primary and secondary users, as described in Borgman et al.'s (2018) study of the Data Archiving and Networked Services of The Netherlands (DANS). Like P15's institutional repository, DANS enables researchers to self-archive their data, but even so, it still requires extensive amounts of manual labor and "human, technical, and policy infrastructure" (p. 898) to operate.

Access, Use, and Reuse.

Providing *access* to data and promoting its *reuse* are also steps in the research data lifecycle. Promoting data access and reuse presented, by far, the trickiest challenges participants reported in our interviews. Archiving data enables researchers to 1) reproduce or verify past work, 2) make the results of funded research available, 3) enable others to answer new questions by using extant data, and 4) advance science (Borgman, 2012). Data producers (researchers) also

Author Manuscript

Author Manuscript

often work with archivists to determine what access restrictions they place on datasets. Qualitative interview transcripts, for example, are hard to anonymize so their reuse increases the risk of participant reidentification. Eight of our 15 study participants discussed problems concerning data access, use, and reuse:

- P7 described their archive's contract to validate findings for publication in a journal. Conflict arose when the journal's editor wanted to publish a paper that archive staff could not replicate, because the researchers used specialized, proprietary software. Staff included a note in the paper saying they could not replicate the study, but they still felt it damaged the journal's reputation.
- P6 discussed difficulties enforcing data access rules. Archivists typically work with researchers to set access rules that balance study participant privacy with data accessibility, but P6 said it was hard to find and sanction researchers who violated these policies. Similarly, several participants (P2, P3, P4) mentioned that some federally funded data archives lacked sufficient information about who they should contact in the event security breaches and access violations happened.
- One participant (P8) described being unable to make a researcher's data freely accessible, despite their wishes, because it contained direct identifiers: "We're kind of banging our head against the wall. The researchers are saying, 'well, IRB says, it's HIPAA deidentified' [...]. I don't know how much you know about HIPAA deidentification, but for qualitative data, that's a necessary, but definitely not a sufficient condition." The IRB also approved sharing the data, but the archivists asked, "What did you really tell participants? And what does that mean for their autonomy, which is this value that we are upholding." Here, archivists and researchers disagreed about the risks associated with

making data "open" and the ethics of reuse. Study participants had not been informed their interviews could be used for future projects.

- P14 talked about balancing their desire to provide unrestricted access to data with study protocols cleared by researchers' IRBs. They said: "In our heart... we really want information to be as open and accessible and discoverable as possible." However, researchers sometimes inadvertently created impediments to data sharing when planning and carrying out research: "On the consent form, they're just thinking like, 'Oh okay the IRB is going to go after me. And I need to be sure that I make them as happy as possible. So, no sharing right off the bat! That'll get me through the IRB process,' without realizing that they're going to hurt themselves later if they're required to, or if they want to share that data."

In cases like those we mention above, frontline data managers could not resolve problems by deferring to policies or rules. Resolving problems required that managers use their professional discretion. Furthermore, it was hard to pin down where in the research data lifecycle some dilemmas fell, because decisions at one point could cause issues later. One example of this is when P7 said a dataset was produced with proprietary software that prevented study replication.

Legacy Data

Our second study finding was that legacy data is hard to manage, especially when deposit agreements are involved. Before the "data deluge" the purpose of data archives was to store data (Hey & Trefethen, 2003). Increasingly, however, the iFields pay attention to repositories' potential to expedite the creation of new scientific knowledge with scientists creating knowledge by reusing that archived data. Our interview with P4 shows how old datasets can create new

Author Manuscript

problems. In this case, P4 described a longitudinal dataset archived by a multi-institution research team. Years later, archivists found a crosswalk file to which the research team reserved the right to unilaterally approve or reject access. A deposit agreement based on outdated data sharing norms permitted this, and subsequently, there were systemic inequities in data access and research outputs within the researcher's discipline.

P3 described providing non-disclosive data in a tightly controlled physical enclave which dramatically limited reuse. The problems found in this case were because of technological innovation and disciplinary norms. Data sharing was not common when the data was archived, and affordable technology did not exist to provide secure, remote access. Today, though, technology does exist to provide secure, remote access, and researchers increasingly feel entitled to reuse data, especially from government-sponsored research. The archive where P3 worked could not provide remote access, because the data deposit agreement was too complex to renegotiate, and the depositor was resistant to change. Borgman et al (2018) observe the value of data archives is in taking a long view. Although P3's archive could not resolve their dilemma, sometimes archives might claim control over data when two parties "do not satisfy their contractual obligations" (p. 901) or when there is another need to do so.

Data Science Education

What are our findings' implications for data science education? We found that diverse and complex dilemmas arise at every stage of the research data lifecycle, and that legacy data in particular create challenges. Informed by this knowledge, we argue data managers and professionals need a broad, sociotechnical understanding of the research data lifecycle to build, use, and sustain infrastructures that make research data accessible and reusable. Effective RDM is also central to realizing the potential of open science by training researchers to find and

Author Manuscript

resolve issues relating to scientific data collection, access, and dissemination, as well as scholarly communication.

Based on the complexity and variety of dilemmas our study participants encountered, we argue iSchools can contribute to the advancement of data science by teaching RDM with problem-based learning strategies. Research examining current iSchool and data science curricula suggest that analytical skills, research methods, and data management are taught, but upper-level skills like problem-solving are not usually emphasized (Si, 2013; Tang & Saeb-Lim, 2016). Teaching skills that relate to the dilemmas faced by research data managers and scientists using problem-based learning would improve data science education by emphasizing both research processes (primary research) and the practices of data archives that ultimately facilitate secondary data reuse.

One of our study participants made a key point that justifies iSchools taking this approach. P6 noted how the 80/20 rule applies: "So [our work is ...] keeping track of [all aspects of RDM and ...] advising on it so that nothing goes too far outside of the boundaries of what [we are ...] comfortable with. And it's, to the extent possible, meeting with the internal stakeholders, so leadership [...] and the managers to 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." Some basic principles apply to all data archives, but communities, expectations, needs, and processes differ. For instance, data from astrophysics, political science, and health science (Caso & Ducato, 2014) each bring with them different ethical issues stemming from the environments where data are collected. Archives create localized policies based on professional best practices and legal requirements that apply to most (e.g., 80%) of their

Author Manuscript

work, but iSchools can help students develop skills to address problems in the remaining 20% of cases.

One curriculum for RDM that helps to address problems is the RDM Librarian Academy (RDMLA) program, a "global professional development program for librarians and other professionals working in research-intensive environments" (RDMLA, 2021). RDMLA was developed through a partnership between Simmons University's LIS program, research institutions, and Elsevier, to be used in both educational and professional development contexts. The program describes best practices in RDM that prepare librarians to work with researchers and enhance institutional capacities. RDMLA was collaboratively created by professionals working in RDM and academic partners, and it is also part of an effort to create an RDM community of practice (Thomas & Martin, 2020; Shipman & Tang, 2019).

iSchools are unique in that they focus on the intersection of people and technology. Ethics are part of this intersection, and our research found data managers must be guided by an awareness of the vulnerabilities and risks to study participants and society manifest in data, beyond those envisioned by IRBs. Knowledge is power: data creators/depositors have a competitive advantage when conducting research using their own data, while data archivists have an intimate knowledge of the data they curate that IRBs may lack (Pasquetto, Borgman, & Wollford, 2019). Thus, RDM education should be not only technical, but sociotechnical to build and support ethical infrastructure enabling the work of data scientists. For instance, we found that:

- Data breaches and malintent or negligence by researchers may require interventions to create more secure, accessible repositories that serve the public and also protect study participant identities; and

Author Manuscript

- Research data research managers sometimes balance competing interests, like a university's claim to intellectual property vs. scientific openness.

These findings, among many others, provide fertile ground for future classroom discussions for practical professional education.

Conclusion

To conclude, research data archives are a growing component of the scientific infrastructure at universities, independent research institutes, and elsewhere. These archives also tend to be expensive to operate. Technological advances and evolving norms about data reuse have the potential to increase secondary data use, and we expect reuse will increase in the future. Researchers employing methods taught in data science programs will bring about social and physical science discoveries, requiring further investment in data curation and to ensure the perpetual availability of research data. iSchools are positioned to facilitate these future discoveries and advance science more generally by focusing on ethical and responsible data management and providing a holistic view of the research data lifecycle, which may impact data scientists' work. Case-based and problem-based learning can facilitate creative thinking in classes that prepare students to contribute to RDM development in their institutions, preserving and enhancing access to research data in a rapidly changing environment.

References

Author & Author (2021). Insert title. In *Proceedings of the 2021 Meeting of the Association for Information Science & Technology*. Silver Spring, MD.

Bates, M. J. (1999). The invisible substrate of information science. *Journal of the American society for* an, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059-1078.

Author Manuscript

Borgman, C. L., Scharnhorst, A., & Golshan, M. S. (2019). Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. Journal of the Association for Information Science and Technology, 70(8), 888-904.

Caso, R., & Ducato, R. (2014). Intellectual property, open science and research biobanks. Trento Law and Technology Research Group Research Paper, (22).

Cragin, M.H., Heidorn, P.B., Palmer, C.L., & Smith L.C. (2007). An educational program on data curation. Retrieved from http://hdl.handle.net/2142/3493

Higgins, S. (2008). The DCC curation lifecycle model. Retrieved from https://doi.org/10.2218/ijdc.v3i1.48

RDMLA. (2021). Research Data Management Librarian Academy. Retrieved from https://rdmla.github.io/

Si, L., Zhuang, X., Xing, W., & Guo, W. (2013). The cultivation of scientific data specialists: Development of LIS education oriented to e-science service requirements. *Library Hi Tech*.

Hayashi, C. (1998). What is data science? Fundamental concepts and a heuristic example. In *Data science, classification, and related methods* (pp. 40-51). Springer, Tokyo.

Higgins, S. (2008). The DCC Curation Lifecycle Model. *International Journal of Data Curation, 3*(1). DOI: https://doi.org/10.2218/ijdc.v3i1.48

Imperial College London (n.d.). Data deposit agreement. Retrieved from https://www.imperial.ac.uk/research-and-innovation/support-for-staff/scholarly-communication/research-data-management/archival-and-preservation/data-deposit-agreement/

Author Manuscript

National Science Foundation. (2021). Dissemination and Sharing of Research Results - NSF Data Management Plan Requirements. https://www.nsf.gov/bfa/dias/policy/dmp.jsp

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and reuses of scientific data: The data creators' advantage. *Harvard Data Science Review, 1*(2).

Shipman, J. P., & Tang, R. (2019). The collaborative creation of a Research Data Management Librarian Academy (RDMLA). *Information Services & Use*, *39*(3), 243-247.

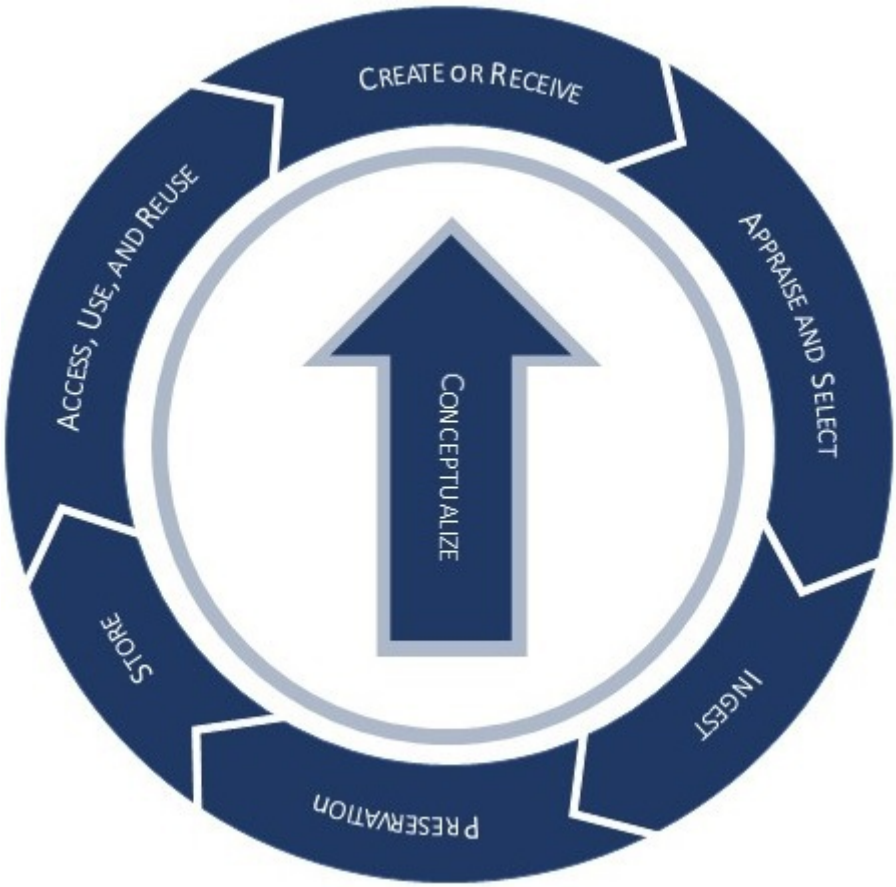
Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information systems research, 7*(1), 111-134.

Tang, R., & Sae-Lim, W. (2016). Data science programs in US higher education: An exploratory content analysis of program description, curriculum structure, and course focus. *Education for Information*, *32*(3), 269-290.

Thomas, A., & Martin, E. R. (2020). Developing a Community of Practice: Building the Research Data Management Librarian Academy. *Medical Reference Services Quarterly*, *39*(4), 323-333.

Author Manuscript

Author Manuscript



ASI_24645_jasist_figure_1_final.jpg