Issues in Science and Technology Librarianship

Understanding Research Data Practices of Civil and Environmental Engineering Graduate Students

Xiaoju Chen*
Librarian
Carnegie Mellon University
Pittsburgh, PA
xiaojuc@andrew.cmu.edu

Emily Dommermuth*
Science & Engineering Librarian and Assistant Professor
University of Colorado
Boulder, CO
emily.dommermuth@colorado.edu

Jessica G. Benner
Librarian
Carnegie Mellon University
Pittsburgh, PA
jbenner@andrew.cmu.edu

Rebecca Kuglitsch
Associate Professor
University of Colorado
Boulder, CO
rebecca.kuglitsch@colorado.edu

Abbey B. Lewis
STEM Engagement Librarian
University of Colorado
Boulder, CO
abbey.b.lewis@colorado.edu

Matthew R. Marsteller
Principal Librarian
Carnegie Mellon University
Pittsburgh, PA
matthewm@andrew.cmu.edu

Katherine Mika
Data Services Librarian
Harvard University
Cambridge, MA
katherine_mika@harvard.edu

Sarah Young
Principal Librarian
Carnegie Mellon University
Pittsburgh, PA
sarahy@andrew.cmu.edu

*Co-first authors, all subsequent authors equally contributed and listed alphabetically.

Abstract

Research data management is essential for high-quality reproducible research, yet relatively little is known about how research data management is practiced by graduate students in Civil and Environmental Engineering (CEE). Prior research suggests that

faculty in CEE delegate research data management to graduate students, prompting this investigation into how graduate students practice data management. This study uses semi-structured interviews and qualitative content analysis to explore how CEE graduate students work with data and practice data management in their research, as well as what resources and support would meet their needs. Many respondents touched on data collection, data management, disseminating research outputs, and collaboration and learning in their interviews. Several themes emerged from the interviews: data quality as a concern, as many CEE graduate students rely on secondary data for research; a gap between values and enacted practices; a connection between disseminating data and reproducibility; and a reliance on peer and self-directed learning for data management education. Based on these themes, the study recommends strategies for librarians and others on campus to better support CEE graduate student research data practices.

Keywords: Data management, Engineering, Graduate students, Qualitative

Recommended citation:

Chen, X., Dommermuth, E., Benner, J. G., Kuglitsch, R., Lewis, A. B., Marsteller, M. R., Mika, K., & Young, S. (2022). Understanding research data practices of civil and environmental engineering graduate students. *Issues in Science and Technology Librarianship, 100*. https://doi.org/10.29173/istl2678

Introduction

Researchers' data practices have changed significantly over recent decades due to innovations in technology and evolving research methods. Effective data management applied across the entire research lifecycle improves the quality and experience of research projects, full research agendas, and general lab environments. Understanding current data practices and researcher needs in these scenarios is crucial for academic libraries' research services to provide useful support.

Results from a 2019 research project (Cooper et al., 2019) showed that faculty researchers in Civil and Environmental Engineering (CEE) at Carnegie Mellon University (CMU) and the University of Colorado Boulder (CU Boulder) were leaving the practice of managing their data up to their graduate students, who were relied on to store, process, and analyze data (Chen et al., 2019; Kuglitsch et al., 2018). This finding prompted the research team to explore graduate student research data practices, challenges, and services that can be provided to better support data practices. To investigate this, the CMU and CU Boulder teams used qualitative methods to explore the experiences and perspectives of graduate student researchers in CEE. CEE is interdisciplinary, highly collaborative, and involves varied methods. Understanding graduate students' data practices in CEE can help us understand the needs of researchers in CEE and related disciplines. We explored two research questions:

1. How do graduate students in CEE work with data in their research and practice research data management?

2. What resources and support related to data are needed by graduate students in CEE?

Literature Review

Role of Libraries in Data Practices

Supporting research data management (RDM) and data sharing practices is a natural fit for libraries given the expertise of librarians in areas such as metadata, data curation, information organization, and research dissemination (Cox & Pinfield, 2014). Academic libraries began to develop formal RDM services starting in the mid-to-late 2000s, in part prompted by funder mandates for data management plans (Antell et al., 2014). Thus, there is now about 15 years of literature on data-related library services, including many case studies highlighting the efforts of individual libraries to meet researchers' RDM needs. To help address the need for funding agency-required data management plans, libraries provide both infrastructure in the form of data repositories (Choudhury, 2008; Witt, 2012) and expertise on describing and preparing data and drafting data management plans (Rolando et al., 2015). A recent report from Ithaka S+R showed that among 120 universities and colleges under study, libraries are important providers of research data services with consultations and training events (Radecki & Springer, 2020). Promoting data sharing has also emerged in the past 10-15 years as an important role for libraries (Kim, 2013).

As these services have matured over the past decade, more research has been done to assess library RDM services broadly. Antell et al. (2014) surveyed Association of Research Libraries member libraries in Canada and the United States to identify the roles and responsibilities of science librarians in RDM specifically, and Pinfield et al. (2014) surveyed librarians in the UK in a similar study. Both studies found that most libraries were only beginning to offer RDM support, and there was a level of uncertainty regarding the need for training, staffing, and infrastructure to support these novel services. A few years later, Cox et al. (2017) found a more developed landscape of RDM services in which libraries demonstrated significant leadership, especially emphasizing advocacy and advisory services.

Newton et al. (2010) explored Purdue Libraries' involvement in data curation through a task force that identified and acquired sample data collections for a prototype data repository. The researchers described several librarian skills essential for collecting data for an institutional repository, including advocating for the value of broad data access through an institutional repository, fluency in data repository capability, and research awareness. Similarly, Lage et al. (2011) produced a case study at the University of Colorado Boulder on scientific data curation, which is the organization and integration of data collected from various sources. The researchers used an interview technique to ascertain the willingness of scientific researchers to accept library contributions to scientific data curation. They found factors such as proximity to data curation, lack of support for data curation, personal views in favor of sharing, and working in certain fields including environmental engineering indicated a willingness to work with libraries on data curation. The researchers noted that graduate student researchers were

often involved in data curation (and thus willing to accept library contributions) because they had inadvertently assumed this responsibility for their research group.

Still, libraries face challenges in implementing integrated data management services and promoting good data-sharing practices among emerging and established scholars. Sufficient staffing, the need for technical expertise, and the cost of infrastructure have all been cited as challenges to the success of well-rounded and integrated RDM services (Latham, 2017; Tang & Hu, 2019).

Data Practices in Graduate Students

Graduate students have been responsible, yet underprepared, for data practices throughout the whole research life cycle (Carlson et al., 2011). Sharma and Qin (2014) surveyed students at a mid-size research university to investigate graduate students' knowledge about data management. The results collected from 173 students in social sciences, natural science, health sciences, and engineering showed that they lacked awareness of data management policies, technologies, and practices. Pasek and Mayer (2019) surveyed graduate students to understand RDM practices across diverse fields. Students reported that ethics and attribution were the most important aspects of data practices, followed by visualization. Data curation and reuse, however, were identified as the skills that most needed improvement. Graduate students also indicated that self-directed learning was their most frequent means of acquiring RDM skills.

Carlson and Stowell-Bracke (2013) showed that graduate students played a significant role as data collectors and generators and concluded that understanding graduate students' needs was essential to address the real-world needs of research communities and labs. These needs, however, could not always be addressed due to constraints such as the absence of support within labs and the lack of a larger disciplinary culture supporting data sharing and reuse. Similar interviews by Valentino and Boock (2015) and Wiley and Kerby (2018) found that students were willing to engage in good data management practices, though they generally lacked knowledge of best practices. In addition, there was a significant gap in communication and collaboration between the principal investigators who manage research projects and the graduate students who perform specific data tasks.

Data Practices in Civil and Environmental Engineering

A few studies have explored research data practices of researchers in fields related to CEE. Those studies highlight challenges regarding different aspects of data practices, including data accessing, processing, and sharing, and proposed solutions (Montgomery et al., 2007; Pejşa & Song, 2013; Satheesan et al., 2018; Shahi et al., 2014). In each case, specific challenges faced by researchers facilitated the recognition of data management needs and created a motivation to value data management activities.

Case studies of RDM are beginning to enter the literature in fields related to CEE. Schröder and Nickel (2020) produced a study of research data management as an "elementary component of empirical studies." Using landscape ecology as an example, the authors demonstrated how to integrate RDM into research design, because sharing

and reusing empirical data requires good RDM practices. Petters et al. (2019) provided an example of a library data consulting service supporting improved data management for long-term ecological research in the fish and wildlife conservation department at Virginia Tech. Carlson et al. (2011) produced a foundational work exploring the extent to which faculty and students are prepared to integrate data management into their workflows. The researchers used semi-structured interviews of research faculty including some from civil engineering to assess graduate student course performance. Results indicate a need for a data information literacy program to prepare students for data management work.

Researchers at the University of Minnesota Libraries conducted a case study review of the data management practices of four graduate student researchers and one faculty researcher from a civil engineering lab. The study highlighted the lack of practice and training in research data management practices (Johnston & Jeffryes, 2014). A related study focused on faculty perceptions of graduate student's data information literacy skills. Faculty who worked in civil engineering and related fields including landscape architecture, hydrology, computer science, and natural resources were interviewed, and interviews highlighted uncertainty and lack of data competency for both faculty and their graduate students. There was a lack of formal training and policies; consequently, students were learning from their advisor at the point of need, and through self-directed "trial and error" (Sapp Nelson, 2015). The current study builds on these two studies but focuses on graduate student data practices from the perspective of graduate students working in CEE at two universities.

Methods

Institutional Contexts

This research focuses on understanding data practices among graduate students in CEE at two doctoral-granting institutions: Carnegie Mellon University (CMU) and the University of Colorado Boulder (CU Boulder). CMU is a private institution with approximately 14,000 enrolled students. In 2020 the College of Engineering enrolled 3,720 students, with 169 graduate students in civil and environmental engineering and 62 in engineering and public policy (Carnegie Mellon University, 2020). CU Boulder enrolls 35,000 students, about 7,500 in the College of Engineering and Applied Sciences, of which about 200 are graduate students in civil and environmental engineering (University of Colorado Boulder, 2020). According to U.S. News & World Report, both institutions' engineering programs are top ranked (U.S. News & World Report, 2021).

Research Methods

The research protocol was submitted to and approved by the Institutional Review Boards at both institutions. Interviewees were recruited via email outreach and were offered a ten-dollar Amazon gift card as a participation incentive. The team collaboratively developed a semi-structured interview guide (Appendix 1). The interviews used open-ended questions to solicit students' research data experiences and practices, including their challenges and learning experiences. The interviews were transcribed, and the research team reviewed the transcripts for accuracy and to remove

identifying information. The team applied a qualitative content analysis approach to code the interviews to identify themes or patterns (Cho & Lee, 2014). A coding scheme was developed in phases. First, each team member asynchronously reviewed three to five transcripts and generated an individual set of open codes. At an in-person research meeting in January 2020, each team member shared and discussed their open codes, and an initial common set of codes was generated. The team also developed a set of descriptors to apply to each interview (Appendix 2).

The draft codes were trialed to ensure team members were similarly applying codes to a common transcript. The team then finalized the coding scheme and used Dedoose qualitative analysis software to apply the codes to the transcripts in a series of rounds. In the first round, each team member coded two to three transcripts. Another team member then reviewed the transcript and codes, made suggestions, and had a conversation to reach consensus on the final codes. Once all transcripts were coded and a consensus was reached, the team used Dedoose and Python to explore and analyze coded passages and descriptor features to identify trends and themes.

Results and Discussion

Our eight-person research team conducted 19 interviews with research-track graduate students during the summer of 2019. Recruitment efforts resulted in 11 interviewees at CMU and 8 interviewees at CU Boulder. Descriptors were assigned to each interview to describe the nature of the data and research methods the interviewee used (Appendix 3).

Graduate Student Research Data Practices

Interviews were analyzed to answer the question: "How do graduate students in CEE work with data in their research and practice research data management?"

Data Collection

Over half of the CEE interviewees worked with self-described small data sets (Figure 1). Interviewees worked with primary data, secondary data, and a combination of both types (Figure 2). Secondary data used by the interviewees came from government, academic, industry, or non-governmental organization sources, and some interviewees worked with multiple data sets from multiple sources.

For those interviewees who worked with primary data, the data were collected through multiple means including surveys, observations, interviews, experiments, and environmental monitoring. Amongst all factors, instrumentation was generally the greatest challenge, followed by survey design and raw data cleaning. Challenges brought by instrumentation included extra time spent due to a poor internet connection at the data collection site, additional work needed with equipment manufacturers, and low-quality instruments collecting low-quality data. Interviewees mentioned the challenge of encountering resistance from survey participants due to over-surveying, and the challenge of time-consuming data cleaning processes.

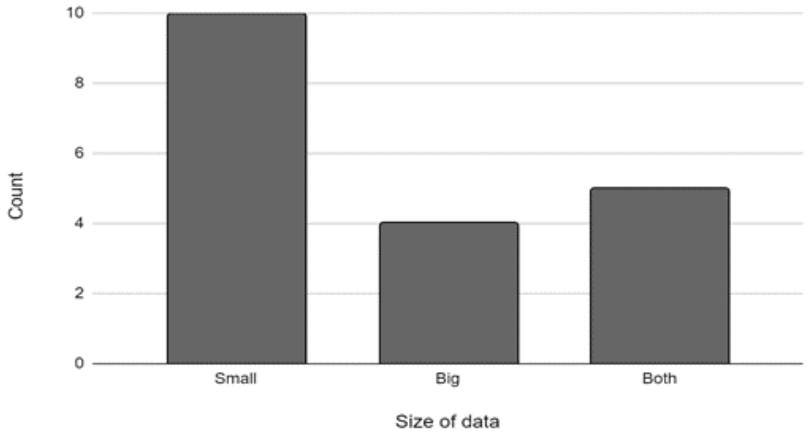

Figure 1. Size of data, as defined by interviewees (n=19)

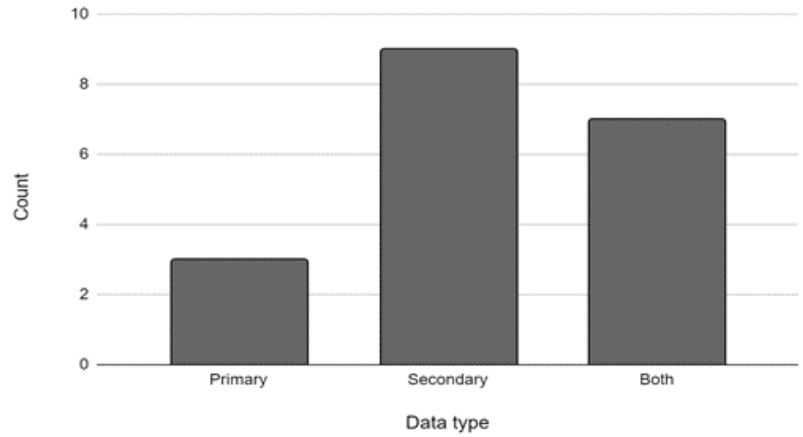

Figure 2. Type of data, as defined by interviewees (n=19)

According to interviewees, acquiring and using secondary data frequently presented challenges, including the cost of data, inability to obtain needed data because it is owned by a private entity, security and public safety concerns, data inconsistency, and not meeting researcher needs.

Several interviewees described leveraging a collaboration or connection to gain access to a secondary data source. A few interviewees mentioned a colleague referring them to an openly available source, and several were involved in collaborations where one person or group provided data. A few interviewees mentioned reusing code or data from within their research group, for example:

> Actually, I learned from some old code that my advisor's student from ten years ago had. [...] He had a .zip file of everything. [...] It is on my advisor's website, but it wasn't formally published. And, my advisor forgot about it, and I found it [...]. And, even though it was in MATLAB, and [...] it was for a different project, but I opened it and I was able to bring some things, like some tricks or some techniques – I was able to help my own research.

Collaboration could range from simply handing over data sets to making formal agreements to provide them their data. Several interviewees mentioned dealing with access to information that is proprietary or not yet public:

> As long as I'm sponsored by my company and they have access to this service to help the business, so they consider my work as a business need. But there was always a lot of personal communication element into this. It was not [a] 100 percent official request. But I sent emails to the people [...] who are the [...] proponents of this project, who use these models in the company [...], because I've known them before. So they helped me and sent me [...] access to everything.

Sometimes advisors brokered the collaboration or served as primary collaborators who disseminated the data to their graduate students. The collaborations could also involve expertise, for example, one CEE interviewee arranged for an electrical engineering student to write code that would pull data from a social media API; another hired a local team to interview community members in another country.

There were only a few mentions of searching the literature to locate data sets for reuse. Incompleteness of available data was mentioned as a challenge with data sets found through literature searching. Sometimes an interviewee discovered publicly hosted data that was referred to in the literature. Targeted web searching for data was another, more frequently mentioned strategy interviewees used. The interviewee would have a specific information or data-producing body in mind, and would then search or browse specific websites for needed data. Sometimes data was scattered over different information producers. Examples of data producers from interviews included state Oil and Gas Conservation Commission, United States Geological Survey (USGS), state geospatial data portals, National Aeronautics and Space Administration (NASA), United Nations' Food and Agricultural Organization (FAO), and the Intergovernmental Panel on Climate Change (IPCC). Interviewees described how data gathering from these sources could be onerous: "And so they have databases that if you kind of dig through their website, you can get to and download and it'll again, just be like massive, unorganized CSV files."

In addition to general difficulty in finding data, interviewees reported the quality and usability of secondary data as particularly challenging:

> It's more like there are these databases [...]. They're just not great, and they're sometimes hidden, but they are there. So it's more like diving into the haystack to pull that out and then downloading all that data. [...] The USGS has better data, but it's [a matter of] finding it. [...] Initially when I download it, [it is] whatever file they offer. Hopefully CSVs, sometimes more like esoteric file types that [...] I have to [...] look up tutorials on how do I even open it? I just go, that's [the] state government [...] and then manipulate it.

Even when interviewees were able to procure data that might be suitable for their purposes, it could be challenging to determine variables, units of measurement, scale, geographical coverage, and more. Un- or under-documented data also made it difficult to learn about the context, assumptions, and other particulars of data collection processes and protocols that could have significantly affected downstream analysis:

> Sometimes it just needs a lot of digging into the assumptions behind it and how it was collected. This is the data that really answers the question or it's something

> else. Sometimes there's a lack of [...] communication or labeling the data that you don't know [...] if this is the specific parameter that you look for or something similar.

The quality of secondary data is usually discussed by interviewees as a challenge to evaluate, prepare, and analyze. Indeed, the availability of high-quality or usable data affected interviewees' research agendas and confidence in their results and that of the field in general:

> I think one of the biggest challenge[s] is the availability and also the quality of the data. 'Cause unlike some other field[s], a lot of times we need to rely on other sources. We cannot do those experiments ourselves and produce our own data. We have to rely on external sources. And so that brings a huge challenge of availability and also just reliability of that data, how certain are you [in] those numbers.

Due to the importance of the secondary data used in CEE research, the quality of the secondary data introduces a unique set of challenges and considerations in this field.

Data Management

CEE graduate student researchers interviewed had a wide range of experiences with, knowledge of, and practices for research data management. Several interviewees described how they define data management. The broadest definition an interviewee shared was:

> I think it's a very broad term, data management, but in my opinion [it] is all of the process of collecting, treating the data, using the data, and producing more data, [...] And, that could include models, statistical tools, raster data that's used through GIS [...] it would also include like saving the data and having it organized. Organizing the data to be able to use it and access it.

Another interviewee stated, "I just think it's about how to get your data and then store it, and then use it and document it," while a third interviewee focused on organizing and naming files, version control, and backup in their definition. Interviewees could share their ideas about what data management is, but they were much more unsure about data management plans. When asked if their research group or grant had one, most expressed uncertainty about if they existed or not. For example, "I haven't seen much in the grant. Yeah. It might be. I'm not super sure." If interviewees are not exposed to the data management plans for the grants they work on, they may not be able to follow the plans and will lack a big picture understanding of long-term research data management.

Many interviewees used commercial cloud storage services (e.g., Google Drive, Microsoft OneDrive, Box) for storing, backing up, and sharing data with collaborators. It was common for interviewees to use cloud storage in addition to a copy on a hard drive, external drive, or flash drive. Other storage/backup strategies included storing data in their email or using their own servers. For code, researchers often put backups

on GitHub, in addition to another location. One interviewee said they were not backing up their data or practicing version control: "because I'm bad at it, not because I think there's not [...] value." Even those who were not practicing backup with their data still understood its value.

Most interviewees also described using the sharing features of their cloud storage tools to share their data with collaborators or so their advisors could access and review their work. Several others shared data via email, leading one interviewee to just leave their data in email and store it there. Those with larger data had more variety in their data sharing methods, including shipping physical hard drives to collaborators and using research computing services at their institution to transfer data to collaborators.

Interviewees valued storage and sharing tools that integrate version control and organizational features, such as Box and GitHub. Interviewees reported a variety of strategies, but overall most interviewees practiced some form of version control. One interviewee shared how it became apparent that they should do version control:

> I guess in research things just change a lot really fast, and maybe I'm starting to [...] realize, oh, I can't assume that it will continue working. I have to save and commit and make sure I have working versions. And then, once I change something, I can figure out what changes made it stop working because there are so many ways to make something stop working.

Interviewees also described variety in their organizational strategies, but most interviewees had developed some strategy to organize their various files. One interviewee described how they developed a file organization system when they were confused by all their files and spent a great deal of time looking for what they needed. Through trial and error, interviewees found version control and organization systems that worked for them.

The strategies interviewees used to document their research and data practices for current use were also varied. Several described a document, spreadsheet, or other files where they kept running notes about their work. Several discussed commenting on their analysis code or scripts, using the commenting features of the tool they are working with, such as GitHub. Several interviewees described a mix of different documentation strategies and places:

> I right now [...] use kind of a pen and paper, like when I'm doing stuff in R, what's working, what code isn't working and what I'm seeing. And as far as like a meta way of analyzing it, if I find something that I think is important for the results then I just have a Google notes sheet going with [...] little notes and [...] things that I'm finding and things that I'm seeing when I'm organizing and analyzing data. R is actually pretty good, too, because you could [...] write notes into the sheet. And so code, [...] you use a code and it doesn't work you don't have to delete it you can just keep it there and say this one didn't work, don't use this one again.

This mix of strategies meant that notes were spread out across different mediums and programs, which could lead to difficulty using or sharing the documentation later. For at least one interviewee using a mix of strategies and places, it also led to a feeling that their work might not be completely documented:

> I find it's hard to document what I've done to the data using those tools. I end up with a bunch of text files or [...] commands copied and pasted into them. And I know there's probably a better way of doing it, [...]. So yeah, not everything is incredibly well documented yet.

Thus, interviewees understood the need for documentation for their own use.

The need to create documentation that others could use was not as clear to interviewees. For example, one interviewee stated, “but I don’t think people require that.” Another interviewee needed to share their work with collaborators, but felt time was a limiting factor so they skipped creating documentation and verbally described their work to collaborators:

> Well, there are two ways. You either spend more time in the document commenting, making notes and make everything as clear as possible, or more time efficiently you get on a conference call and would just share your screen explaining what you did.

Documenting their work for others was not a consistently high priority for interviewees.

Interviewees discussed their data management practices for their personal needs and current use of the data. However, they largely had not planned for what would happen to their data after they finished their graduate studies. When asked about long-term data management, several interviewees stated that they had no plans for long-term management, while several others stated that the current data storage they were using operated as their long-term data management strategy. Other interviewees guessed that their advisor or lab group would manage the data for the long term, for example, one interviewee said: “I'm sure that we do. I'm not sure what it is. Yeah. I would guess that we do just because my advisor is usually really on top of things like this.” Another interviewee said: “I'm sure they'll store and manage the data,” highlighting their uncertainty over if there was a plan for long-term data management.

Several interviewees discussed their need to develop a plan to pass data on to new graduate students joining the research group, but still indicated uncertainty around how this would be done. For example:

> I haven't thought too much about it. We do have that new student I mentioned who's coming in. I think if she begins to be interested in it then probably sharing it with her and letting her look through it.

This highlighted that students who see the need to preserve their data for long-term use did not know best practices for long-term storage and preservation.

Finally, several interviewees had considered that they themselves might want their data in the future. One interviewee shared: "I'll keep them with me for sure. This is a year's worth of work that I spent here [...] in case I needed them after for my post-school career or projects, or if I need to extend the work afterwards." This interviewee was mostly considering their personal future use, but another had considered the possibility that someone else might want the data and stated:

> But I think I will definitely preserve that for future use. I'll keep a good documentation of where data comes from. [...] I [...] clear[ly] reference [...] where these data come from, and make some notes of the whole spreadsheet if I store everything in spreadsheets so that I can come back to it later maybe for sharing in the future.

This interviewee put thought into their research data practices so that they might be able to effectively disseminate their data for others to use in the future.

Disseminating Research Outputs

Students were asked about their practices for disseminating outputs of their research, including their data. Interview questions focused more on data, code, models, visualizations, etc., rather than traditional research papers (Appendix 1). Some interviewees referenced norms for disseminating or publishing in their research community. When students discussed disseminating, they mainly shared graphs, visualizations, or documented workflows. A few mentioned a willingness to disseminate data that is available elsewhere as opposed to unique data they collected. Some disseminated models or code; however, almost no one mentioned that they had disseminated a data set. Some reasons for the lack of disseminating data sets included the need to clean up the data or code and wait until there was a demonstrated value. For example, one interviewee mentioned:

> When they get to be good enough, of course, yeah. I'm still working on them. Of course if I can get to the point that I'm confident to put them online and people would benefit from them, if they can add value, by all means I can make it publicly available.

Others felt ambivalence for several reasons such as territoriality:

> But at this point, my raw data, like very territorial, which I don't know if that's a good thing or bad thing or just kind of [...t]he nature of academia [...] But I think, you know, after [...] after we've published [...], might be more willing to share bits and pieces of my survey data set. But I also...like my advisor is also kind of on the same page.

And another interviewee stated:

> But so far, I'm just thinking just follow the routine. What other people are doing, I'm just doing the same thing. You cite where my data come from, publish my

> final results, the graphs, the model, how I did my calculation, without necessarily showing all the steps in the middle.

Even with the time investment barriers to make data publishable, territoriality, or the hesitancy to do something new, many of the interviewees expressed that disseminating data and code was important for reproducibility and improving their research area. Some specific motivations included getting information to practitioners, discovering unintended uses of data, and increasing the credibility/reputation of the research group. As readers of other people's research, some interviewees felt they did not have enough information to understand the method, analysis, or results presented in a published paper without access to data. Reproducible data outputs are particularly useful to graduate students because publications often lack sufficient information to reapply a study or method in a new context—a common strategy for graduate student research. One interviewee summed up the need to disseminate information beyond just text and equations in a publication:

> Well, replicate, learn, and improve on it. An article is just text and equations, but I think nowadays what we do is really hard to capture in just text and equations. …there's a lot of hidden methods and understandings that are hidden in the code that somebody can't pick up just from the paper.

While interviewees were not asked about their understanding of or skills in reproducible research, nearly half identified it as an important concept.

Collaboration and Learning

Interviewees described working with various colleagues, including their advisors, lab group members, collaborators at other institutions, and other graduate students. Interviewees detailed communication practices for collaborating, including through formalized lab meetings, document sharing through platforms like Google Drive and Box, and presenting work and receiving peer feedback. Communication and collaboration often resulted in learning opportunities for interviewees.

Learning from peers, such as fellow graduate students in their own or related programs, was a prominent method for learning research data practices based on our interviews. Interviewees described learning skills such as version control, code documentation, and practices related to reproducibility from their peers. Interviewees leveraged collaborations to access expertise they did not personally have, such as in machine learning or experimental design. A notable avenue for communication and learning from peers was the informal and serendipitous interactions that take place day-to-day based on physical proximity. One interviewee noted:

> So it's very informal, and we talk to each other all the time about research. It's been actually a great thing. If I would have one recommendation for faculty members or [...] departments, is to group students in their research group if it's a dynamic group because it has made us all much more productive, because instead of trying to figure out on our own, we can ask each other, and everybody has ideas.

Similarly, another interviewee stated:

> I guess [...] in the informal collaboration, that's probably the people who are both in my lab, [...] under my advisor and also in the same physical office space…[W]e all have different projects but...most of us have to use stats and do a bunch of regression models. And none of us are great at it. It's all collaboration in that regard. More like problem-solving on a day to day level.

These interviewee's statements highlighted the value of proximity to peers and advisors for conversation and learning.

Advisors often tasked more senior students with the role of training incoming students. On the other hand, a senior interviewee also noted learning new skills from an incoming student:

> I was eating some humble pie when just working with this undergrad. His commits were just so clear that... we wouldn't have to meet…[I]t really just streamlined our communication just because everything was right there in front of me. So...I learned a lot from him and then I was able to incorporate those ideas with helping me understand my previous or past work that I had done.

Learning was not necessarily based on seniority, but on skills individuals bring to the lab group.

While peer-to-peer learning was common, interviewees acknowledged that it might not be the best way to learn effective methods and best practices:

> I don't know any grad student who's thought of this who actually knows. I think we're all just quoting someone who graduated four years ago who was probably doing the same thing. And that's just kind of how a lot of the grad student experience is, I feel like.

Peer-to-peer learning of data-related practices featured quite prominently compared to learning these skills directly from advisors. In general, we found that advisors and advisees played different roles in the research life cycle. Advisors were often in charge of the project and responsible for the long-term preservation of the research outputs, and graduate students were the ones who collected the data, built the models, and dealt with details in the research. For example, one interviewee said: "A lot of the data practices fall on the graduate students and [...] not the advisors." One interviewee mentioned that their advisor did not really look at code written by the interviewee because they trusted the interviewee's research ability. A typical communication practice was that the interviewees generated research outputs and shared plots, summaries, or reports with their advisors through meetings, progress reports, or informal communication. The advisors did not necessarily want to see their raw data. However, tools such as Dropbox, Box, Google Drive, and GitHub were used to store and share data, scripts, and results, so advisors could access interviewee work as needed. Some interviewees mentioned that as they progressed in their program they became even more independent from advisors.

Despite the different roles advisors and interviewees were playing, and the heavy reliance on peer learning, the interviewees did describe learning some practices, skills, or knowledge from their advisors. Some interviewees mentioned that their advisors provided guidance on where to find data and how to manage data and encouraged good data management and data dissemination practices. One interviewee said: "[M]y advisor was the one who encouraged me to use [our institution's data repository], for example, to put everything available from the past project. … there is this initiative of making everything open source and available to others." The advisors have a significant amount of experience with the scientific process, and interviewees absorbed their advisor's approach to guide their own research practices. For example:

> There is a blog…started by… my academic grandfather [who] was my advisor's advisor. … [M]y advisor added to it in his graduate school. People in our research group also add to this blog. … it covers everything from how to do literature review ..to .. very specific plotting package… I could very much see… continuing [...] those kind[s] of traditions as far as … not just the things I've learned, but the actual tools … contributing in the future.

Even in an environment with robust peer support, interviewees emphasized the value of the advisor's inputs.

Interviewees also described self-directed learning. This was especially notable in areas related to managing data for current use. Interviewees referred to simply working through problems and challenges as they arose, expressing a habit of self-directed learning and trial and error. For example, one interviewee noted, "[I'm] just Googling stuff…" and "mostly just self-taught I guess," while another stated, "it definitely takes practice, and trying it, and messing it up." Another interviewee stated:

> Definitely it has been something that I have learned over, [...] my research experience [...] from undergrad research, it's past research experiences and, [...] learning what works and what doesn't. Through trial and error, which is not the greatest way.

One interviewee described their reasoning for self-directed learning as "...mental block of, like, you're going to ask dumb questions, and you don't want to because you're in grad school." They framed this approach as part of "the research learning curve," indicating interviewees' view that self-directed learning via trial and error was part of the graduate school experience.

Resources and Support Needed for Research Data Practices

This section addresses the research question: "What resources and support related to data are needed by graduate students in CEE?"

Data Quality

The number of interviewees who used secondary data in CEE is notable. A major issue associated with using secondary data is evaluating the quality of that data and its

fitness for use, as illustrated by Cai and Zhu (2015). Navigating data quality challenges provides students opportunities to develop essential data skills for future research and professional work. Variable data quality reinforces the importance of data literacy and data evaluation skills. Choosing what data to rely on for a particular use and how to apply them is a fundamental part of graduate-level education (Carlson et al., 2011). Data are never perfect and are always created in a particular context. Learning how to interpret results from limited, flawed, or problematic data is realistic for many industries and domains. Implementing transparent and reproducible data processing and analysis pipelines is part of a suite of critical thinking and problem-solving skills that contribute to high-quality research.

Students may alleviate the burden of evaluation by leveraging an academic network to access data shared among researchers, suggested by advisors, or obtained via industry or public sector collaboration. For example, one student worked with a government agency to acquire specific data tailored to meet research requirements. While a bespoke data set may not necessarily be of higher quality objectively, it is likely more fit for use in this student's particular context. In other words, collaborating with the original data collector may be considered a more reliable and much faster evaluation method.

Another technique students may use to identify data with a minimum standard of usability is to use a data set that was used by another researcher, identified either in the literature or through personal or informal exchanges. Data found on websites or in portals do not usually have the same presumption of quality. Further research is required to identify whether data shared among academic networks is indeed of higher quality and usability; however, these interviews demonstrated that many graduate students perceived this to be the case. There may be a need for instruction and resources that help students independently evaluate data more efficiently or consider what characteristics might indicate higher quality data.

Gap Between Data Management Values and Practice

Limitations such as lack of knowledge, tools, and time, prevented interviewees from always following best practices, indicating a gap between values and practice.

The results highlight how convenience and immediate needs drive graduate student researchers' data management practices. The ubiquity and simplicity of cloud storage tools and GitHub ease the adoption of practices for backup, sharing with collaborators, organization, and version control. Graduate student researchers may also be motivated by experience, such as specific challenges or impactful successes. Indeed, interviewees talked about how not being able to find files prompted them to develop an organization system or how frequent changes in their research necessitated saving working versions of their code as they progressed, so they could go back and figure out when things stopped working.

The results also illustrate a variety of documentation strategies, practices, and motivations. Some interviewees were motivated to document for their own future use, such as the reward of seeing their progress and using their documentation when working on presentations and visualizations. Some needed to document their work for

close collaborators, which may result in documentation, though in at least one case an interviewee chose to share their work via screen sharing in a virtual meeting because creating documentation was too time-consuming. Therefore, even when there is an immediate need for documentation, time and immediate convenience may still influence graduate student researchers' adherence to best practices.

Several interviewees had inherited data from others with documentation that varied in usefulness, and this documentation, or lack of documentation, motivated interviewees' desire to document their research, illustrated by an interviewee who stated:

> [Data analysis] gets confusing when you get other people's files because you don't know exactly what parameters they had when they ran their model or [...] something like that. So yeah, documenting stuff is so important. So I would say that is something that I spent a lot of time thinking about.

While their experiences may motivate documentation, seeing its value does not automatically translate into knowing best practices for documentation. Thus, graduate student researchers may be stuck in a loop: experiencing poor documentation when they work with secondary data, then valuing but not having the support and skills to effectively document their own data. As a result, the next student or researcher who uses their data also receives poor documentation and gets stuck in the same loop. There is a need to connect students' values to learning opportunities that allow them to act on their values. Good data management practices take time and thought to develop. Students need additional support and guidance for building habits and opportunities to practice using tools and systems that integrate good documentation and data management strategies.

The overall lack of planning for long-term data management reported by interviewees may indicate that long-term data management is not valued as much as data management practices that have more immediate benefits. This observation agrees with findings in a similar study on researchers in the social sciences (Jahnke et al., 2012). Still, some see the value in keeping their data for themselves and some recognize they may need to pass data onto another student, and at least one thought someone else might have interest in their data in the future. This observation indicates they value their work and understand its reuse value and may represent a desire for improved access to tools for and instruction on long-term data management practices. These findings echo Valentino and Boock (2015) and Wiley and Kerby (2018), who found that students were willing to engage in good data management practices but generally lacked knowledge of best practices.

Disseminating Data and Reproducibility

Interviewees also described their values towards disseminating, or not disseminating, their data or other research outputs more widely. Their values related to this were often connected with their opinions on reproducible research. Reproducibility was not on the interview guide; however, it was a topic that nine interviewees brought up without prompting and therefore became a significant theme. A few interviewees had not gotten far enough in their research to think about the logistics of disseminating data; however,

for others there was a tension between the need to publish and the motivation to disseminate that often prevented early dissemination of data or code.

Disseminating data and code takes work beyond what has been expected from researchers in the past. This additional time investment was also a barrier to data dissemination amongst civil and environmental engineering faculty (Cooper et al., 2019). Two interviewees did not report seeing any benefits of data dissemination as a student researcher. Even those who thought dissemination is critical for good scientific practice and reproducibility, including one who considered themselves a "reproducibility evangelist," found it hard to devote the energy to regularly documenting their work:

> I was just talking to the guy that sits next to me and I was joking that I would be such a hypocrite because I'm such a reproducibility evangelist. [...] It's probably going to take me a week to get this thing to where it would get fully documented and reproducible. And I was like, you know what, I really could just not do this. But yeah, I agree that it's good to actually practice what you preach.

For the purposes of this discussion, reproducibility is largely used by students to refer to a general spectrum of computationally reusable data outputs. A data package may be more or less reproducible depending on the ability of a user to reproduce results from a paper, apply a computational method in a new context or with new data, or reuse parts of code. Reproducibility for CEE students appears to be closely tied to whether a specific model can be applied using different data.

As noted, a majority of interviewed students used secondary data. Several interviewees explicitly cited the difficulty in using secondary data as a primary influencing factor in valuing reproducibility. Data are collected for various purposes and in various contexts and are often published or released without sufficient metadata, documentation, or processing guidelines to support reuse (Sadiq & Indulska, 2017). While messy, obscure, and poor-quality data products are difficult to use, even high-quality and reproducible data sets may be insufficient for specific reuse applications due to scale, precision, format, or other fundamental data set properties (Trisovic et al., 2021). Data that are difficult to reuse or are irreproducible affected interviewees' perception of data quality and the importance of reproducibility:

> [A]nother reason why I really wanted to do that, like putting all the data and code and everything is because for that project, too, I spent a very long time trying to reproduce someone else's results, and it was just crazy. [...] [J]ust by following what they did in the paper[,] I couldn't get anything close [to] their results. And I was not doubting [...] that their results were wrong; it's just that there are intermediate steps that they did not say in the paper that were actually very important to get those results.

Interviewees' direct experience with other researchers' insufficient research data practices influenced their own research data practices.

Other reasons for valuing reproducibility include verifying results, journal requirements, establishing a credible reputation, the high value of reproducible outputs in several academic and software communities, and the usefulness of understanding past work on one's own project. While interviewees' discussed journal data dissemination requirements as partially responsible for their exposure to reproducibility, funding agencies' requirements are conspicuously absent. This finding may be because graduate students are less likely to be primarily responsible for large grants that require data management and data dissemination plans. Students are more likely focused on publications and, therefore, more acutely aware of their data dissemination responsibilities in the context of journal requirements.

Interviewees reported learning about reproducibility concepts and value from academic associations, institutional support organizations, peers, and personal experience. Interestingly, none of the students identified either an advisor or any formal data management coursework as contributing to their understanding of reproducibility. This finding may suggest a need for practical, reproducibility-specific training in graduate curricula, which could include instruction on best practices for disseminating data and other outputs.

Learning and Communicating Data Management Practices

The graduate students in our study commonly referred to peers—other graduate students—in the context of learning about data analysis and management. For example, our interviews indicated the early manifestations of skill-based collaboration amongst graduate students in CEE. As discussed in previous work on research collaboration (Beaver, 2001), collaboration often occurs to bring analytical skills to a project that a team may lack. This points to a need for time and space for graduate students to discuss their ideas and questions with their peers.

It was notable that several interviewees mentioned the value of face-to-face and serendipitous interactions with peers and faculty. This finding raises the vital question of how peer-to-peer learning may be affected by remote work environments and limited interaction. Recreating these types of interactions is challenging in an online environment; nevertheless, there is also a need for online spaces outside of formal group meetings facilitating opportunities for peer-to-peer learning.

Findings indicate different roles for faculty advisors and graduate student researchers throughout the research project. These findings, combined with findings from previous research that CEE faculty researchers were relying on their graduate students to manage their data (Chen et al., 2019; Kuglitsch et al., 2018), demonstrate that there may be a need to provide instruction and guidance to faculty researchers in research data management. Such instruction could help improve research and lab workflows and guide CEE graduate students.

Graduate students interviewed expressed a sense that figuring out questions for themselves was part of the graduate experience, as well as a desire not to expose ignorance or inexperience. This conclusion follows Gardner's (2007) findings of the push and pull between guidance and learning to become an independent scholar as part

of the graduate experience. On the other hand, students also sometimes experienced frustrations related to self-directed learning, especially around lost time, redoing work, and a sense of uncertainty. One interviewee expressed this frustration: "[I]f you don't know exactly what you need [...], you don't know what to ask for, that [...] becomes a bad spiral for your first year or so of research." A part of this challenge is caused by the desire not to look uninformed despite acknowledging the benefits of reaching out. One interviewee struggled with the desire not to ask dumb questions, but also mentioned that consulting with senior graduate students had been helpful. Thus, there is a need for resources that support students efficiently "figuring things out for themselves," as well as spaces where graduate students feel safe to ask questions and expose that they do not know something.

Recommendations

Liaison librarians for STEM fields provide a variety of research supports to graduate student and faculty researchers. This could include data support, or liaison librarians could be working with data services librarians to provide expertise and tools to support data practices. Additionally, our findings highlight the range of resources and individuals who support graduate students' research data practices. This is echoed by Radecki and Springer's (2020) findings that research data services are offered by a variety of groups on campus, including information technology and research computing, academic departments, independent research centers and facilities, and professional schools. Librarians can familiarize themselves with the landscape of research data support providers at their institution, and the departments in which graduate students do their research, to understand services and curriculum, seek collaborations, and support the education of graduate student researchers. In this section, we provide recommendations for these librarians based on the findings from our research.

A high proportion of the interviewees were working with secondary data. These secondary data may have variable quality and fitness-for-use, and graduate student researchers can be encouraged to critically evaluate the quality and usability of the secondary data they are using, and select the most feasible data to meet their needs. Additionally, secondary data often require cleaning and processing, and these were tasks consistently identified as pain points and thus a space for guidance. One approach we recommend to meet this need is that institutions or libraries offer The Carpentries workshops (https://carpentries.org/) to teach software development and data science skills to researchers. The Carpentries can help fill the gap in data cleaning and processing education for graduate students, as well as other gaps including data management.

The gap between data management values and practice suggests a need for tools, education, and services that encourage graduate student researchers to efficiently implement research data best practices. Librarians can seek out and share tools that ease best practices, such as backup, sharing with collaborators, organization, documentation, and version control. Documentation is an especially notable pain point for graduate students in our findings. Basic guidelines, tools that integrate documentation, and connecting with graduate students early in their careers can help build habits that

ensure documentation is created as the research progresses and is created in manageable chunks.

Many students shared misconceptions about what might be considered long-term data storage and reliable backup storage. Still, interviews show that graduate students would value this kind of expertise and the misconceptions speak to a real need. At many institutions, libraries and research computing services may be well positioned to meet this need with existing infrastructure and support, such as institutional repositories and large-scale data storage. We recommend that libraries and research computing ensure graduate students are aware of these existing services, and how these services can be used for data management.

As more funders require data publishing, and more graduate student researchers value open data as illustrated by our interviewees, libraries can help make it easy to distribute and publish data. Specifically, we recommend providing institutional repositories, Digital Object Identifier (DOI) registration services, and support for reviewing data submissions for alignment with the FAIR Data Principles, which outline data should be findable, accessible, interoperable, and reusable (FORCE11, 2014). It is also important to ensure graduate student researchers are aware of these services.

Our findings suggest that reproducibility-specific training is another crucial gap in graduate curricula that students need and would appreciate. This combined with interest in data dissemination, highlights the value of outreach, programming, and materials around Open Science practices. Such practices allow scientific information, data, and outputs to be more widely accessible and reliably harnessed (UNESCO, 2020). Education around reproducibility practices specific to disciplines may also help fill this crucial gap.

CEE graduate students showed a strong tendency to learn data management practices from each other. Libraries can facilitate this type of learning through train-the-trainer models, where graduate students learn data management from librarians and then go back to their lab groups to train their peers, and their faculty advisors. Libraries can also host events that bring graduate students together to share their expertise in working with data. CEE students also reported a high amount of self-directed learning, so it is recommended that libraries provide asynchronous learning tools and self-teaching supports, such as online guides, tutorials, and readings. These resources could fill graduate students' self-directed learning practices with higher quality information and advice, to build data practices that adhere more closely to best practices. In planning peer and self-directed learning opportunities, we recommend especially marketing them to new graduate student researchers. If students learn best data practices early on in their research, they can build good data habits for their entire career and pass those habits on to more of their peers.

While interviewees illustrated the nature of peer and self-directed learning opportunities, courses where students learn and practice research data skills may exist. Course-embedded librarians would be able to assist faculty in developing projects that effectively teach data management concepts, while also supporting students as they learn to put these skills into practice. The GIS course described by Ivey et al. (2012)

provides an example of how data management and other data practices can be integrated into the graduate curriculum.

Limitations and Future Work

This study focuses on understanding the data practices among CEE graduate students at two doctoral-granting research institutions. The nature of qualitative research includes small samples and allows for the richness of the data to be examined, but findings cannot always be generalized. Further studies employing different methods with larger sample sizes from multiple institutions could provide a broader, more generalized understanding of graduate student data practices. In addition, although this study focused on graduate students in CEE, we found interviewees conducting multidisciplinary work which likely incorporates views and knowledge from other disciplines. Future work can explore the data practices in other areas of engineering and other fields related to CEE to gain a broader picture of these multidisciplinary data practices. In addition, research can focus on data practices for specific data types, such as geospatial data.

Reproducibility came up incidentally across our interviews, indicating its importance in students' data practices. Designing studies that dig deeper into graduate student researchers' understanding and reproducibility practices could produce richer data about this topic. Similarly, future work can also explore graduate students' understanding of and values toward Open Science and elucidate how it influences their research data practices. Finally, norms around remote and hybrid work are changing, and the effects of these changes on graduate student research practices, especially around learning, collaboration, and serendipitous encounters, are another area for future work.

Conclusion

This study aims to understand CEE graduate students' data practices and needs, as well as provide recommendations to support those needs. Results from this study confirmed the finding from previous studies (Chen et al., 2019; Kuglitsch et al., 2018) that graduate student researchers are in large part responsible for managing research data. Findings indicate that graduate student researchers in CEE are frequently working with secondary data, and practice different aspects of good data management to varying extents. Willingness to disseminate data and code varies, as there is a need to balance personal priorities, such as time and level of control over data, with the desire for openness and reproducibility. Learning about data and data management most often occurs through peer networks, with some additional support from advisors and self-directed learning. CEE graduate students' needs center around (1) data quality—both a need to improve the quality of available secondary data and instruction on evaluating data quality, (2) tools and resources that help build good data management habits, (3) instruction on enacting reproducible data practices, and (4) information and learning opportunities that connect to students' established patterns for communication and learning. Librarians and other campus services and departments are well-positioned to meet these needs and improve graduate student expertise and research data practices.

Acknowledgments

Our thanks to Prof. Pingbo Tang from CMU and Prof. Andrew Johnson from CU Boulder for their comments on a draft of this paper. We also appreciate Ithaka S+R for connecting us via the original research study. Finally, our gratitude to the interviewed graduate students who shared their time, experiences, and perspectives with us.

References

Antell, K., Foote, J. B., Turner, J., & Shults, B. (2014). Dealing with data: Science librarians' participation in data management at Association of Research Libraries institutions. *College & Research Libraries, 75*(4), 557–574. https://doi.org/10.5860/crl.75.4.557

Beaver, D. Deb. (2001). Reflections on scientific collaboration (and its study): Past, present, and future. *Scientometrics, 52*(3), 365–377. https://doi.org/10.1023/A:1014254214337

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal, 14*, 1–10. https://doi.org/10.5334/dsj-2015-002

Carlson, J., Fosmire, M., Miller, C. C., & Nelson, M. S. (2011). Determining data information literacy needs: A study of students and research faculty. *portal: Libraries and the Academy, 11*(2), 629–657. https://muse.jhu.edu/article/428877/summary

Carlson, J., & Stowell-Bracke, M. (2013). Data management and sharing from the perspective of graduate students: An examination of the culture and practice at the water quality field station. *portal: Libraries and the Academy, 13*(4), 343–361. https://muse.jhu.edu/article/522641

Carnegie Mellon University. (2020). *Enrollment Fall 2021.* Institutional Research and Analysis. http://www.cmu.edu/ira/Enrollment/index.html

Chen, X., Benner, J., Young, S., & Marsteller, M. R. (2019, June 15-19). *Understanding the research practices and service needs of civil and environmental engineering researchers – A grounded theory approach* [Paper presentation]. 2019 ASEE Annual Conference & Exposition, Tampa, FL, United States. https://doi.org/10.18260/1-2--33483

Cho, J. Y., & Lee, E.-H. (2014). Reducing confusion about grounded theory and qualitative content analysis: Similarities and differences. *The Qualitative Report, 19*(32), 1–20. https://doi.org/10.46743/2160-3715/2014.1028

Choudhury, S. (2008). Case study in data curation at Johns Hopkins University. *Library Trends, 57*(2), 211–220. https://muse.jhu.edu/article/262031

Cooper, D., Springer, R., Benner, J., Bloom, D., Carrillo, E., Carroll, A., Chang, B., Chen, X., Daix, E., Dommermuth, E., Figueriredo, R., Haas, J., Hafner, C., Henshilwood, A., Krogman, A., Kuglitsch, R., Lanteri, S., Lewis, A., Li, L., ... Yu, S. H.

(2019). *Supporting the changing research practices of civil and environmental engineering scholars.* Ithaka S+R. https://doi.org/10.18665/sr.310885

Cox, A. M., Kennan, M. A., Lyon, L., & Pinfield, S. (2017). Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *Journal of the Association for Information Science and Technology, 68*(9), 2182–2200. https://doi.org/10.1002/asi.23781

Cox, A. M., & Pinfield, S. (2014). Research data management and libraries: Current activities and future priorities. *Journal of Librarianship and Information Science, 46*(4), 299–316. https://doi.org/10.1177/0961000613492542

FORCE11. (2014). *The FAIR data principles.* FORCE11: The Future of Research Communications and e-Scholarship. https://force11.org/info/the-fair-data-principles/

Gardner, S. K. (2007). "I heard it through the grapevine": Doctoral student socialization in chemistry and history. *Higher Education, 54*(5), 723–740. https://doi.org/10.1007/s10734-006-9020-x

Ivey, S. S., Best, R. M., Camp, C. V., & Palazolo, P. J. (2012, June 10-13). *Transforming a civil engineering curriculum through GIS integration* [Paper presentation]. 2012 ASEE Annual Conference & Exposition, San Antonio, TX. United States. https://doi.org/10.18260/1-2--22130

Jahnke, L., Asher, A. D., Keralis, S. D. C., & Henry, C. (2012). *The problem of data.* Council on Library and Information Resources and Digital Library Federation. http://www.clir.org/pubs/reports/pub154

Johnston, L., & Jeffryes, J. (2014). Data management skills needed by structural engineering students: Case study at the University of Minnesota. *Journal of Professional Issues in Engineering Education and Practice, 140*(2), 1–8. https://doi.org/10.1061/(ASCE)EI.1943-5541.0000154

Kim, J. (2013). Data sharing and its implications for academic libraries. *New Library World, 114*(11/12), 494–506. https://doi.org/10.1108/NLW-06-2013-0051

Kuglitsch, R., Dommermuth, E., & Lewis, A. (2018). *Research practices of civil and environmental engineering scholars.* University Libraries University of Colorado Boulder. https://scholar.colorado.edu/libr_facpapers/132

Lage, K., Losoff, B., & Maness, J. (2011). Receptivity to library involvement in scientific data curation: A case study at the University of Colorado Boulder. *portal: Libraries and the Academy 11*(4), 915-937. https://muse.jhu.edu/article/452638

Latham, B. (2017). Research data management: Defining roles, prioritizing services, and enumerating challenges. *The Journal of Academic Librarianship, 43*(3), 263–265. https://doi.org/10.1016/j.acalib.2017.04.004

Montgomery, J. L., Harmon, T., Kaiser, W., Sanderson, A., Hass, C. N., Hooper, R., Minsker, B., Schnoor, J., Clesceri, N., Graham, W., & Brezonik, P. (2007). The WATERS Network: An integrated environmental observatory network for water research. *Environmental Science & Technology, 41*(19), 6642–6647. https://doi.org/10.1021/es072618f

Newton, M. P., Miller, C. C., & Bracke, M. S. (2010). Librarian roles in institutional repository data set collecting: Outcomes of a research library task force. *Collection Management, 36*(1), 53–67. https://doi.org/10/d65c3b

Pasek, J. E., & Mayer, J. (2019). Education needs in research data management for science-based disciplines: Self-assessment surveys of graduate students and faculty at two public universities. *Issues in Science and Technology Librarianship, 92*. https://doi.org/10.29173/istl12

Pejşa, S., & Song, C. (2013, July 22-26). *Publishing earthquake engineering research data* [Paper presentation]. Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, Indianapolis, IN, United States. https://doi.org/10.1145/2467696.2467758

Petters, J. L., Brooks, G. C., Smith, J. A., & Haas, C. A. (2019). The Impact of targeted data management training for field research projects – A case study. *Data Science Journal, 18*(43), 1–7. https://doi.org/10.5334/dsj-2019-043

Pinfield, S., Cox, A. M., & Smith, J. (2014). Research data management and libraries: Relationships, activities, drivers and influences. *PLoS One, 9*(12), 1–28. https://doi.org/10.1371/journal.pone.0114734

Radecki, J., & Springer, R. (2020). *Research data services in US higher education: A web-based inventory.* Ithaka S+R. https://doi.org/10.18665/sr.314397

Rolando, L., Carlson, J., Hswe, P., Parham, S. W., Westra, B., & Whitmire, A. L. (2015). Data management plans as a research tool. *Bulletin of the Association for Information Science and Technology, 41*(5), 43–45. https://doi.org/10.1002/bult.2015.1720410510

Sadiq, S., & Indulska, M. (2017). Open data: Quality over quantity. *International Journal of Information Management, 37*(3), 150–154. https://doi.org/10.1016/j.ijinfomgt.2017.01.003

Sapp Nelson, M. (2015). *Data strategies: The research says.* Purdue University. https://doi.org/10.5703/1288284315525

Satheesan, S. P., Alameda, J., Bradley, S., Dietze, M., Galewsky, B., Jansen, G., Kooper, R., Kumar, P., Lee, J., Marciano, R., Marini, L., Minsker, B. S., Navarro, C. M., Schmidt, A., Slavenas, M., Sullivan, W. C., Zhang, B., Zhao, Y., Zharnitsky, I., & McHenry, K. (2018, July 22-26). *Brown Dog: Making the digital world a better place, a few files at a time* [Paper presentation]. Proceedings of the Practice and Experience on Advanced Research Computing, Pittsburgh, PA, United States. https://doi.org/10.1145/3219104.3219132

Schröder, W., & Nickel, S. (2020). Research data management as an integral part of the research process of empirical disciplines using landscape ecology as an example. *Data Science Journal, 19*(26), 1–14. https://doi.org/10.5334/dsj-2020-026

Shahi, A., Haas, C. T., West, J. S., & Akinci, B. (2014). Workflow-based construction research data management and dissemination. *Journal of Computing in Civil Engineering, 28*(2), 244–252. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000251

Sharma, S., & Qin, J. (2014). Data management: Graduate student's awareness of practices and policies. *Proceedings of the American Society for Information Science and Technology, 51*(1), 1–3. https://doi.org/10.1002/meet.2014.14505101130

Tang, R., & Hu, Z. (2019). Providing research data management (RDM) services in libraries: Preparedness, roles, challenges, and training for RDM practice. *Data and Information Management, 3*(2), 84–101. https://doi.org/10.2478/dim-2019-0009

Trisovic, A., Mika, K., Boyd, C., Feger, S., & Crosas, M. (2021). Repository approaches to improving the quality of shared data and code. *Data, 6*(2), 15. https://doi.org/10.3390/data6020015

UNESCO. (2020). *Open science: Making science more accessible, inclusive and equitable for the benefit of all.* https://en.unesco.org/science-sustainable-future/open-science

University of Colorado Boulder. (2020). *University of Colorado Boulder IR.* Tableau Public. https://public.tableau.com/app/profile/university.of.colorado.boulder.ir

U.S. News & World Report. (2021). *Best engineering schools ranked in 2021.* https://www.usnews.com/best-graduate-schools/top-engineering-schools/eng-rankings

Valentino, M., & Boock, M. (2015). Data management for graduate students: A case study at Oregon State University. *Practical Academic Librarianship, 5*(2), 77–91.

Wiley, C. A., & Kerby, E. E. (2018). Managing research data: Graduate student and postdoctoral researcher perspectives. *Issues in Science and Technology Librarianship, 89*. https://doi.org/10.5062/F4FN14FJ

Witt, M. (2012). Co-designing, co-developing, and co-implementing an institutional data repository service. *Journal of Library Administration, 52*(2), 172–188. https://doi.org/10.1080/01930826.2012.655607

Appendix 1: Interview Protocol

Research Support Services for the Field of Civil and Environmental Engineering

Instructions: In answering our questions, please do not provide any identifiable or private information about another person, for example do not use another person's name, or reveal anything private about another person through which their identity could be inferred.

Background

- Describe your research focus and projects.
- Do you do research as part of a group or lab, or do you mainly work one-on-one with your advisor? Do you have more than one advisor?
- What does communication with your advisor(s) look like? Do you have group meetings, individual meetings?
- How do you communicate within your research group? Do you have regular meetings?
- What stage of your program are you in?

Data collection

- Can you describe the kind of data that you primarily work with? (Tabular/spreadsheets, text files, encoded data, models etc.)
- Do you typically produce data?
    - What kinds of data does your lab's research typically produce?
        - Did you personally produce this data or are you using "inherited data"?
        - How long have you been working with these data set(s)?
    - How do you collect your data?
        - What tools do you use to collect and record data? / where does the data live (spreadsheet, ELN, Instruments or models)?
- Does your research involve working with data produced by others?
    - What kinds of data produced by others does your lab typically work with?
    - How is that data procured?

Active Data

- What's a typical size of the data you work with?
- How do you manage and store data for your current/active use?
    - How do you backup your data?
    - If using other people's data do you keep a copy? How do you manage (store, etc.) that copy of the data?
    - Version control
    - (what cloud?)

- How do you analyze the data? (tools, methods, campus resources/collaborations)
    - Parallel processing?
    - Databases?
- How is the data incorporated into the research outputs?

Data preservation

- What are your plans for managing the data and associated information beyond your current use?
    - Does your lab have a policy for long term data storage and preservation/archiving?
    - Do you collect, maintain, or document your data or analysis processes? (Store metadata anywhere?)
- How do you plan to preserve your data for future use? (probe: would you preserve it in an institutional repository, a disciplinary repository, FigShare, etc., a dark archive, or a cloud storage system like AWS)
    - Specifically for others, i.e. reuse, public
    - For future you
- What's your plan for the data you produced after you graduate?

Data sharing, publication, reuse

- Who do you collaborate with?
    - What do you do to prepare the data to be used by someone other than yourself? [Probe: who are these others--colleagues, following grad students, external scientists]
    - How do you share your data with collaborators? (Advisor, PIs, other grad students, other researchers)
- Do you publish your data (or models)? Do you want your data to be easily discoverable?
    - What parts of your data do you publish? (Raw data? Research data? Scripts or code? Models? Documentation?)
    - Where do you publish your data?
    - If you don't publish your data, why not?
- Is your research supported by any grant? Does the funding agency request you to share the data? (data ownership)

Challenges in working with data

- Have there been any challenges in the process of working with the data your lab's research produces? Or from others?
- Are there any resources, services, or other supports that would help you or the lab as a whole more effectively work with the data produced by others?
- How did you learn about data management practices that you apply in your research? Have you received any help in terms of data management practice? What are they?

Wrap up

- What are you interested in doing after you finish your studies here? Do you think the data skills that you've learned here will help you in your future career?
- Do you have anything else you'd like to add?

Appendix 2: Codes, Descriptors and Definitions

Table 1. Theme codes used in our analysis with definitions and exclusions

Theme Code Name Subcodes	**Definition**	**Exclusion**
Working with others External collaboration Working with advisor Working with non-academic entities Working with peers/colleagues Working with support staff	Anytime the student describes interacting with another person, group, or lab. Interacting can include getting help, communicating, sharing results or data internally. Share research outputs within research group, with advisors, with collaborators.	
Getting data Evaluating data quality Leveraging collaboration Leveraging personal connection Searching the literature Searching the web	Anytime the student describes about how they obtain, collect, produce, find, discover, or access data to do their research.	Any mention of how they store or manage the data.
Processing/Analyzing data Applying statistical analysis Building models Processing data Using models Visualizing data	Anytime the student describes how they clean data to prepare for data analysis, use and manipulate data, write scripts, develop or use models, conduct exploratory data analysis.	Any mention of how they store or manage the data.
Managing data for current use Backing up data Documenting for current use Organizing files Practicing version control Sharing data for collaboration Storing data Working with a data management plan	Anytime the student describes how they store/ backup their data, label or add metadata, create codebooks, manage data workflows and version control or document data processing for their current use. This includes mentions of lab policies/practices regarding managing data for current use.	
Sharing research outputs Citing data Documenting for external sharing Licensing issues Navigating mandates to share Publishing code/models Publishing data Publishing results	Anytime the student discusses making their data, code, scripts, models or visualizations available in a repository, publication, or to external individuals or funders upon request. This includes mentions of funding requirements or lab	Sharing within research group, with advisor, with collaborators.

Theme Code Name Subcodes	**Definition**	**Exclusion**
	policies/practices regarding data sharing.	
Long-term Storage / Preservation	Anytime the student discusses their plans (or lack thereof) for storing and managing their data for longer than the duration of their project or plans to document their data for future use. This includes mentions of lab / funder policies/practices regarding managing data for future use.	Discussing plans for current use.
Reproducibility	Anytime the student discusses a desire (or lack thereof) for others to replicate or reproduce their work OR discusses reproducibility or lack thereof in others work.	

Table 2. Cross-cutting codes used in our analysis with definitions and exclusions

Cross-cutting[a]	**Inclusion**[b]	**Exclusion**
Challenges	Anytime the student describes something that was a 'challenge', 'difficult', or a 'barrier' to their work.	
Opportunities	Anytime the student specifically asks for something or *describes a gap in service* or a situation in which the institution can be of assistance.	
Existing Support Services	Anytime the student discusses supports they are already receiving/accessing from the library or university.	Mention of a tool provided by a service, unless support/help with tool is mentioned.
Tools	Anytime the student mentions specific off-the-shelf software, platforms (e.g., Google Drive), programming languages, scripting tools or hardware.	Research methodology, tools that the student is developing.
Values	Anytime the student mentions their values toward Open Access, Open Science, RDM, or Reproducibility OR their *perception* of RDM as easy or hard or just not applicable to their work, OR how they define RDM.	
Learning	Anytime the student mentions getting training or building skills in a formal setting, mentoring from others or teaching themselves about a topic.	

Cross-cutting[a]	Inclusion[b]	Exclusion
[a] Add as many of these as you want. [b] Always requires co-coding.		

Table 3. Descriptors used in our analysis with definitions and accepted values

Characteristic Name	Definition	Values
Year in program	The current year of study for the student.	Early (1-2 years; no results yet); middle (3-4 years; some results; maybe a publication); late (4+ years; results and maybe a publication)
Career path	The student's stated career interest.	Industry, academia, government, NGO, doesn't know, not mentioned.
Size of data	How the student describes the size of their data.	Big, small, both
Type of data	Are they producing data (primary) or collecting it from others (secondary)?	Primary, secondary, both
If secondary, pay or free?	If using secondary data is it free available or do they have to pay for it?	Pay, free
If secondary, organization?	If using secondary data, what kind of organization is it from?	Industry, academia, government, NGO
If primary, how is it collected?	What methods are used to create the primary data?	Experimental, observation
Research methods	What type of research method is the student using?	Open text, common values included: machine learning, statistical analysis, experimental, observational, modeling
Number of advisors	How many advisors are working with the student?	One, two, more

Appendix 3: Descriptors Assigned to Each Interview

Interview Number	Data Size	Data Type	If Secondary, Cost	If Secondary, Source	Research Methods Used
CU1	Small	Primary & Secondary	Free	Academia	Modeling & Experiment
CU2	Big	Primary			Modeling & Observation
CU3	Big	Primary & Secondary	Unclear	NGO	Modeling, Observation & Survey/Interview

Interview Number	Data Size	Data Type	If Secondary, Cost	If Secondary, Source	Research Methods Used
CU4	Small	Primary & Secondary	Unclear	Government	Modeling, Experiment & Observation
CU5	Small	Primary			Observation & Survey/Interview
CU6	Big & Small	Secondary	Unclear	Government	Modeling & Statistical Analysis
CU7	Big & Small	Primary & Secondary	Free	Government	Survey/Interview & Social Media Analysis
CU8	Big & Small	Secondary	Free	Government	Modeling
CMU1	Big	Secondary	Free	Academia & Government	Modeling
CMU2	Big & Small	Primary & Secondary	Unclear	Academia, Government & Industry	Modeling, Observation & Social Media Analysis
CMU3	Big & Small	Secondary	Free	Academia & Government	Modeling, Statistical Analysis & Case Study
CMU4	Small	Primary & Secondary	Free	Government, Industry & NGO	Modeling, Statistical Analysis, Experiment & Survey/Interview
CMU5	Small	Secondary	Free	Government, Industry & NGO	Modeling & Statistical Analysis
CMU6	Small	Primary & Secondary	Free	Government & Industry	Modeling, Statistical Analysis & Observation
CMU7	Small	Secondary	Free	Academia & Government	Modeling & Statistical Analysis
CMU8	Big	Primary			Modeling & Experiment
CMU9	Small	Secondary	Free	Government & NGO	Modeling
CMU10	Small	Secondary	Free	Academia	Modeling
CMU11	Small	Secondary	Pay	Industry	Modeling