Data Management Skills Needed by Structural Engineering Students: A case study at the University of Minnesota

Johnston, Lisa.¹ and Jeffryes, Jon.²

Keywords: Data Communication, Data Collection, Management (NT Information Management), Engineering Education

Abstract

The current era of e-science and big data calls upon researchers to ensure the security and management of their research data. In this case study researchers from the University of Minnesota Libraries interviewed a graduate student research team and faculty advisor to provide an understanding of the educational gaps that need to be filled in order to graduate data-savvy engineers. This paper presents the needed data management skills with thoughts on how the engineering discipline and academic research libraries might partner on a shared training solution.

¹ Lisa Johnston (ljohnsto@umn.edu) is the Research Services Librarian at the University of Minnesota, Twin Cities. 108 Walter Library, 117 Pleasant Street SE, Minneapolis, MN 55455.

² Jon Jeffryes (jeffryes@umn.edu) is an Engineering Librarian at the University of Minnesota, Twin Cities. 108 Walter Library, 117 Pleasant Street SE, Minneapolis, MN 55455.

Introduction

The current era of e-science and big data challenges engineers to face not only research problems, but also cope with and manage digital results that can be replicated and preserved for future use. In structural engineering, projects like the Network for Earthquake Engineering Simulation (NEES), have begun to shape the digital curation skills required by researchers in the field. In 2010 The National Science Foundation (NSF) included a requirement for writing a data management plan into grant proposals. This addition requires researchers to consider, plan, and implement methodologies for creating, collecting, organizing, processing, representing, archiving, preserving and ethically sharing research data (NSF 2010). Large research universities also have, often lesser known, expectations for data retention that require researchers to maintain their data in secure and accessible formats (PEL Program 2009). For the purposes of academic educational objectives, it becomes important to determine what data management (DM) skills could be identified as necessary to graduate “data literate” engineers who are able to properly manage research data in the digital age.

The task of managing research data for scholarly use and reuse has evolved quickly over the past decade; becoming increasingly important to both the disciplinary communities that rely on large data sets and networked scientific collaborations (e.g., E-Science, interdisciplinary informatics), but also to the information professions that collect, archive, disseminate, and preserve scholarship. Therefore DM best practices, bringing a potential for data reuse and re-analysis to solve new research problems, highlight a role that library expertise will fill in this research environment. A report by the American Research Libraries describes “data stewardship” as a critical facet for researchers in this era of E-science and the libraries that support them (Lougee, 2007). Data stewardship encompasses both preservation and curation of well-managed digital data for the purposes of long-term data archiving, dissemination and reuse. The report signified a call-to-action for research libraries to undertake assessment methods

surveying the nature of DM and research behavior within their local domain in order to develop the tools and education needed to help researchers in this pursuit.

The Data Information Literacy Project (<http://www.datainformationliteracy.org>), funded by a grant from the Institute of Museum and Library Services (IMLS), will result in models of scientific and engineering data literacy drawing on the expertise from the library science and information fields. The models -- including information organization and representation, dissemination and preservation -- will better serve the scientific and engineering disciplines. This project includes five disciplinary-based case studies, underway at four different institutions, each concurrently look at the DM skills needed by graduate students in science and engineering disciplines. As co-PIs on this project, the authors conducted in-depth interviews with a faculty member and four graduate students in a structural engineering research group housed in the Civil Engineering department at the University of Minnesota (UMN) to identify the DM skills needed to fully prepare students for their post-graduate research.

Here we summarize the case study findings and present the observed needs and DM skills for graduate students in structural engineering and outline potential learning objectives to meet these needs. As a follow up to this case study, our team will respond by developing and implementing a DM curriculum for our partnering research group to address these local needs (planned implementation for Fall 2012). A follow-up paper will report our educational intervention, subsequent student response, and assessment results of our approach.

Literature Review and Environmental Scan

The literature shows the importance of data management (see specifically the literature reviews of Haug, et al. 2011 and Marsh 2005). Reporting from industry Marsh reports "If the accuracy and security

of a company's data are in doubt, stakeholders will lose confidence, trading will be affected and performance will suffer." (2005, p. 107). Specifically in the discipline of civil engineering the call for thorough data management has been around for twenty years (Port 1992). Port wrote, "Therefore, the way in which engineers handle their data is of central importance to their efficiency and effectiveness – this makes data the 'lifeblood' of engineering activity." (1992, p. 133).

Currently the civil engineering disciplinary expectations regarding teaching data management to its students remains poorly defined. In the sciences more generally the primary importance of skills surrounding data has been associated with introductory research methods courses and through related topics such as research ethics or laboratory sections. The topic of data literacy must be inferred into existing learning outcomes or other standards that touch upon data tangentially, usually under outcomes that focus on the overall experimentation process.

The most recent version of the American Society of Civil Engineers (ASCE)'s reconstruction of a successful engineering curriculum, *Civil Engineering Body of Knowledge for the 21st Century: Preparing the Civil Engineer for the Future* (BOK 2) does not explicitly address the concept of data literacy. But data management skills can be inferred in many of the outcomes focused around its seventh outcome group, "Experiments." The relevant outcomes include:

Identify the procedures...to conduct civil engineering experiments
Explain the purpose, procedures...of experiments
Conduct experiments...according to established procedures
Analyze the results of experiments... (ASCE 2008, p.106)

A report produced between iterations of the BOK 2, *Development of Civil Engineering Curricula Supporting the Body of Knowledge For Professional Practice* found room for improvement in the depth of students' engagement with data, citing one example where "students are not able to take an open-

ended real world situation and design the experiments that would provide the necessary data to solve the problem.” (ASCE Curriculum Committee 2006, p.23)

Data literacy can also be inferred into the outcomes regarding communication (BOK 2, Outcome 16) which calls for students to “...**use** appropriate graphical standards in preparing engineering drawings” [emphasis included in original work] and “**[o]rganize** and deliver effective...graphical communications” (ASCE 2008, p.110). The skills can be implicitly read into the Outcome 13: Project Management, if the new standard procedure for conducting experiments includes creating a plan to manage data, including organization, security and preservation, now mandated by some funding agencies (NSF 2010).

The engineering field, more widely, shares this opacity of expectation. The outcomes suggested in the BOK2 echo those already implemented by the Accreditation Board for Engineering and Technology (ABET) in their outcome, “an ability to design and conduct experiments, as well as to analyze and interpret the data” (ABET 2012).

Locally, UMN students and faculty receive somewhat scattershot DM training. For example, the University requires all Primary Investigators (PIs) of research grants to complete one of two web-based instructional modules on “Research Data Management” (FIRST 2011) that cover some aspects of data control and intellectual property concerns. However these Responsible Conduct of Research (RCR) modules, last updated in 2008, are extremely weak on the issues inherent in organizing, sharing, and preserving information in a digital environment. Beginning in 2010, researchers could supplement that training with workshops taught by the libraries on “Creating a Data Management Plan for Your Grant Application” or “Introduction to Data Management for Scientists and Engineers” available as drop-in library workshops and online video recordings (University Libraries 2012). The former workshop reached

over 300 faculty members and is offered for RCR continuing education credit (Johnston, Lafferty, and Petsan 2012). However, both RCR training and library-led workshops are designed specifically for faculty PIs and therefore not targeting the graduate student population.

It is possible that data management skills are being addressed, along with other information literacy skills, in student research experiences such as UROPs, research assistantships, or cooperative educational programs, but the current literature focuses primarily on information retrieval skills (Jeffryes and Lafferty 2012). One student in our study mentioned receiving some data management skills in an introductory research methods class, but considered it too early in their student career to be useful to their current research. The current integration of DM skills into the graduate curriculum are neither constant nor at their point of need. To address this gap in DM training for graduate students, Carlson et al (2012) expressed the need for a data literacy training program geared toward graduate students in their academic discipline. This research formed the Data Information Literacy Project, and with it, this case study.

Methodology

In the spring of 2012, the authors participated in a grant funded project to study the needs of graduate students in data management skills in five disciplinary case studies run at four research universities. The UMN case study interviewed members of a structural engineering research group consisting of one faculty member and four graduate students ranging in experience from a first year graduate student to a student in their final semester. The interview instrument used by the project is a modified version of the Data Curation Profiles Toolkit (Carlson n.d.) a tool used to gather detailed information about the practices, limitations, needs and opportunities for improving DM practices from the perspective of both the faculty member and graduate students in the subject area. The instrument and interviewee

handbook are available for reuse (DIL n.d.). In addition, relevant documentation, including data set examples and supporting research practices, were collected and evaluated.

The interviews were conducted between March 13th, 2012 and April 20th, 2012. Structured one- to two-hour interviews took place in a library conference room using two handheld Samson H4 Zoom recorders each producing a .WAV audio file that a graduate assistant transcribed into Microsoft Word documents for analysis. In addition to the verbal questions of the interview, the interviewees each filled out a paper-based worksheet with corresponding questions to the interview questions. The observation data, including the data examples from the research group, the interview transcripts and audio files, and the interview worksheets, were analyzed, anonymized and compiled into an excel file for sharing among the five Data Information Literacy Project members.

Results of the Needs Assessment

Our case study generated four in-depth graduate student interviews, one faculty interview, and associated quantitative worksheet data providing a snapshot of the DM skills needed for structural engineering graduate students at the UMN. The analysis revealed several needs at various stages throughout the data life-cycle. It was clear that the students had no formal training in DM, something that could either hinder them in future research positions or in the least, create inconsistency with the generation and maintenance of structural research data during their student career at the university.

In the course of interviewing the graduate students, we noticed that certain steps in the data lifecycle were present regardless of the specific details of the research project (see Figure 1).

(Insert Figure 1: Data lifecycle synthesized from interviews)

The students did not use a consistent vocabulary when describing these steps, but the same characteristics of activity appeared in each of the student interviews (Table 1).

(Insert Table 1: Data Life-Cycle Stages as described by the graduate students)

To analyze the skills and needs described in the interviews, the results were viewed in the context of each stage of the data life-cycle. A final stage (Stage 6: "Preservation") is included in our analysis below. Although the students did not identify preservation as a step in their data life-cycle, critical aspects of this topic were mentioned throughout the data interview. These observations provide a foundation for a generalized approach to understanding the data interactions of structural engineering graduate students in a research group.

Raw Data

In the first module of the interview we asked the graduate students to describe the type of data with which they were working. All graduate students reported using sensor data as the crux of their research projects. Three out of four of the graduate students collect data for projects that generate real-time sensor data to monitor the performance of local bridges, while one graduate student generated experimental data and simulations on concrete column performance in simulated earthquake conditions.

The faculty interview highlighted the need for students to understand the potential hazards of collecting "bad" data and the faculty member thought a better understanding of how sensors collect data might help. Several students mentioned knowing about potentially disruptive elements such as temperature

conditions or scheduled construction/testing; however their processes and documentation did not merge these events with the data that was collected. Additional points of interest from the raw data stage:

- two of the students inherited data from a previous graduate student.
- data collected from bridge monitoring projects can generate up to 400 individual files that can be as large as 10MB depending on the frequency of data collection.
- file formats collected by the students varied between ASCII text files to proprietary data files only readable with specialized, proprietary software.

Collection and Organization

In discussions regarding data collection and organization more trends emerged, appearing throughout each of our graduate student interviews. For example,

- students used date-based file naming structures, even when they weren't familiar with the concept of a file naming structure, as one student commented: "I've never even heard of a file naming system."
- students did not consider data security an issue and felt that they had adequate protections in place.
- back-up of their data was often sporadic or non-existent. Two of the students interviewed displayed some confusion about the concept of data backup versus data redundancy. For example, one student described their backup process as copying files to a separate folder on their desktop (which would not protect against theft or computer damage).
- students agreed that they had no formal DM instruction but had to rely on their peers, family, and previous experience for direction. One student described "I've had many projects with Excel files and stuff that I've needed to save, and I guess [I learned DM] just out of habit, mainly."

(Insert Table 2: Interview Results Summary of Data Management Practices by Structural Engineering Graduate Students)

Formal and informal documentation practices recording the data collection process and changes made to the data were ad hoc and varied (see Table 3). Only one student was familiar with the term metadata, responding “It means data captured and saved during the test.” The other students each responded negatively when asked if they were familiar with the term. Regardless, all students were providing some level of metadata to the data they were working with, but the majority were not collecting or applying it in an intentional or formal manner.

When asked if the student had any means of documenting the DM steps for someone else to repeat, the inefficiencies of their own system were described. One student admitted, “I guess if I were to repeat [the research project], I would probably do it in a different way. I could probably document what I’ve done and I probably will do so, but then I’ll also suggest maybe keeping things a little less complicated.”

(Insert Table 3: Documentation and Metadata Practices Reported by Structural Engineering Students)

Processing/Analysis and Results

Each of the graduate students described a process for analyzing, visualizing and making conversions of the data beyond the original raw data stage. The majority of the graduate students spoke of a process of converting ASCII text files into Excel for further manipulation and sense-making. One graduate student reported using a proprietary sensor program that allowed for data manipulation within their web-based software. Regardless of format, a process of further manipulation of the data was described, such as a

process of removing “bad” data (i.e., bridge sensor readings contaminated due to construction), synthesizing the rough data using equations, and creating graphical representations of the data (“plotting”), all to better communicate their findings.

The faculty member held the graduate students’ facility with Excel and Matlab in high esteem, but had some concern that students weren’t receiving all the support they needed in more advanced data analysis, saying: “It’s the relational databases...and their capabilities for statistical analysis that are a little weak. And there are courses they can take on campus for the statistical and the relational databases, so maybe it’s something that we should be requiring. The problem is that if they’re going to do a Master’s thesis, they take only seven courses.”

He echoed the sentiment for further development of student skills in this area by noting that further education on the strategy behind the data plotting would be beneficial. His ideal would be for graduate students to demonstrate an “ability to take the data and come up with a way of conveying it so that the reader can pick it up very quickly.” Indeed one student described their process of creating data visualizations in Excel as “mostly trial and error.”

The faculty member also specifically called out the need for students to be able to identify and track the quality of the data they are collecting, which may have been compromised by outside forces, such as construction on the bridge from which they are collecting sensor data. The professor commented that the students aren’t currently tracking this aspect of their data analysis in the documentation, but “it would be nice, especially when they’re collecting huge amounts of data, if we could somehow get measures of the quality of the data, statistically. And if we could use these measures to keep track of getting good data and when we’re not getting good data.”

Sharing and Archiving

Each of the four students share their data in some way. One student shared their data in a formal process through the mandatory data archiving protocol of the NEESHub program, while the other students shared their data with state contractors, their advisor, and passed along findings to the next graduate student continuing the project.

Although students had little to no experience with data citation, when asked if they think it important, they reported an understanding of the value of this practice. A student explained, "because you need to know where this data is coming from, and obviously if it's not your own, then I feel like it's important to make other people aware that it is not data that you actually collected yourself."

As to the potential for other researchers to reuse their data, only one student felt their analyzed data was unique and therefore of potential value. The other students had a harder time imagining how their data might be useful to researchers outside of their specific project. Additionally, the graduate students demonstrated little to no knowledge of data repositories in their field or experience using someone else's data from outside their lab. One student did mention looking at another researcher's data in the literature review leading up to their experiment, but the data was found by chance and the repository was not a standard destination.

The graduate students also did not see the value in archiving similar data sets together in a subject-based repository structure. Referencing the 35W bridge in Minneapolis which was rebuilt after the tragic 2007 collapse with sensors measuring strain in a similar way to the data obtained by our interviewee, the student noted "unless you could come up with some good way to compare the two sets of data, I

don't know really what use it would be to collect them all into one place." The student did see the value of data repositories to save on space however, so that "there aren't 50 external hard drives floating around."

Issues around privacy and confidentiality were a complex issue for students working in state-contracted work analyzing bridge sensor data. Students knew to contact their advisor with requests to share the data owned by the state agency. One student described their caution with presenting the state funded data results at a conference, "I had to get permission from [the state contractor] first before I could even do that." However the reasons beyond "ownership" were unclear. The faculty member was able to explain the sensitive nature of the data when asked if the state agency had any specific interests in sharing this data beyond the agency. The professor replied: "That's a really good question. They would like to share data, as long as they can protect their interests. And I don't mean any advantage in having that data. What they're afraid of is this data represents measurements that are taken off of real bridges, and that can very easily be misinterpreted and used to undermine a bridge that's actually not in bad shape, and then present a bloated and incorrect scenario about how bad the bridge problem is. Or the claim that a bridge is in great condition, when in fact it needs to be replaced. For that reason, they are very, very, very unwilling to have anything like open access."

(Insert Table 4: Reusing and Sharing Data Responses by Structural Engineering Graduate Students)

Preservation

In discussing the long-term access to data, several preservation concerns arose. The subjects identified multiple issues not currently being addressed in the research group such as physical storage (desktop

computer used by each graduate student will eventually be recycled) and file migration (saved in proprietary and future incompatible version of Excel) for data stored in the lab.

Students were unclear who held the responsibility to preserve the data for long term access. Additionally they were uncertain of the steps needed to be taken to preserve data for 20-50 years, or the life of the bridge. For example, one student suggested the contracting state agency held the responsibility for preserving the data and that they would keep the data “forever.” When asked to identify the steps needed to preserve the data and if the state currently implemented those steps, the student responded: “I think that's just sort of what they do. Again, because they've had issues in the past where people have completed projects and then others have wanted to repeat them or go more into depth with them and then haven't been able to find any of the original data for it, so I think that's kind of just their policy.” When asked for steps to preserve the data set the graduate student reported, “Just putting them onto that hard drive and making sure it doesn't melt I guess.”

(Insert Table 5: Preservation Responses by Structural Engineering Graduate Students)

In our conversation with the faculty member the issue of data versioning for long-term access and preservation arose. Along with identifying and implementing steps to preserve and store data in the long-term, researchers must also choose which versions of their data should be preserved for future use and authenticity. The professor responded to the issue of versions: “This is an interesting problem. There are actually multiple stages and multiple things that you do [to the data], and so how many data sets do you store? Clearly, you want the raw data. That's the purest form. And clearly you want the data that you think has been completely digested as you think it needs to be. But how many of the intermediate stages do you want to keep?”

Summary Data Management Skills Needed by Graduate Students in Structural Engineering

Based on our analysis the following DM skills were uncovered as important educational needs for graduate students participating in structural engineering advanced degree programs.

(Insert Table 6: Summary of Needed Data Management Skills)

Conclusions

As mentioned above, the Data Information Literacy Project will result in five instructional approaches that pair faculty and librarians to educate graduate students in DM skills, with the ultimate outcome of producing data literate researchers. The results of this local study will be used to develop and implement an educational model based on a set of learning outcomes targeting the perceived greatest needs of graduate students that arose in the interviews. In this mutually beneficial partnership, the librarians bring their expertise in organizing and managing information and to apply to the curation of research data. At the same time, the civil engineering faculty member can bring examples and target the skills that will speak to the students' experiences and fit within the disciplinary norms.

Currently the integration of these skills into the graduate level curriculum remain completely voluntary. Students graduating have no guarantee of receiving formal education in the best practices of data management. Many students only learn through informal word-of-mouth instruction or address the problem when they suffer their own data loss. As the expectation for data accountability grows, the discipline of civil engineering needs to formally address this gap in the core competencies outlined in the BOK. As shown above in the literature review in the current edition of the BOK, data management skills can only be inferred (and then inferred with creativity) in Outcome 13 and Outcome 16. The authors

recommend using this case study as a starting point in the conversation of disciplinary norms. A replication or adaptation of this process administered more widely would gauge the DM needs of students across institutions in the civil engineering field. Once the educational gaps have been identified, the BOK should be updated to explicitly address these skills, either incorporating them into one of the existing outcomes or including an additional outcome devoted to this emerging facet of responsible, high quality research.

This study provides a snapshot of the DM skills needed by graduate-level civil engineering students that not currently being met in a formal manner by the education outcomes of either the engineering discipline generally, or civil engineering specifically. This paper provides a starting point in the conversation of formalizing the skills necessary to produce data literate researchers across the discipline.

Twenty years ago Port wrote, "In the computer age, the scope for such confusion has been increased, as there might be plots and prints made, data files on hard-disc, back-ups on floppy discs, archived data on magnetic tape and data files transferred around local-area or wide-area networks." (1992, 134).

In 2012, our faculty partner reported, "The skills that they need are many, and they don't necessarily have it and they don't necessarily acquire it in the time of the project, especially if they're a Master's student, because they're here for such a short period of time. "

It is time for the civil engineering community to stop talking about the gaps in this skill set of data management and start making progress towards formally addressing them in the educational outcomes of the next generation of engineers.

References

Accreditation Board for Engineering and Technology (ABET). (2012) "Criteria for Accrediting Engineering Programs, 2012-2013." ABET. <<http://www.abet.org/engineering-criteria-2012-2013/>>. (May 23, 2012).

American Society of Civil Engineers(ASCE). (2008) *Civil Engineering Body of Knowledge for the 21st Century:Preparing the Civil Engineer for the Future*. 2nd Edition. Reston, VA: American Society of Civil Engineers.

ASCE Curriculum Committee. (2006) *Development of Civil Engineering Curricula Supporting the Body of Knowledge for Professional Practice*. <<http://www.asce.org/Content.aspx?id=2147484541>>. (June 29, 2012).

Carlson, Jacob. (n.d.) *Data Curation Profiles Toolkit*. <<http://datacurationprofiles.org>>. (May 23, 2012).

Data Information Literacy. "Interview instruments." (n.d.) *Data Information Literacy*. <<http://wiki.lib.purdue.edu/display/ste/materials>>. (June 29, 2012).

Data Information Literacy. "Project overview" (n.d.) *Data Information Literacy*. <<http://wiki.lib.purdue.edu/display/ste/About>>. (July 3, 2012).

FIRST. (2011). "Electronic courses." Office of the Vice President for Research of the University of Minnesota. <<http://cflegacy.research.umn.edu/first/electroniccourses.htm>>. (July 2, 2012)

Jeffryes, Jon and Lafferty, Meghan. (2012). "Gauging workplace readiness: Assessing the information needs of engineering co-op students." *Issues in Science and Technology Librarianship*. No. 69. <<http://www.istl.org/12-spring/refereed2.html>>. (July 3, 2012).

Johnston, Lisa and Lafferty, Meghan. (2011). "Data management 101: Planning checklist." <<https://netfiles.umn.edu/ul/Divisions/AcaProg/PhysSciEng/PSEShared/Instruction/Data%20Management/Public/Data%20Management%20Checklist%20Workshop.pdf>> (May 31, 2012)

Haug, Anders, Zachariassen, Frederik, and van Liempd, Dennis. (2011). "The costs of poor data quality." *Journal of Industrial Engineering and Management*. 4(2). 168-193.

Johnston, Lisa, Lafferty, Meghan, and Petsan, Beth. (2012). "Training researchers on research data management: A scalable cross-disciplinary approach." *Journal of eScience Librarianship*. 1(2):Article 2, <http://dx.doi.org/10.7191/jeslib.2012.1012>.

Marsh, Richard. (2005). "Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management." *Database Marketing & Customer Strategy Management*. 12(2). 105-112.

Port, S.R. (1992). "Data management in engineering" *Proceedings of the Institution of Civil Engineers: Civil Engineering*. 92(3). 133-137.

President's Emerging Leaders Program (2009). Implementing Research Cyberinfrastructure for the 21st Century. University of Minnesota Digital Conservancy. <http://purl.umn.edu/91772>

National Science Foundation (NSF). (2010). "Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans." *National Science Foundation*.

http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928&org=NSF. (May 23, 2012).

University of Minnesota Libraries. (2012). "Managing Your Data" *University of Minnesota Libraries*.

<http://www.lib.umn.edu/datamanagement>. (July 2, 2012).

Accepted Manuscript
Not Copyedited

Table 1: Data Life-Cycle Stages as described by the graduate students

Student Response	Initial	Second	Third	Fourth	Fifth
Grad #1	Raw sensor data	Processed Data	Processed with figures	Comparison (with other research)	Share the data (stages 1 and 2)
Grad #2	Raw	Excel 1	Excel 2	Stress Calculation / Force and Moment Calculation	Final Excel File
Grad #3	Raw numbers	Organization	Analysis and conclusion		
Grad #4	Data download from RMA website	Organize data into test folders and regular activity of bridge folder	Analyze data	Create Alarms to warn of potential problems on the bridge	
Generalized Data Stages	1. Raw Data	2. Collection & Organization	3. Processing and Analysis	4. Results	5. Sharing and Archiving

Accepted Manuscript
 Not Copyedited

Table 2: Interview Results Summary of Data Management Practices by Structural Engineering Graduate Students

Interview Question	Results of Student Interviews	Notes
Data Storage: How and where do you store your data?	"Desktop Computer" (100%)	All students stored the data on their workplace desktop computer. Half the students also stored a copy to an external drive attached to their computer.
Formal Policies: Are there formal procedures that you follow in managing your data set?	I don't know (50%)	Students were not aware of a formal procedure for managing the data they produce in their lab.
File Naming Systems: Do you use a file naming system to assign names to your data files?	Always (50%) "by month of date collected" Sometimes (25%) "download files then make a copy and rename them according to when they happened" I don't know (25%) "I've never even heard of a file naming system."	Students were not aware of a standard way to name and organize their files to allow someone else in the lab to navigate their directories and find their data.
Backup: Do you make backup copies of your data?	Always (50%) "occasionally and sporadic" Never (50%) "sort of by accident I do make backup copies because I have of course these multiple Excel files that have the same data in them."	Students do not always have a good foundation in data backup best practices that will protect their data from loss. One student described their intended backup practice as "Oh I will. I guess I will just store a copy in multiple locations on the same hard drive."
Security: Do you take security measures to protect your data?	Always (50%) "password protection to computer" Sometimes (25%) "authenticated computer and locked lab door" Never (25%) "computer has security and the web site is password"	Whether students were aware or not, they each had security measures in place to protect their data according to the university policies. However the students had to be prompted (eg. "Do you use a password to log into your work computer") to better illustrate they were using them.

Accepted Manuscript
 Not Copied

	protected”	
Version Control: Do you follow a system to identify and track different versions of your data?	<p>Always (75%) “ by date, labeled in the filename”</p> <p>Never (25%)</p>	
Authenticity: Do you provide a means to identify the “official” or “authoritative” version of your data?	<p>Always (25%) "place ‘Actual’ in the file name to indicate the official version”</p> <p>Sometimes (25%) “email current version to other group members”</p> <p>Occasionally (25%) “just remember it as the one that you work with the most”</p> <p>Never (25%)</p>	Students have poor file naming practices to manage the authoritative version of their data.

Accepted Manuscript
Not Copyedited

Table 3: Documentation and Metadata Practices Reported by Structural Engineering Students

Interview Question	Results of Student Interviews	Notes
Documentation: Please describe how this data set is documented and described.	Varies. "We write our own descriptions of the data within the excel files when the data are compiled."	Students described their ad hoc processes to document the data, such as labeling columns in excel. Additional information, such as the bridge sensor locations, were held in multiple and locations separate from the data files (eg. email correspondence or schematic drawings).
Reproduction: Is there sufficient documentation to reproduce the data?	Yes (100%)	One student described the experience with a NEES project and explained why documentation was required, "Because when we share those data on NEES, it's not gonna work if someone wants to share the data that nobody else can use. There is a standard way, which is stated by NEES."
Reuse: Is there sufficient documentation to understand and use the data?	Yes (100%)	The faculty member was not as positive as the students and rather thought that the current documentation would be "mostly" sufficient to allow a researcher in the lab to use the data, but only "maybe" sufficient for someone outside the lab.
Standards: Are there standard practices in your discipline to organize, document, or describe data?	I don't know (75%) No (25%)	One student may not have understood the practice of data management standards and commented "I think if i were to do it differently. I would just have fewer excel files."
Metadata: Are you familiar with the concept of metadata?	No (75%) "actually, I've never heard that word before" Yes (25%)	The necessary metadata to enable search, discovery, and interoperability are not understood by a majority of students interviewed.

Accepted Manuscript
 Not Copyedited

Table 4: Reusing and Sharing Data Responses by Structural Engineering Graduate Students

Interview Questions	Results of Student Interviews	Notes
External Data: Have you ever made use of data produced by a source outside your lab?	Yes (50%) No (50%)	
Citation: Do you know how to cite a data set?	No (75%) "I've never done that before" Yes (25%)	Interestingly the faculty member felt that, Yes, the students knew how to cite data according to disciplinary standards.
Data Repositories: What are the major repositories or sources of data in your discipline (if any)?	No Answer (75%) "NEES and Berkeley (BEER)" (25%)	Although not aware of existing repositories, one student felt that data repositories might be useful "so there aren't 50 external hard drives floating around"
Archiving: Have you ever deposited any data into a data repository?	Yes (50%) No (50%)	
Future Archiving: Are you planning on depositing your data in the future?	No (50%) Yes (25%) I don't know (25%)	The student working with NEES planned on depositing the data into NEESHub in the future. One student who answered No explained: "I would not because this data belongs to the state, so it doesn't belong to me."
Sharing: Would you be willing to share your data with someone outside the lab?	Yes, with conditions (75%) No (25%)	The students' who answered Yes said that they would do so after checking with their faculty advisor first. The students did not feel ownership with the data.
Publishing: Is data accepted for publication, as a supplement for example, in the journal in your field?	No (50%) I don't know (25%)	The students were unfamiliar with data publishing practices and one student commented "I could see them accepting it if it was a relatively small amount. I don't know that they would accept my many, many thousands of data points for any reason."

Accepted Manuscript
 Not Copied

Audience for Reuse: Who would you imagine might be interested in the data you are producing?	Researchers in the field (50%) State Contracting Agency (50%)	<i>"this particular bridge is a very specific design, so others would be interested because not a lot of research available. But the analysis itself, not the raw data" One graduating student also mentioned "the graduate student taking over my research" as a likely audience.</i>
Value: How might the data be reused by others?	<i>Varies</i>	One student described: "They could look at it again and perform different analysis. The MATLAB program assumes certain things about the concrete behavior, so if those aren't correct, they could make better assumptions than mine and come up with more accurate calculations." Also, another student thought that the state agency might "put the data in the archives for potential future use..." or "if there was a problem in the future."
Post-Graduation: Will you leave a copy of your data with your advisor after you graduate?	Yes (100%)	"the previous student left all the data on my computer desktop. I will probably leave this data on the external hard drive since computers get recycled throughout various students."

Accepted Manuscript
 Not Copyedited

Table 5: Preservation Responses by Structural Engineering Graduate Students

Interview Questions	Results of Student Interviews	Notes
How long would your data set be useful if it were to be preserved?	15-20 Years (75%) Other “For the life of the bridge” (25%)	
Are you or anyone else in the lab taking steps to preserve this data set?	No (50%) Yes (25%) I don’t know (25%)	The student who answered Yes felt that the funding agency was taking care to preserve the data set. Further information from the faculty member revealed that the agency preferred the data to reside in the lab.

Accepted Manuscript
Not Copyedited

Table 6: Summary of Needed Data Management Skills

Data Stage	DM Skills Needed by CE Graduate Students
1. Raw Data Generation	<p>All students need:</p> <ul style="list-style-type: none"> ○ Understand how sensors work and respond to physical phenomenon ○ Download sensor log files securely ○ Understand privacy issues associated with data (i.e. real-time bridge sensor data). ○ Track external events and understand how these affect raw data generation (temperature, weather, maintenance). ○ determine the best way to manage data collected over time (Data management plan) <p>Some students need:</p> <ul style="list-style-type: none"> ○ Writing a workplan/experimental design ○ Create/read schematic representation of sensor locations on physical bridge ○ Working with experimental laboratory personnel ○ Troubleshoot issues with sensor hardware attached to structures ○ Find/Reuse existing data
2. Collection & Organization	<p>All students need:</p> <ul style="list-style-type: none"> ○ organize data with temporal component (i.e. 15 min increments) ○ collect/organize data from multiple sensor sources into a single file ○ Track multiple versions of a data file shared with multiple people ○ Create documentation about data collection ○ Backup data appropriately ○ Co-locating metadata and processing actions with organized data ○ Create a custom file naming schema that can be easily understood by others ○ Emailing documents securely/efficiently (so not to make multiple versions) ○ Ability to track versions of the data and maintain authority control ○ Awareness of university security policies (using laptop for remote data collection) ○ Understand how to separate out data that was affected by external events <p>Some students need:</p> <ul style="list-style-type: none"> ○ Manage media files generated by MAST instrument (video, images)
3. Processing and Analysis	<p>All students:</p> <ul style="list-style-type: none"> ○ Create documentation of analysis steps for future graduate student or data reuse ○ Use known engineering theories to process data (temperature, age) ○ Apply equations to transform data into results (e.g. sensor frequencies)

Accepted Manuscript
Not Copyedited

	<p>into stresses)</p> <ul style="list-style-type: none"> ○ Compare data to simulation results ○ Ability to identify trends ○ Generate graphs and plots to visualize the data <p>Some students need:</p> <ul style="list-style-type: none"> ○ Understand how to find, use standards and code books ○ Utilizing programming tools such as Matlab to create simulation data. ○ Analyze data in instrument-specific software program
4. Results	<p>All students:</p> <ul style="list-style-type: none"> ○ Report results of data analysis ○ Ability to explain trends in the data ○ Creating/plotting graphs that accurately convey meaning of data <p>Some students:</p> <ul style="list-style-type: none"> ○ Report data to funding bodies (NSF, state contracts) ○ Show results in presentation format
5. Sharing & Archiving	<p>All students:</p> <ul style="list-style-type: none"> ○ Understand the potential for others to reuse data ○ Acknowledge the implications of accessing/sharing data using proprietary software (RMA website for AEWIn). ○ Understand the scientific value of sharing data ○ Archive data in the lab <p>Some students:</p> <ul style="list-style-type: none"> ○ Deposit data into discipline repository (NEES) ○ Archive locally on external hard drive
6. Preservation	<p>All students:</p> <ul style="list-style-type: none"> ○ Use file formats that allow long-term access ○ Create preservation backup copies ○ Understand the funder requirements for maintaining access to data ○ Understand the issues/problems associated with the preservation of data ○ Co-locate the data and documentation

Accepted Manuscript
Not Copyedited