Exploring Data Practices of the Earthquake Engineering Community

Shuheng Wu[1], Adam Worrall[2], and Besiki Stvilia[3]
[1]Queens College, The City University of New York
[2]University of Alberta
[3]Florida State University

Abstract
There is a need to compare and contrast data practices of different disciplines and groups. This study explores data practices in earthquake engineering (EE), an interdisciplinary field with a variety of research activities and dynamic data types and forms. Findings identify the activities of typical EE research projects, the types and forms of data produced and used in those activities, the project roles played by EE researchers in connection with data practices, the tools used to manage data in those activities, the types and sources of data quality problems in EE, and the perceptions of data quality in EE. A strong relation exists among these factors, with a stronger role for test specimens and high quality documentation and more blurring of project roles than in other fields. Suggestions are provided for resolving contradictions impeding EE researchers' curation and archiving activities and for future research on data practices.

Keywords: data practices; data curation; earthquake engineering; data quality; data repositories
Citation: Editor will add citation

Contact: Shuheng.Wu@qc.cuny.edu

1 Introduction

Modern science is characterized by technologies producing data at a rate exceeding scientists' ability to process, analyze, interpret, use and reuse, and manage it. As the quantity and diversity of scientific data are growing tremendously, preserving and archiving data can no longer be treated as post-project activities, but should be seen as part of daily research activities (Anderson, 2004). This has led to scientific collaboration on data management and curation. *Data curation* can be defined as "the activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and reuse" (Lord & Macdonald, 2003, p. 12). Digital libraries and institutional repositories should not only collect, organize, and preserve scientific literature, but also expand the scope of their services to meet the changing data management needs of their institutions and users, become involved in scientific data curation, and assist scientists with their daily archiving activities (Gray, 2007; Heidorn, 2011). To help manage or curate data, researchers study scientific data practices to gain an understanding on what constitutes data in a particular domain, the characteristics of data (e.g. provenance, quality, and ownership), and the activities centered around data. Data practices are contextual and vary by individual researcher, lab, project team, institution, discipline, and community. Previous studies (e.g. Borgman, 2012; Campbell et al., 2002; Palmer & Cragin, 2008; Stvilia et al., 2015; Witt, Carlson, Brandt, & Cragin, 2009) have identified the need for comparing and contrasting data practices of different disciplines and groups.

Earthquake engineering (EE) is an interdisciplinary field involving researchers from seismology, structural, mechanical, and geotechnical engineering concerned with saving lives and preventing damage from earthquakes and tsunamis (Pejša & Song, 2013). Engineers are a unique population; while "not always easy" to distinguish from natural scientists, they normally focus on "doing" instead of "knowing" (Petroski, 2010, p. 17). The interdisciplinarity, complexity, and diversity of the EE community lead to a variety of research activities and dynamic data types and formats (Pejša & Hacker, 2013), posing great challenges for data management and curation in EE. To address seismic risks in the United States, the National Science Foundation (NSF) founded the George E. Brown, Jr. Network for Earthquake Engineering Simulation (NEES, 2009), consisting of 14 geographically distributed laboratories performing different types of experiments (e.g. shake table tests, geotechnical centrifuge research, large-scale structural testing). NEES developed a cyberinfrastructure platform called NEEShub (https://nees.org/) to facilitate distributed collaborations, offer data curation services, preserve data in a repository named NEES Project Warehouse, and provide open access to experimental data and documentation (Pejša & Hacker, 2013). To enable data sharing and long-term preservation, NEES (2013) requires NSF-funded research teams to submit corrected data with necessary documentation to the NEES Project Warehouse within six months after an experiment ends. Twelve months after completing the experiment, the data will be made public after at the Warehouse.

This study explores data practices in EE, examining research project activities of EE researchers, their perceptions of and requirements for data quality, their data management and curation practices, and their interactions with NEEShub. Findings can inform the formulation of data management and curation policies; build knowledge for further developing and maintaining NEEShub to deliver data services and educate data curators and users; provide insight into building knowledge organization systems (metadata schema and ontologies); facilitate librarians, archivists, and curators' processes of appraising, selecting, and depositing data; and support the new tasks of digital libraries and data repositories.

2 Literature Review

There are a number of previous studies examining scientific data practices. Borgman, Wallis, and Enyedy (2007) adopted an ethnographic approach to study data practices of habitat ecologists and their collaborators gathered around the Center for Embedded Network Sensing (CENS). Guided by Activity Theory, Borgman et al. identified types of data produced by that community, their data sharing and publication practices, and their concerns on data quality and ownership. Stvilia et al. (2015) conducted a mixed-methods study to examine project tasks, perceptions of and priorities for data quality, and data management practices of the condensed matter physics community gathered around the National High Magnetic Field Laboratory in Florida. Campbell et al. (2002) conducted a nationwide survey to study data sharing and withholding practices of academic genetics researchers and compare with those of other academic life scientists. Their study identified types of data withheld by genetics researchers and the reasons and consequences of their withholding. Instead of focusing on a particular scientific community, Tenopir et al. (2011) used an online survey to explore data sharing and withholding practices of scientists from a variety of disciplines, including environmental sciences and ecology, social sciences, biology, physics, computer science, atmospheric science, and medicine. Tenopir et al. identified the tools and cyberinfrastructure supporting data sharing and scientists' perceptions of barriers and enablers of data sharing. Paine, Sy, Piell, and Lee (2015) applied three qualitative research methods to investigate data processing practices across four scientific research groups: atmospheric science, marine geophysics, microbiology, and empirical cosmology.

Data quality assurance is an indispensible part of data management and curation. *Data quality* can be defined as “the degree to which the data meet the needs and requirements of the activities in which they are used” (Stvilia et al., 2015, p. 247). Data quality encompasses a duality, incorporating subjectivity (meeting individual expectations) and objectivity (meeting activity requirements) (Stvilia, Gasser, Twidale, & Smith, 2007). Data quality is contextual, dynamic, and multidimensional. When data quality is evaluated at the individual level, one's domain knowledge and familiarity with the data repository can affect one's quality evaluation (Stvilia, Jörgensen, & Wu, 2012). When aggregated to the team, discipline, or community level, data quality can be understood differently, as seen in Stvilia et al.'s (2007) case studies of a digital library and of Wikipedia. Data quality can be affected by changes to the data, the underlying object it describes, or in the context of its creation and use. Data quality can be measured directly by examining the data; or indirectly by analyzing data provenance, the data creator's reputation, and the process of creating and (re)using the data. Perceptions and assessments of data quality vary by individuals, and within and across teams, disciplines, institutions, and communities (Ball, 2010).

Data quality control in the NEES Project Warehouse focuses on ensuring the validity of uploaded data through assessing the technical quality of documentation and supplied metadata describing the research workflow (e.g. project, experiment, trial), and the completeness of documentation (Pejša & Hacker, 2013). Technical quality is concerned with the format, integrity, and location of uploaded data. Other data quality dimensions such as accuracy, authority, precision, and relevance are not taken into account. Whether the data quality requirements set forth by NEES meet the requirements of the community and any potential data users is not known. Based on semi-structured interviews with 14 EE researchers, Faniel and Jacobesen (2010) studied how EE researchers assessed data quality when reusing their colleagues' experimental data for model validation. That study was concerned only with quality assurance during reuse of data for model validation, and did not address quality control for other types of data (e.g. specimen, documentation, computational) or in the context of other activities (e.g. performing experiments, analyzing data, or disseminating results).

Few studies have investigated the data practices of specific scientific disciplines or communities formed around lab facilities. To the best of our knowledge, no systematic studies have examined the data practices and research activities in EE. To facilitate effective data management and better support NEEShub and other cyberinfrastructure, in-depth understanding is needed of the EE community's current data practices, research project activities, and perceptions of and requirements for data quality.

3 Research Method

Guided by Activity Theory (Engeström, 1990; Kuutti, 1996; Leont'ev, 1978) and an Information Quality Assessment Framework (Stvilia et al., 2007), this study employed documentary analysis and semi-structured interviews (Blee & Taylor, 2002) to answer six research questions:

1) What are the typical activities of an EE research project?
2) What are the types and forms of data these activities produce and/or use?
3) What are the project roles the EE researchers play in those activities?
4) What are the tools the EE researchers use to manage data in those activities?
5) What are the types and sources of data quality problems in EE?
6) What are the perceptions of data quality in EE?

The researchers first conducted documentary analysis on the research data, documentation, and other relevant documents stored on NEEShub. This documentary analysis helped identify the community's data curation policies and guidelines, types of research data produced, and metadata standards (e.g. DataCite, PREMIS, PRONOM, Dublin Core) adopted; it also allowed an interview questionnaire to be developed. After documentary analysis, the researchers conducted semi-structured interviews with nine EE researchers from three research institutions between July 2014 and June 2015 regarding their research project activities, data practices, and data quality perceptions. Of the nine interviewees, one was an assistant professor, three were postdoctoral researchers, and five were doctoral students. Doctoral students and postdoctoral researchers were purposefully selected as interviewees because (a) they self-identified as responsible for data management and curation in their project teams, and (b) NEES perceives young researchers to be of special importance in archiving data and communicating with curators (Pejša & Hacker, 2013). All interviews, ranging from 25 to 68 minutes, were audio recorded, transcribed, and coded with NVivo 10. Two researchers independently coded all the interviews using an initial coding scheme based on Activity Theory (Engeström, 1990; Kuutti, 1996; Leont'ev, 1978), an Information Quality Assessment Framework (Stvilia et al., 2007), and documentary analysis. After comparing, discussing, and resolving any differences in their coding, the researchers formed a new coding scheme with emergent codes and subcategories and recoded all interviews.

4 Research Findings

4.1 Activities

Based on the documentary analysis and semi-structured interviews, the authors developed a typology of research project activities with specific tasks in EE (see Table 1): conceptualization, preparation, experiment, analysis and interpretation, archiving, publication and dissemination, administration, communication, and education.

Activities	Tasks
Conceptualization	Writing grant proposals
Preparation	Designing experiments, developing computational models, validating models, constructing specimens, writing construction summaries, creating testing protocols, installing sensors and cameras, pretesting sensors and cameras, recording sensor positions
Experiment	Observing tests, taking notes, taking photos, capturing data, storing data, backing up data
Analysis and interpretation	Processing data, assessing and/or improving data quality, analyzing data, running computational models, improving computational models
Archiving	Writing reports, providing documentation, organizing data, uploading data to the NEES Project Warehouse
Publication and dissemination	Writing articles, presenting in conferences
Administration	Construction management, project management
Communication	Group meeting, negotiating with industry partners, communicating with data curators
Education	Advising students, training interns, educational outreach activities

Table 1. Typology of Research Project Activities in EE

Some tasks, such as assessing or improving data quality and providing documentation, repeated in different activities. Before performing experiments, researchers develop computational models to simulate how the test specimen (e.g. house, tower) responds to earthquakes or tsunamis. To validate their models, they reuse data produced from other experiments to run on their models, compare the results with those of others, and then calibrate their models. One interviewee, developing a computational model, stated “sometimes I feel no confidence on this numeric model, so I have to use some data from other experiments, trying to calibrate [it].” After performing their own experiments, they rerun the models on the experimental data, compare simulation results with experimental results, and improve the models.

Similarly, creating or providing documentation occurs in different activities. One interviewee described how she created documentation during the preparation process:

> We had an initial set-up construction drawing with structural plans, but they changed a little bit. As the project was going and the building was being constructed, we would change the drawing to reflect what was actually being built. So it was kind of a continuous process while it was being built. As the sensors were installed, we were documenting exactly where they were, especially because in our test the big building was going to get destroyed after testing. So it was hard to go back afterwards and measure things.

Another interviewee mentioned he created documentation (laboratory notes) during the test, such as “drawing maps of cracks, and recording where there is a sound coming out.” All the interviewees, including computational researchers who are not performing experiments, stated that they would write reports and provide documentation after finishing experiments or close to the end of a project.

4.2 Types and Forms of Data

The authors also developed a typology of data produced or used by EE researchers. Corresponding to the identified research project activities, the data can be categorized (see Table 2) as experimental data, computational data, documentation, test specimens, secondary data, publications, and presentations. In terms of state, the data can be classified as raw data, processed data, analyzed data, verified data, certified data, and archived data. Certified data are particularly the experimental data meeting the curation criteria set forth by NEES (Pejša & Hacker, 2013). Archived data are those accepted to the NEES Project Warehouse and made accessible to the public.

Data Types	Data
Experimental data	Sensor measurements, videos, images
Computational data	Simulation models, software, programming code
Documentation	Grant proposals, project executive summaries, specimen design drawings, specimen structural plans, construction drawings, construction summaries, instrumentation plans, sensor metadata, experiment notes, project reports, experimental setup reports, presentations, meeting minutes
Test specimen	Buildings, columns, walls, nonstructural building components
Secondary data	Earthquake data, online databases, government data, published papers, reports, conference proceedings, experimental data produced by others, simulation models developed by others
Publications	Journal articles, theses
Presentations	Conference presentations
Communication data	Emails

Table 2. Types of Data Corresponding to the EE Research Project Activities

The forms of data produced and used by EE researchers are diverse, including but are not limited to data in ASCII format captured by the data acquisition systems, images, videos, digital drawings (e.g. AutoCAD files), simulation models, software or programming code, test specimens, statistics files, spreadsheets, laboratory notes, text documents, presentation files, databases, and web sites.

4.3 Project Roles

Of the nine interviewees, seven indicated that they did both computational and experimental research; one did both computational and theoretical research; and one did purely computational work. Most of the interviewees implied that they played multiple roles in their project teams. The assistant professor identified his project role as principal investigator (PI) and student advisor. Two doctoral students and two postdoctoral researchers indicated their role as project lead, being in charge of writing proposals; designing, preparing for, and performing experiments; processing, analyzing, managing, and archiving data; and writing and publishing articles. Most of the doctoral students and postdoctoral researchers interviewed saw themselves in the lead role in their own project, but with necessary assistance and guidance coming from their advisor. One doctoral student participating in a large-scale project, which took more than a year to build a test specimen, indicated that one of her project roles was construction manager, responsible for "making sure that the construction was on track, the schedule was in place, and everything was getting done." Another doctoral student who was part of the same project, though identifying his role as graduate research assistant, was involved in nearly every task of the project: designing and constructing the building; writing weekly construction summaries; creating and maintaining the project web site; performing the experiment; developing computational models; managing and archiving data; and writing test reports, presentations, and journal articles.

4.4 Tools

According to Activity Theory (Engeström, 1990; Kuutti, 1996; Leont'ev, 1978), tools can be defined as the external objects or internal symbols that researchers use in their research project activities. The EE researchers used various types of tools in their data-related activities. Experimental researchers used different types of sensors (e.g. accelerometers, strain gages, linear and string potentiometers, cameras, linear variable differential transformers) to measure the acceleration, displacement, strain, and force of structural and nonstructural components during experiments. Some laboratories have a data acquisition system (e.g. LabVIEW) to automatically collect data from those sensors. Computational researchers interviewed mostly used OpenSees and LS-DYNA software to develop computational models to simulate responses of structural and geotechnical systems to earthquakes. MATLAB, Mathcad, and Microsoft Excel are other popular software the EE researchers used to process or improve the quality of experimental data, such as for noise filtering. Besides the tools mentioned above, some computational researchers used C++, Python, Fortran, Tool Command Language (TCL), and Linux shell scripting to write their own code to process data. One computational researcher who was heavily involved in software development used version control software (CVS, SVN) for version management of his code. Images and videos are one of the main categories of data produced by this community. Some EE researchers interviewed used Adobe Premiere, Adobe Photoshop, ParaView, and Final Cut Pro to process those data for analysis, presentations, papers, and reports.

When asked whether they disposed data or not, all the interviewees emphasized that they kept almost everything and rarely deleted data, especially experimental data, because it was nearly impossible to rerun the experiments. The only situations in which they would delete data were when sensors were not working or malfunctioned, or computational models were incorrect. Most interviewees used external hard drives, portable drives, personal or lab computers, local servers, and cloud storage systems to store and back up data. Four interviewees pointed out that they had insufficient storage space; for example:

> [If] I run my computational model once, it generates more than 50 gigabytes of data. So if I do the parametric study, I will run the model several times and a lot of data will be generated. I will run out of memory [*sic*] very soon.

To resolve the storage problem, sometimes researchers would store the data in different locations. One computational researcher revealed he used at least four cloud storage systems (Dropbox, Google Drive, Baidu Cloud, and 360 Cloud) at the same time.

In terms of creating or providing documentation, most of the interviewees indicated they followed the Data Sharing and Archiving Guidelines proposed by NEES (2013) to provide the required documentation and reports. Some used the same software as for data analysis (e.g. MATLAB) to provide metadata (e.g. source code comments). In terms of archiving data, NEEShub has developed some software and system tools for project teams to upload different types of data to the Project Warehouse. For researchers who were not working on NEES/NSF funded projects, they stored or archived the data on their personal computers, external hard drives, and local servers.

Computational researchers used secondary data (e.g. publications, experimental data produced by others) to calibrate and validate their models. Similarly, experimental researchers used secondary data (e.g. earthquake data) from online databases and government web sites to help with experiment design. For example, one interviewee described his use of secondary data collected from online databases:

> We have to select a few ground motions for this test because we're doing the shake table test. You have to go to the online databases to collect previous data regarding ground motion excitations, and then analyze it.

4.5 Data Quality

Data quality problems occur when the data cannot meet the needs and requirements of the activities in which the data is used (Stvilia et al., 2007). The data quality problems encountered by interviewees included inaccurate, incomplete, and inconsistent data or metadata. The sources of those data quality problems included incomplete or missing documentation or metadata, instrument errors, lack of instruments, imprecise instruments, human errors, external environmental interferences, and lack of version control. For example, a computational researcher described a data quality problem (inconsistent data) she had encountered, which was caused by lack of metadata or documentation:

> I had the experience of using data from a database. They changed the data a few months later, and later they changed it back. For the user I don't know what happened and which data is accurate, because the data was basically changed back and forth. If they have a log recording when the changes happened and why, it would be very helpful.

Another interviewee described how some experimental data became useless to his project team because of human errors and lack of metadata:

> We take pictures everyday. We kind of have a rule [to describe the pictures]: date and [photographer's] name, date first and then your name. Somebody just dumped thousands of pictures without any characterization. It will be hard to find out when you took these pictures and why you took them. And those become kind of useless.

EE researchers can have difficulty in reusing others' experimental data because of incomplete data or documentation provided in publications and reports:

> I'm writing some papers. Because I only did my test on two specimens, if I want to verify some equations I proposed, I have to find other people's research data, and then use that to verify my theories. Because I didn't do their test, what I can do is just to read their report or published paper. And hopefully they have some description of the data, but usually it's difficult. They may not publish the one [description] you need.

When asked about their perceptions of data quality, interviewees were provided with a list of 14 quality criteria (dimensions) with definitions adapted from Stvilia et al.'s (2015) previous study of the condensed matter physics community. Interviewees were asked to identify whether those criteria were applicable in their work context and if any criteria were missing from the list. None of the criteria in the list were perceived inappropriate by all the interviewees, except for currency. Interviewees emphasized they did not care about the age of data, with one interviewee explaining:

> The age of data is not important. Old data doesn't mean it's bad. But if it's what you need, for example, an earthquake record from the 1960s, it's just as valid as the one last week.

Another interviewee expressed similar viewpoints:

> [Earthquake engineering] was started, like, [in the] early 1960s. And people are still doing work in that field, but most of the premier works were actually done, like, 50 years ago. So at least for me old data is also very important.

The interviewees perceived accessibility, accuracy, authority, completeness, consistency, redundancy, reliability, validity, and verifiability as relatively more important. Nearly every experimental researcher pointed out sensors not working or malfunctioning as a source of data quality problems—incomplete or inaccurate data—they had encountered. They would install more sensors to the test specimens than needed to ensure the completeness of data. One of them explained why experimental researchers value redundancy in their experimental data:

> If the sensors malfunction, we get very bad data. We pretty much just don't use the data because we have a lot of redundancy. If this sensor is not working, we have another sensor adjacent to it.

5 Discussion

5.1 Types and Forms of Data

Based on the findings of this study, the types of data produced or used by EE researchers largely depend on whether they are doing experimental or computational work. Experimental research produces raw data (e.g. sensor measurements, images, videos) and a variety of documentation and metadata to describe the specimens, instruments, and experiment settings, such as specimen drawings, structural plans, construction summaries, instrumentation plans, sensor metadata, and project reports. Computational research generates simulation models and documentation explaining those models, and software or programming code for processing, analyzing, and transferring the data. Borgman et al. (2007) categorized data in habitat ecology by their states as raw data, processed data, verified data, models, software, and algorithms. Stvilia et al. (2015) identified the data types in condensed matter physics, and extended Borgman et al.'s typology to include three more types: text documents, presentations, and visualization data. The current study developed three typologies of data, based on EE researchers' project activities, and the state as well as form of their data. Compared to previous studies, the additional data types in EE include secondary data, test specimens, archived data, and documentation. Archived data means data that has been accepted to the NEES Project Warehouse and is accessible to the public; this data is the product of NEES's data sharing and archiving policies. To be archived data, the experimental data has to meet the curation criteria set by NEES (including necessary documentation and in a certain format) to become certified data first, and then becomes public at the NEES Project Warehouse and can be reused by the EE community. The process of data being changed from experimental data to archived data involves the data curation and archiving activities of EE researchers and NEEShub curators.

Test specimens—such as houses, buildings, and columns—emerged as one of the more unique data types produced and used by EE researchers. Researchers in ecology (Borgman et al., 2007), genetics (Campbell et al., 2002), and condensed matter physics (Stvilia et al., 2015) do use specimens as data (albeit under different names, e.g. samples) that are subject to an experimental condition or observation. However, the specimen itself is a key data source in EE; it is designed, constructed, tested, observed, measured, and operated on at almost every stage and in many of the activities of EE research. It is also dynamic, continuously being changed before the start (preparation) until the end of experiments. Documentation activities in EE have to be continuous, starting from before the experiments, to reflect the changes in test specimens. In other fields, documentation may not be created until the experiments are performed; for example, documentation in condensed matter physics is mostly created when the researchers are collecting and analyzing data (Stvilia et al., 2015). The role of test specimens in EE researchers' data practices is perhaps unique in its strength, centrality, and importance to all in the field.

Documentation in EE comes in a variety of types and forms, and is perceived as extremely important, particularly for experimental researchers. The diversity and complexity of their research project activities (see Table 1) and the irreproducibility of some of their experiments leads to this importance. Some test specimens (e.g. houses, buildings), which took months and cost hundreds of thousands of dollars to build, were destroyed after the experiments. Without accurate and complete documentation, other researchers may not be able to interpret, use, and reuse the experimental data, which may be impossible to reproduce. This finding echoes the research of Borgman et al. (2007) and Stvilia et al. (2015) in different fields, although the stages and activities where EE documentation took place vary from documentation in other fields. One experimental researcher explained that without the documentation required by NEES, others could not interpret his data:

> NEES has very specific guidelines on what sort of data is a minimum requirement when you're uploading your experiment data to the web site. They need the overall photographs of the specimen; a plan view, a profile view, and a site view of the specimen; a layout of the instrumentation plan, where you put the actual instrumentations. Otherwise if I don't give the instrumentation plan, nobody knows where strain gage 1 and strain gage 2 are.

While the rules put in place by NEES drove most documentation activities, one interviewee mentioned that his project team created additional documentation for internal use to ensure data consistency:

> We have a lot of professors and students [in our project team]. So in order to keep consistency in our future publications, I prepared a summary of the building response data, like a table for people to use, [indicating] what's the peak acceleration at the roof, what's the peak displacement at the first floor. Just to make sure in our future publications we don't conflict with each other.

Documentation also took place at multiple stages before, during, and after experiments. It emerged as one of the most important activities in EE researchers' data practices, and one of the key factors in maintaining high accessibility, completeness, consistency, and verifiability for the data collected.

As Borgman (2012) stated, whether documentation is kept often depends on the cost and reproducibility of an experiment. Experiments in EE are often of high preparatory cost—large buildings and houses taking time to construct and set up—and not easily reproduced without careful explanation of the construction, set up, and data collection processes. NEES requires such documentation and engages in quality control and assurance processes surrounding it (see Pejša & Hacker, 2013) because they understand its importance to this specialized field, as interviewees did in this study, and know that publication venues (e.g. conferences, journals) do not always request or include such documentation (see also Borgman, 2012). Necessary documentation may be missing from publications—"they may not publish the one [description] you need"—and human error can lead to certain conditions not being documented, units of measurement being left out, or documentation rules being followed inconsistently. Faniel and Jacobsen (2010) also found difficulty in reusing experimental data, but the high monetary and time costs for EE research stress the importance of quality documentation.

5.2 Project Roles

As detailed above, interviewees often played multiple roles in their project teams, even when their official status was a doctoral student or postdoctoral researcher. While advisors and PIs were in charge of the projects, it often fell to the junior researchers to design and construct specimens, maintain project documentation (both public- and private-facing), manage and archive data, and write substantial portions of reports and publications. Many self-identified as project leads, despite the presence of advisors. This blurring of roles for junior researchers, and their potential to take over many data management and curation tasks, appears more common in the EE field than in condensed matter physics (Stvilia et al., 2015), although in both fields students are more involved in day-to-day experimental activities than their advisors (see Burnett et al., 2014). While work on project roles in scientific collaborations exists (see e.g. the review by Sonnenwald, 2007), the authors are not aware of significant additional work on the project roles played by scientific researchers in specific connection with data practices, and believe further cross-discipline and cross-context research in this area is necessary.

5.3 Data Quality Perceptions

The EE researchers perceived accessibility as an important data quality criterion. One interviewee admitted the data sharing and archiving policies enforced by NEES enable easy access to experimental data. He pointed out the difficulty of gaining access to data produced by non-NEES/NSF funded projects:

> For all the NSF funded projects, they [NEES] require you to upload data to NEEShub. But there are some projects, which are not funded by NSF and not supported by NEES. So in those cases, it might be hard for somebody else to actually get access to those data.

Two interviewees, who were not experimental researchers, had not participated in any NEES/NSF-funded projects. When asked how they archived data, both indicated they stored data in their personal computers and backed it up to external hard drives. NEEShub accepts research data produced by non-NEES/NSF funded projects provided the data are approved by the NEES curators and meet NEES's minimum requirements (Pejša & Hacker, 2013). However, neither interviewee indicated they would bother to do so.

5.4 Contradictions and Suggestions

According to Activity Theory, *contradictions* refer to historically accumulated tensions or instabilities within or between activities, playing a central role in changing, developing, and learning those activities (Allen et al., 2011; Roos, 2012). Contradictions may exist within each component (i.e., subject, tool, objective, rule, division of labor, and community) of an activity, between components of the activity, between different developmental phrases of the activity, and between different but interconnected activities (Engeström, 1990). Contradictions are sources of problems and development for activities (Kuutti, 1996). This study found contradictions within the curation/archiving activities, and conflicts between curation/archiving activities and research activities. Based on the identified contradictions, the authors provide suggestions in the following section for resolving the contradictions impeding the curation/archiving activities.

NEES requires project teams to upload experimental data and necessary documentation to the Project Warehouse for the purpose of long-term preservation and future reuse by the community. However, three interviewees implied that the tools developed by NEES could not help them upload data to the Warehouse with efficiency and in a timely manner. One interviewee who knows Linux explained why he would rather write his own code to upload data than use the tools developed by NEES:

> NEES requires people to use those tools they developed. But they are not convenient for data uploading and far from robust from a decent Linux user's point of view. You spend more time to communicate with them [NEES curators] and wait for them to fix the bugs for a simple task, such as uploading a few videos.

Another interviewee indicated his project team had difficulty with NEEShub tools and guidelines:

> When I'm uploading data using the NEEShub tool, there are a lot of problems. We have a lot of difficulties to upload data correctly. For example, for Excel files, it requires a certain format. If you didn't follow exactly the same format, you're not able to upload the data. We've been doing a lot of things to address this issue.

NEEShub may consider allowing more flexibility in the file format and developing additional software to automatically convert file formats on the NEEShub side to save project teams' time and efforts. NEEShub may also enable researchers to use their own tools to upload data where possible, and collaborate with computational researchers with programming skills to improve the current data-uploading tools.

One interviewee, a doctoral student responsible for data management and curation in her project team, implied that the archiving policy (i.e., uploading data and documentation within six months after the experiment is completed) formulated by NEES prevented her from doing research activities:

> Data management and archiving is a very time consuming process. If someone's doing a research project, just because it's a large project, it'll have a lot more data and need a lot more data management. So having to do that as a student takes a lot of research time away. I think that these tasks may could have been performed by staff, or if there was an easier software or some type of equipment that made it easier, somehow!

As suggested by this interviewee, NEES may consider investing more resources to develop data archiving tools to ease the process of data management and curation for project teams. Since the laboratories of NEES are located in 14 universities across the United States, NEES may collaborate with the libraries or institutional repositories of those universities, and train subject or metadata librarians to be local facility curators to help project teams with their data curation and archiving activities. Four of the 14 universities have iSchools, which may lead to educational and support opportunities in data management.

6 Conclusions and Future Research

This study explores data practices in the EE community based on nine semi-structured interviews, uncovering a clear and strong relation between the typical activities of EE research projects, the project roles EE researchers play in those activities, and the types and forms of data produced and used in those activities. This paper does not report findings on data ownership and data sharing; these will be reported in future publications. This study is limited in that most interviewees were postdocs or doctoral students. More interviews should be conducted with researchers holding other academic or research positions (PIs, professors, lab managers, curators) to gain different perspectives. Another limitation is that seven of the nine interviewees, who claimed to be doing both experimental and computational research, were all participating in NEES/NSF-funded projects. The other two non-experimental researcher interviewees, did not work on any NEES funded projects. Future research will interview more non-experimental researchers and compare their data management and curation practices and quality perceptions with those of experimental researchers, to learn more about NEEShub's influence on data practices.

7 References

Allen, D., Karanasios, S., & Slavova, M. (2011). Working with Activity Theory: Context, technology, and information behavior. *Journal of the American Society for Information Science and Technology, 62*, 776-788. doi:10.1002/asi.21441

Anderson, W. L. (2004). Some challenges and issues in managing, and preserving access to, long-lived collections of digital scientific and technical data. *Data Science Journal, 3*, 191-202.

Ball, A. (2010). *Review of the state of the art of the digital curation of research data* (ERIM Project Report No. erim1rep091103abl2). Bath, UK: University of Bath. Retrieved from http://opus.bath.ac.uk/19022/

Blee, K. M., & Taylor, V. (2002). Semi-structured interviewing in social movement research. In B. Klandermans & S. Staggenbory (Eds.), *Methods of social movement research* (pp. 92-117). Minneapolis, MN: University of Minnesota Press.

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*, 1059-1078. doi:10.1002/asi.22634

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries, 7*, 17-30. doi:10.1007/s00799-007-0022-9

Burnett, G., Burnett, K., Kazmer, M. M., Marty, P. F., Worrall, A., Knop, B., Hinnant, C. C., Stvilia, B., & Wu, S. (2014). Don't tap on the glass, you'll anger the fish! The information worlds of distributed scientific teams. In P. Fichman & H. Rosenbaum (Eds.), *Social informatics: Past, present, and future* (pp. 118–134). Newcastle, UK: Cambridge Scholars Publishing.

Campbell, E. G., Clarridge, B. R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A., & Blumenthal, D. (2002). Data withholding in academic genetics. *Journal of the American Medical Association, 287*, 473-480.

Engeström, Y. (1990). *Learning, working and imagining: Twelve studies in activity theory*. Helsinki, Finland: Orienta-Konsultit Oy.

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work, 19*, 355-375.

George E. Brown, Jr. Network for Earthquake Engineering Simulation (NEES). (2009). *About NEES*. Retrieved from https://nees.org/about

George E. Brown, Jr. Network for Earthquake Engineering Simulation (NEES). (2013). *NEEScomm Data Sharing and Archiving Policies*. Retrieved from https://nees.org/resources/6218/download/ Data_Sharing_and_Archiving_Policy_20130501.pdf

Gray, J. (2007). Jim Gray on eScience: A transformed scientific method. In: T. Hey, S. Tansley, & K. Tolle (Eds.), *The fourth paradigm: Data intensive scientific discovery* (pp. 5-12). Edmond, WA: Microsoft Research.

Heidorn, P. B. (2011). The emerging role of libraries in data curation and e-science. *Journal of Library Administration, 51*, 662-672. doi:10.1080/01930826.2011.601269

Kuutti, K. (1996). Activity Theory as a potential framework for human-computer interaction research. In Nardi, B. A. (Ed.), *Context and consciousness: Activity theory and human-computer interaction* (pp. 17-44). Cambridge, MA: MIT Press.

Leont'ev, A. (1978). *Activity, consciousness, personality*. Englewood Cliffs, NJ: Prentice–Hall.

Lord, P., & Macdonald, A. (2003). *E-Science curation report: Data curation for e-Science in the UK: An audit to establish requirements for future curation and provision*. Bristol, UK: JISC.

Paine, D., Sy, E., Piell, R., Lee, C. P. (2015). Examining data processing work as part of the scientific data lifecycle: Comparing practices across four scientific research groups. In G. Olson (Chair), *Proceedings of 2015 iConference*. Retrieved from http://hdl.handle.net/2142/73644

Palmer, C. L., & Cragin, M. H. (2008). Scholarship and disciplinary practices. *Annual Review of Information Science and Technology, 42*, 163-212. doi:10.1002/aris.2008.1440420112

Pejša, S, & Hacker, T. (2013). Curation of earthquake engineering research data. In *Proceedings of Archiving Conference 2013* (pp. 245-250). Springfield, VA: Society for Imaging Science and Technology.

Pejša, S, & Song, C. (2013). Publishing earthquake engineering research data. In J. S. Downie & R. H. McDonald (Chairs.), *Proceedings of the 13th ACM/IEEE Joint Conference on Digital Libraries*. New York, NY: ACM.

Petroski, H. (2010). *The essential engineer: Why science alone will not solve our global problems*. New York, NY: Alfred A. Knopf.

Roos, A. (2012). Activity theory as a theoretical framework in the study of information practices in molecular medicine. *Information Research, 17*(3). Retrieved from http://www.informationr.net/ir/17-3/paper526.html

Sonnenwald, D. H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology*, *41*, 643–681. doi:10.1002/aris.2007.1440410121

Stvilia, B., Gasser, L., Twidale, M., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology, 58*, 1720-1733.

Stvilia, B., Hinnant, C. C., Wu, S., Worrall, A., Lee, D. J., Burnett, K., Burnett, G., Kazmer, M. M., & Marty, P. F. (2015). Research project tasks, data, and perceptions of data quality in a condensed matter physics community. *Journal of the Association for Information Science and Technology, 66*, 246-263. doi:10.1002/asi.23177

Stvilia, B., Jörgensen, C., & Wu, S. (2012). Establishing the value of socially created metadata to image indexing. *Library and Information Science Research, 34*, 99-109. doi:10.1016/j.lisr.2011.07.011

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

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. *The International Journal of Digital Curation, 3*(4), 93-103.