This is a preprint of an article accepted for publication in *Library Hi Tech*. (Wu, S., & Worrall, A. (2019). Supporting successful data sharing practices in earthquake engineering. *Library Hi Tech*. Retrieved from https://doi.org/10.1108/LHT-03-2019-0058)

Supporting Successful Data Sharing Practices in Earthquake Engineering

Shuheng Wu
Graduate School of Library and Information Studies, Queens College, 65-30 Kissena Blvd., Queens, NY 11367-1597. Email: Shuheng.Wu@qc.cuny.edu

Adam Worrall
School of Library and Information Studies, University of Alberta, 5-168 Education North, Edmonton, Canada. Email: worrall@ualberta.ca

Abstract

Purpose – Prior studies identified a need for further comparison of data sharing practices across different disciplines and communities. Towards addressing this need this study examined the data sharing practices of the earthquake engineering (EE) community, which could help inform data sharing policies in EE and provide different stakeholders of the EE community with suggestions regarding data management and curation.

Design/methodology/approach – This study conducted qualitative semi-structured interviews with 16 EE researchers to gain an understanding of which data might be shared, with whom, under what conditions, and why; and their perceptions of data ownership.

Findings – This study identified 29 data sharing factors categorized into five groups. Requirements from funding agencies and academic genealogy were frequent impacts on EE researchers' data sharing practices. EE researchers were uncertain of data ownership and their perceptions varied.

Originality/value – Based on the findings, this study provides funding agencies, research institutions, data repositories, and other stakeholders of the EE community with suggestions, such as allowing researchers to adjust the timeframe they can withhold data based on project size and the amount of experimental data generated; expanding the types and states of data required to share; defining data ownership in grant requirements; integrating data sharing and curation into curriculum; and collaborating with library and information schools for curriculum development.

Keywords: Data sharing, Data ownership, Data practices, Data repositories, Earthquake engineering

Paper type: Research paper

1 Introduction

Modern scientific practices are characterized by computational technologies generating data at a rate beyond researchers' abilities to process, analyze, use, manage, and share them. Many government agency and university-based initiatives have aimed at the challenge of scientific data management, bringing together librarians, archivists, scientists, and system developers to reach specific scientific communities, study their disciplinary activities, and address their data management and sharing problems (Borgman, 2012; Witt *et al.*, 2009). Funding agencies believe data sharing can enhance scientific research. For example, in the United States the National Endowment for the Humanities (NEH, 2019), the National Science Foundation (NSF, 2011), and the National Institutes of Health (NIH, 2003) require applicants to submit data management plans specifying how they will disseminate and provide access to their data. Journals (e.g., Nature, Science, PLoS ONE) have implemented data sharing or depositing policies, requiring authors to make data underlying their findings available.

To allow for data sharing in earthquake engineering (EE), NSF founded the George E. Brown, Jr. Network for Earthquake Engineering Simulation (NEES), developing a cyberinfrastructure platform named NEEShub to provide experimental facilities, data curation services, and open access to experimental data and documentation (Pejša and Hacker, 2013; Pejša *et al.*, 2014). NEES required NSF-funded research projects to submit corrected data with necessary documentation to NEEShub within six months after an experiment ends. The data would become public at NEEShub twelve months after completing the experiment. To broaden the support for other natural hazards (e.g., windstorms, tsunamis, coastal flooding) engineering research, NSF recently founded the National Hazards Engineering Research Infrastructure (NHERI) to replace NEES, while continuing its emphasis on the EE research previously supported by NEES (Rathje *et al.*, 2017). To succeed NEEShub, NHERI built a new cyberinfrastructure platform named DesignSafe consisting of experimental facilities located at eight universities in the United States. At the heart of DesignSafe is a central open data repository named Data Depot that now hosts NEEShub-published data and supports the full lifecycle of research in natural hazards engineering, from data creation and analysis to curation and publication. Unlike NEES's 12 month requirement, NHERI does not set firm deadlines for research projects performed at NHERI's experimental facilities to publish data in Data Depot, but recommends timelines for publishing different data types (DesignSafe, n.d.a).

Data sharing can be defined as "the release of research data for use by others" (Borgman, 2012, p. 4). Besides releasing data in open repositories like Data Depot, data sharing in EE may include private exchanges between researchers; publishing data in journals, websites, wikis, or blogs; and presenting data in conferences. Data sharing practices vary by individuals, and within and across teams, disciplines, institutions, and communities. Prior research (Borgman, 2012; Cragin *et al.*, 2010; Faniel and Zimmerman, 2011; Fecher *et al.*, 2015; Kowalczyk and Shankar, 2011; Tenopir *et al.*, 2015; Tenopir *et al.*, 2018) has identified a need for further study and comparison of data sharing practices of different science and engineering disciplines, including identification of discipline-specific enablers and barriers for data sharing that are of ever-increasing importance due to the requirements from funding agencies and publications.

Towards addressing this need, this study examined the data sharing practices of the EE community, gaining an understanding of which data might be shared, with whom, under what conditions, and why alongside their responses to the data-sharing policies by funding agencies. Answers to these questions can inform the formulation of data sharing and curation policies, the further development and maintenance of cyberinfrastructure platforms, the delivery of services by data or institutional repositories, and the education of data producers, curators, and users.

2 Literature Review

Data sharing is a sociotechnical practice (Kowalczyk and Shankar, 2011; Van House, 2003). It relies on the technical infrastructure (e.g., data repository, metadata schema, management software) to ensure the persistence, longevity, security, and quality of data. From the data curation perspective, the U.S. National Science Board (NSB, 2005) categorizes three types of infrastructure supporting the collection, curation, analysis, and sharing of digital data: (a) research data collections produced from individual researchers or research projects, (b) community data collections serving a specific science or engineering community, and (c) reference data collections serving a diverse set of communities (e.g., students, scientists, or educators). King (2007) introduced the Dataverse Network as an infrastructure for data sharing

within scientific communities to meet their requirements for recognition, public distribution, authorization, validation, persistence, ease of use, and legal protection.

Besides technical infrastructure, data sharing practices are influenced by complex social, organizational, cultural, and ethical factors such as research ethics, institutional policies, and disciplinary norms (Borgman, 2012; Fecher *et al.*, 2015; Kowalczyk and Shankar, 2011; Van House, 2003). Some of the motivations for data sharing include requirements from funding agencies and journals, an increase in citation rates, receiving feedback from other researchers, enhancement of reputation, expectations of reciprocity, networking with other researchers, and altruism. Data sharing supports scientific, educational, and socio-economic benefits; verifying, reproducing, and refining results produced by others; applying new tools to extant data; asking new questions; and conducting interdisciplinary and longitudinal research. It reinforces open inquiry, advances the impact of research, and reduces the cost of repetitively producing data. Data sharing also allows citizens and educators to use publicly funded project data, furthering public understanding of scientific research, promoting scientific education, and making citizen science possible.

Despite the benefits of and motivations for data sharing, researchers may be concerned about others misusing or misinterpreting their data, losing absolute control over them, incorrect or insufficient citations of them, exposing data collection methods for criticism, disclosing sensitive and private information, violating intellectual property rights, losing competitive advantages, spending substantial time and effort to make data reusable, losing ownership of data after depositing them, and insufficient technical support and documentation provided by repositories (Birnholtz and Bietz, 2003; Borgman, 2012; Borgman *et al.*, 2007; Fecher *et al.*, 2015; King, 2007; Kowalczyk and Shankar, 2011; Park *et al.*, 2018; Zimmerman, 2008). These concerns can present barriers to data sharing and to the roles data and their documentation can serve as boundary objects for translation of existing knowledge and creation of new knowledge across communities (Star and Griesemer, 1989); as Van House (2003) and Zimmerman (2008) found, both local and global contexts of data must be documented within a standardized infrastructure for successful knowledge sharing.

Prior work examined data sharing practices in different disciplines and communities. A 1993-1994 survey found data withholding commonly occurred among academic life scientists (Blumenthal *et al.*, 1997); geneticists were more likely to deny requests for data due to increasing competitiveness and opportunities for industry funding. Campbell *et al*. (2002) found the reasons for this withholding, in order of popularity, were (a) effort required to produce, (b) to protect publishing ability, (c) financial costs to provide, (d) concerns the requester would never reciprocate, (e) industrial sponsors' requirements prohibiting sharing, (f) to protect human subjects' privacy, and (g) to protect commercial value. Borgman *et al.* (2007) studied the data practices of habitat ecologists and their collaborators around the Center for Embedded Network Sensing (CENS). CENS scientists were more willing to share published, sensor-collected, and contextual data than unpublished, hand-collected, and experimental data. Concerns with commercial reuse, the state of data, efforts to collect data, data recipients, reciprocity, data citations, and temporal conditions influenced CENS scientists' willingness to share data. Tenopir *et al.* (2018) used an online survey to examine the motivations, attitudes, and data sharing practices of geophysicists from the American Geophysical Union. Although geophysicists had positive attitudes toward data sharing and ruse, they had concerns over data misuse and lack of proper citation and acknowledgement.

Others have compared data sharing practices across disciplines. Tenopir *et al.* (2011) identified perceptions of barriers and enablers of data sharing across six disciplines. Cragin *et al.* (2010) interviewed 20 scientists from 12 disciplines in small sciences, finding no field-wide norms for data sharing. Birnholtz and Bietz (2003) compared how researchers in three disciplines—EE, HIV/AIDS, and space physics—used data to inform the development of computer-supported cooperative work (CSCW) systems to support data sharing. They found EE researchers were more willing to share abstractions of experimental data than the full dataset for at least six months, as they wanted to maintain control of the data and findings. Birnholtz and Bietz (2003) also found EE researchers used and shared mostly confirmatory, event-driven data, versus greater use of data providing new or unexpected results in HIV/AIDS and space physics research. Fecher *et al.* (2015) surveyed 603 secondary data users of the German Socio-Econic Panel Study (SOEP). Coupled with a literature review, they developed a conceptual framework consisting of six categories to describe the process of data sharing in academia: data donor, research organization, research community, norms, data infrastructure, and data recipients. Based on surveying 1,317 researchers from 43 STEM disciplines, Kim and Stanton (2016) found regulative pressure from journals, normative pressure within disciplines, perceived career benefits (e.g., citations, co-authorship), and scholarly altruism had significant positive relationships with researchers' data-sharing behaviors; regulative pressure from funding agencies, the availability of data repositories, and perceived career risks (e.g., misuse of data, losing publication opportunities) had significant negative impacts.

Few studies have investigated the data sharing practices of individual disciplines and around infrastructures. To the authors' best knowledge, no systematic studies have fully examined data sharing practices in EE, an interdisciplinary, complex, and diverse community with a variety of research activities and dynamic data types and formats (Pejša and Hacker, 2013), and thus great challenges for data sharing. To fill this gap, this study examined the data practices of the EE community, including the typical activities of EE research projects, the types and forms of data generated and used in those activities, the project roles EE researchers play in those activities, and their perceptions of data quality and ownership. This article reports findings on EE researchers' data sharing practices and perceptions of data ownership.

3 Research Method

As a meta-theory to study human behavior, activity theory is often used to deconstruct an activity into six related concepts or components: subject, objective, tools, community, rules, and division of labor (Engeström, 1987; Kaptelinin and Nardi, 2012). Activity theory can also be used as a methodological framework to help formulate research questions, guide methodological decisions, develop research instruments, provide concepts and structure to analyze data, and situate practices in the broader context of activities to offer insight into the interaction and contradiction between different activities (Roos, 2012; Wu, 2014). In this study, activity theory allowed situating EE researchers' data practices in the broader context of their research project activities and deconstructing them into six related components. Using activity theory as an overarching framework, this study employed qualitative semi-structured interviews (Blee and Taylor, 2002) to answer three research questions regarding data sharing:

- What types of data do EE researchers share, with whom, and under what conditions?
- How do EE researchers perceive the ownership of data?
- What are the factors that may influence EE researchers' willingness to share data or not with a particular researcher or institution?

ID	Sex	Highest Degree	Academic Seniority	Research Approach	Data Ownership Perception
S1	F	PhD	Doctoral student	Computational	Research group Research project
S2	F	MA	Doctoral student	Experimental Computational	Funding agency
S3	M	MA	Doctoral student	Experimental Computational	Funding agency
S4	M	PhD	Assistant professor	Experimental Computational	Don't know Research institution PIs Co-PIs Researchers
S5	M	MA	Doctoral student	Experimental Computational	Funding agency Research institution
S6	M	PhD	Postdoctoral researcher	Experimental Computational	Funding agency
S7	M	MA	Doctoral student	Experimental Computational	Funding agency
S8	M	MA	Doctoral student	Experimental Computational	Funding agency
S9	M	MA	Doctoral student	Computational Theoretical	Don't know Funding agency Research institution
S10	F	PhD	Postdoctoral researcher	Computational	Research group
S11	M	MA	Doctoral student	Experimental Theoretical	Don't know
S12	M	MA	Doctoral student	Computational	PIs
S13	F	MA	Doctoral student	Experimental Computational	PIs Co-PIs
S14	F	MA	Doctoral student	Experimental Computational	Research group
S15	M	PhD	Assistant professor	Experimental Computational	Funding agency
S16	M	PhD	Postdoctoral researcher	Experimental Computational	Funding agency

Table 1. Participant demographics

Research began with documentary analysis of the research data, documentation, data sharing and publication guidelines, and other relevant documents preserved at EE's two cyberinfrastructures: NEEShub and DesignSafe. This informed development of an interview instrument used for qualitative semi-structured interviews with 16 EE researchers from five research institutions in the United States about their research project activities and data practices. Only two of those five research institutions house experimental facilities funded by NEES or NEHRI. Two interviewees were assistant professors, three were postdoctoral researchers, and 11 were PhD students (see Table 1). Postdoctoral researchers and PhD students were purposefully

sampled because (a) they self-identified as responsible for data management and curation in their project teams, (b) they usually create documentation playing key roles in data sharing and reuse, and (c) NEES perceives young researchers to be of special importance in archiving data and facilitating data sharing and reuse (Pejša and Hacker, 2013). Interviews, ranging from 25 to 85 minutes, were audio recorded, transcribed, and coded with NVivo 11. The two authors independently coded all interviews using an initial coding scheme based on the literature review, activity theory, and documentary analysis. After comparing, discussing, and resolving any differences in their coding, a new coding scheme was formed with emergent codes and subcategories, and used to recode all interviews. Despite subtle differences existing, the authors found no significant discrepancies; minor discrepancies were resolved through further discussion to obtain agreement (Bradley *et al.*, 2007). This coding and analysis process also informed the development by the two authors of typologies of data in EE and factors influencing EE researchers' willingness to share data. The first author developed initial versions of these that were then refined through further discussion of the typologies and broader coding and analysis with the second author.

4 Findings

4.1 Data to share

All interviewees indicated they had shared data with others or were willing to share, as they believed data sharing allowed for validating findings, reusing data, avoiding repetitive experiments, and realizing a common good. Corresponding to EE researchers' research project activities reported in Wu *et al.* (2016), this study developed a typology of data in EE, first seen in Wu *et al.* (2016) and revised in Table 2 with one additional data type, field data. The data can further be classified by state as raw data, processed data, analyzed data, verified data, certified data, archived data, and published data. Certified data are particularly the data meeting the curation criteria set forth by NEES (Pejša and Hacker, 2013) or NHERI (DesignSafe-CI, n.d.a). Archived data are those accepted to the NEEShub or Data Depot. Forms of data produced and used in EE are diverse, including but not limited to data in text (ASCII) format captured by the data acquisition systems, images, videos, audio recordings, 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.

4.1.1 Types of data to share

4.1.1.1 Experimental data

In considering the sharing of these different types of data in EE (see Table 2), researchers are more willing to share experimental data (e.g., sensor measurements, videos) than computational data (e.g., simulation models) and documentation. This may be partially due to the previous NSF/NEES data-sharing policy requiring submission of experimental data to NEEShub within six months after the end of an experiment (Pejša and Hacker, 2013). Although NSF replaced NEES with NHERI recently, one interviewee explained the data-sharing policy remains similar to that of NEES:

> For every project that is funded through the NHERI program, the policy remains very similar to NEES. So you need to share your experimental data. But there is a curation period for researchers to have the privilege to use the data first, usually [for] one year or sometimes [it] can be two years. (S16)

Data Types	Data
Experimental data	Sensor measurements, test recording videos, test recording images
Computational data	Simulation models, software, programming code, statistics data, simulation results
Field data	Infrastructure performance data, remote sensing data, field observations, photos, videos, human experiential data (e.g., interview data), reports
Documentation	Grant proposals, project executive summaries, specimen design drawings, specimen structural plans, construction drawings, construction summaries, constructing recordings, instrumentation plans, sensor metadata, experiment notes, experimental setup reports, meeting minutes, project reports
Test specimen	Buildings, columns, walls, bridges or bridge components, 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, conference proceedings, theses
Presentations	Conference presentations, presentations within the project team/research group
Communication data	Emails

Table 2. Types of data corresponding to the EE research project activities

In relation to the NSF policies and their impact on them as researchers, seven interviewees were working or had worked on NSF funded projects: six affirmed the value and benefits of NSF's data-sharing requirement, but one expressed ambivalence:

> … [the NSF requirement] is really good if you want to share it. But then there are also competing interests. So if you're not required to share, why should you share it? (S4)

4.1.1.2 Computational data

Some EE research focuses on empirical, hands-on experiments and field investigation; other projects focus on computations, simulations, and theories, generating computational data to predict results. Interviewees who conducted both experimental and computational research were inclined to keep computational data to themselves and not share them with others, at least before publication. One admitted:

> For the experimental data you have to do [share] it. It's been required [by NSF]. So that's clear… But for my numerical data, that's kind of something very vague. You don't have any rules on that. So my personal preference is I don't share it, unless I'm being asked. Even if asked, I would prefer to share it after things are published. (S7)

A purely computational researcher (S12) indicated he had shared "some generated figures, time history [data], and some internal forces of the structure" with people outside of his project team (e.g., visiting scholars, researchers from other universities). However, when asked about the simulation models he developed, he was cautious of sharing them even with colleagues, expressing concern on losing competitive advantages:

> [For] the code [simulation models] that I generated, I share partially with my colleagues. And normally if my advisor didn't ask, I don't share it with the guys from other universities or visiting scholars, because maybe that's what contains something confidential. That's my own concern…preventing others from stealing ideas. (S12)

When asked why they withheld simulation models, another interviewee who conducted both experimental and computational research explained in his view it would be meaningless for others to reuse the data without participating in the experiment and having the publications to describe the experimental process:

> You don't want to publish your data right away after a test, because at that time they are still raw data. And you are the only one knows what that is. So it's meaningless to share data at that time. And you want to publish the paper before you really share the data. (S3)

Compared to their caution in sharing simulation models, EE researchers were relatively more flexible with sharing simulation results. One interviewee indicated willingness to share simulation results within the research group, but not the simulation models he developed:

> For simulation data I won't share it with anyone…because I might make mistakes. I want to further improve it, unless I get to a level where I'm very confident. So I don't provide my models, unless I publish them. But the result, if they want to use it, yes, maybe in the small scale, within the group. (S7)

Computational researchers not conducting experiments expressed a willingness to share software and programming code they developed for analyzing data, despite NSF not requiring such sharing. For example, one purely computational researcher's data sharing was due to her advisor's practice of intellectual generosity:

> In our group we have [developed] our own Java software to do the analysis. And some of the software is open to the public. But this is our advisor's decision. He wants other people to use our software…it's good for teaching, for students. (S1)

4.1.1.3 Documentation

Along with the potential sharing of experimental and computational data, successful data sharing may also require documentation that can provide details of how the data were generated and collected, processed, analyzed, and used for a given project. In terms of this documentation, especially internal documentation, one interviewee clearly indicated it would not be shared outside the project team, since funding agencies did not require it and concerns over taking responsibility for any issues with reuse:

> … [I]n order to keep consistency for our future publications, I prepared a summary of the building response data like a table for people to use… But this is just within our research team. We won't release it to the general public… We don't want to be responsible for it when they have a potential problem… The documentation is not required by NSF. (S7)

4.1.1.4 Field data

Unlike experimental data, field data are those collected outside of laboratories and experimental facilities, usually after disasters and using different research methods (e.g., surveys, interviews, observations, reviewing documents and records) and tools (e.g., mobile devices, remote sensors, drones, lidar). Three interviewees reported the research project activity of visiting field sites to collect field data after earthquakes. One was a purely computational researcher who had

collected interview data from stakeholders of a hospital to learn the redundancy of the hospital's supply chain system after earthquakes, to inform a healthcare supply chain model for the hospital. She revealed her project team had shared all the data publicly, except for interview data:

> We haven't shared the interview data with other people. But for other data, it's all publicly available…There are some policies about the interview [data]. It's about the people [human subjects], so there are a lot of rules out there. (S10)

4.1.2 States of data to share

When researchers were willing to share some data, it did not mean that they would share data from all stages of the project, with distinctions drawn between raw data, processed data, partially analyzed data, and fully analyzed data ready to be published. In terms of the state of data to share, interviewees' perceptions varied. One interviewee (S5) expressed the preference of sharing partially analyzed data; two others preferred sharing raw data since others may reproduce and validate the original findings or reuse raw data in their own way. One explained:

> It's best to provide raw data…[so] that people can manipulate and easily reproduce the results that you get, but not necessarily just give them the results, because if they don't agree with the way you processed it, they won't be able to reprocess it. (S2)

However, when speaking of videos and images, this interviewee preferred providing others with more usable "processed" data:

> I've shared a lot of video data, mostly processed videos, because people don't necessarily want to process videos themselves… But it's a kind of raw data, because I'm not manipulating it. I've made it in a format where you can view the video, rather than the format that comes out from the camera. (S2)

Another interviewee only shared published data in repositories (e.g., NEEShub, Data Depot) or journals, because he considered the quality check in repositories or through peer review a guarantee of providing others with credible data. From his perspective, unpublished data may not be appropriate for others to use:

> For data not published, we won't share it actually. Because when you share data, you're supposed to have something approved [or] published in my opinion. Otherwise, if you do not publish [data], that's only for personal use, not for other people. (S6)

4.2 When to share data and with whom

When asked under what conditions they would share data, interviewees stated *whom* they would share data with was an important factor. Most interviewees found it necessary to share data with people in the same research group, collaborators, and even sponsors or industry partners *before* publication to collaboratively process and analyze the data, verify findings, and control data quality. One interviewee commented:

> You should share data with others [of the project team] if your research is collaborative. There is no reason you hide data from everyone [in the team]. The consequence is, when people make mistakes or misinterpret the data, they publish those things, and they might compromise your research somehow. (S7)

This researcher shared data within the project team to ensure the quality of team publications. Although he perceived the necessity of sharing data within the project team, he made an

exception for his simulation models (see above), which he would not share with anyone, even people within the same project team. This researcher, participating in a large-scale cross-institutional collaboration, revealed an incident where a collaborator disclosed the simulation models he developed in a conference too early and without consent:

> Sometimes if you don't want to release anything [simulation models], but somehow your collaborator presented it somewhere else [at a conference]… There is no consensus [on when to release the data]. (S7)

If people outside the project team requested data, interviewees indicated a preference for sharing data only *after* publication, to protect their ability to publish and maintain competitive advantages:

> We do sometimes not share the data with specific researchers. The reason is we feel these people want to copy or reproduce our work before we publish it… That happened before. If this is after publication, we will do that. (S10)

Besides sharing after publication, one computational researcher indicated openness to data sharing after graduation to improve data quality:

> I think after I graduate or after the publications are out, it's fine for me to share. And it's good to share with others. Probably they can help us find out some bugs or mistakes in the data. (S9)

The common exception to sharing data outside the project team before publication is friends or colleagues whom EE researchers trust and have no conflicts of interest. One interviewee claimed he had shared data before publication with a peer who had no intentions to publish or misuse the data:

> I've been sharing a lot of things with this guy [his colleague]… with my peers it's not a problem like sharing data, given the fact that I know that they won't go ahead and publish my data without my name. (S5)

One computational researcher mentioned an experience of sharing field data before publication outside of the project team due to reciprocity:

> We went to Mexico together. We collected different datasets, and shared them as the equipment we bought to collect data are different… We wanted to put the data together to see the differences. Another reason [for data sharing with researchers from other universities] is some of the structures are really big. We didn't have enough sensors to install the whole building. (S13)

4.3 Perceptions of data ownership

When asked about the ownership of data produced from their research, interviewees expressed much uncertainty or vagueness. For example, one principal investigator interviewed admitted:

> That's actually a very good question. And I would like to know the answer to that myself. In my case, I believe the data is somewhat owned by the institution that won the grant along with full control by the PIs or the Co-PIs …a student use I think it's OK if it's the data you collected and processed. (S4)

Another interviewee who was also a professor (S15) showed the same uncertainty, and guessed the funding agency owned the data. One purely computational researcher (S10) involved in

collecting interview data perceived the research group owned most of the data the group generated. However, she specifically pointed out the interview data were collectively owned by the researchers who conducted the interviews and the interviewees.

Three interviewees clearly indicated they had no idea who owned the data; one guessed they belonged to the research institution, PIs, co-PIs, and the researchers (e.g., students) who collected and processed the data; another guessed the funding agency and the research institution (see Table 1). Eight interviewees believed their data were owned by the funding agencies; three of these eight considered them a public asset because the funding agency (i.e., NSF) had released them to the public. Three other interviewees thought data were owned by the research group or project, while two assumed they belonged to PIs and/or co-PIs. Finally, two interviewees also mentioned the copyright or ownership of data published as part of journal articles was transferred to the journal publishers.

4.4 Data sharing factors

This study identified 29 factors influencing EE researchers' willingness to share data with a particular researcher or institution, categorized into five groups as follows:

- *Internal factors*: Intellectual generosity
- *External factors*: Requirements from funding agencies, requirements from journals, requirements from sponsors or industry partners, requirements from property owners, competing interests, complementary knowledge or skills, data recipients' reputation, reciprocity, the size or state of data
- *Future outcomes*: Receiving citation, co-authorship, expectation of future collaboration, awareness of how data will be used by recipients, early release of data by collaborators, misuse of data, commercial use of data, protecting future publications, maintaining competitive advantages, protecting human subject's privacy
- *Purposes for sharing*: Education, validating data or findings, ensuring or improving data quality, studying a new problem
- *Social and organizational ties*: Academic genealogy, friendship or familiarity, people within the same research group, people of the same institution, current or past collaborators

Interviewees working currently or previously on NSF funded projects all mentioned the external requirements from the funding agency. All had or would share experimental data via NEEShub or Data Depot due to the NSF requirement to do so. One experimental and computational researcher stressed the importance of NSF's data-sharing requirement:

> Personally I think that it's important to share data… But the problem with that is if there is no requirement by the funding agency, most of the people won't do it, actually. (S5)

Another frequently mentioned data-sharing factor was academic genealogy. Seven doctoral students were less clear on data ownership, and relied on their advisors, PIs, and project team to make a decision when asked to share:

> …[data sharing] is not really up to the students. It's more up to the advisors. So I think the advisors have their own opinions and the students just kind of go along with what the advisors think. Even if we did have an opinion, it's not really up to us. (S2)

5 Discussion

5.1 Data sharing practices

The 29 data sharing factors identified in this study were categorized into five groups: internal factors, external factors, future outcomes, purposes for sharing, and social and organizational ties. The interviews suggest that without NSF's data-sharing requirements, EE researchers may be less likely to share experimental data before publication. This corresponds to the findings of Birnholtz and Bietz (2003) that EE researchers were more willing to share abstractions of experimental data than the full dataset; maintaining control is important. NSF may continue to enforce this data-sharing mandate, and encourage other (non-NSF) research projects to share data in Data Depot and provide them with data curation support or consultation. Barriers to data sharing seen in the literature (Borgman *et al.*, 2007; Campbell *et al.*, 2002; Tenopir *et al.*, 2011; Tenopir *et al.*, 2015; Tenopir *et al.*, 2018) were also present here, including concerns over potential misuse or commercial reuse, losing competitive advantages, needing to protect future publications, ownership and rights concerns, risks associated with violating the confidentiality of human subjects, and the lack of external requirements to share. Enablers included those found in Tenopir *et al.* (2011) and Borgman *et al.* (2007): receiving citations, the potential for future collaborations, co-authorship, and reciprocity. EE researchers also consider social and organizational ties as key enablers for data sharing. Similar to the findings of Cragin *et al.* (2010) and Wallis *et al.* (2013), EE researchers may share data with researchers outside of the project team who are immediate or known colleagues, current or past collaborators, and well-known people in the field, due to friendship, trust, complementary skills or knowledge, and/or anticipated reciprocity.

Unlike some other scientific disciplines, the EE community has strict data-sharing policies established from NSF specifying the types, formats, and quality of data to share and when to share them. These data-sharing policies impact not only whether EE researchers share data or not, but also the types and state of data they share as well as when they share. Policies and norms intersecting with the discipline (Borgman, 2012; Kowalczyk and Shankar, 2011; Van House, 2003), such as those of NSF, have greatly influenced their practices, with documented policies and norms serving as a form of boundary object (Star and Griesemer, 1989). Interviewees complied with NSF's requirements to share raw experimental data, but most did not share computational data and internal documentation outside of the project team since they were not specifically required. Two interviewees would not even share all of their simulation models with collaborators because of concerns about them releasing the data early. This is likely contradictory to the standard practice of sharing within the project to have collaborators process, analyze, and verify the data to ensure quality (Cragin *et al.*, 2010). EE researchers perceive their computational data, especially simulation models, as their key asset, keeping them close to their chest to ensure their ability to publish. Considering the key role documentation plays in data reuse (Faniel and Jacobsen, 2010) and the various types and forms of documentation existing in EE (Wu *et al.*, 2016), NSF or NHERI may consider specifying more types of documentation (in addition to project reports) that should be shared in Data Depot and providing project teams with financial support, if possible, to create and share documentation. Such documentation and policies, as successful boundary objects (Star and Griesemer, 1989), should have common structure and standardization across different research teams, research communities, and funding agencies (Zimmerman, 2008), while allowing for sufficient flexibility for these teams to conduct their own research and make their own interpretations of findings. The authors believe such a view acknowledges that documentation, policies, and other aspects of the sociotechnical

infrastructure for data sharing are often social constructions, "subject to … local tailoring" (Star, 2010, p. 603), even as NSF, NHERI, and other organizations may exert more power and influence over common, standardized elements of the infrastructure in this particular case. The success of such infrastructure—and of boundary objects that are part of it—relies on its social construction successfully matching local processes, needs, and activities (Star and Ruhleder, 1996; Van House, 2003), not solely on firm mandates from NSF or NHERI that ignore teams' and researchers' need for flexibility.

Academic genealogy is an important factor in both whether students' data are shared and in their learning of data sharing practices within the EE community. They legitimately, but peripherally, participate in the collection and management of data of varying types, working closely with senior experimental, computational, and theoretical researchers. This enables their indoctrination into a community of practice (Lave and Wenger, 1991), as they are given "a real opportunity to act as part of the [EE] community" (Birnholtz and Bietz, 2003, p. 344) by being an active part of a research project. Their advisor or the project PI still hold many of the keys to how and when data are shared and to determining data ownership. New students may not be given much opportunity to participate in existing data practices by more experienced students, postdoctoral researchers, and PIs (Birnholtz and Bietz, 2003), despite their key roles in data management in many projects in EE and elsewhere. Even experienced researchers do not always understand who owns the data they are working with or whom they should be sharing them with. Advisors and PIs should consider working more closely with students in both education and practices surrounding data sharing, ownership, and management, given students self-identify in these roles and are of special importance to NSF in project teams (Pejša and Hacker, 2013). Giving new students the chance to at least observe such practices may also help them get up to speed and participate more legitimately and centrally over time. NSF and other funding agencies should provide encouragement towards this closer involvement. These students can then better serve as potential boundary spanners for data sharing and management practices and, as their careers develop, as legitimate and informed gatekeepers of data in research projects. As suggested by Fecher *et al.* (2015), data sharing should be integrated in the curriculum for university students. Since NHERI's experimental facilities are located in eight universities across the United States (DesignSafe-CI, n.d.b), NHERI may consider collaborating with the libraries of those universities to provide EE students with educational opportunities to learn the proper practices of data sharing, citation, and curation. Graduate schools in library and information science are increasingly offering courses in data management and curation; two of the eight universities housing NHERI's experimental facilities have a library and information school (the University of Texas at Austin and the University of Washington). This creates more opportunities in further curriculum development and collaboration.

As indicated above, NSF's data-sharing policy had an influence on the data practices of the seven interviewees currently or previously working on NSF-funded projects. NSF also influenced data practices across most of our interviewees, even those not currently or previously working on NSF-funded projects, through their provision of the sociotechnical infrastructure of NEES (and NEEShub) and NHERI (and DesignSafe), and the guidelines, rules, and expectations set out for data sharing practices and research project activities, which given NSF's central role in natural hazards engineering research have influence beyond NSF-funded projects and researchers. The current study further explored EE researchers' responses to NSF's data-sharing requirements in comparison to previous studies (e.g., Birnholtz and Bietz, 2003; Faniel and Jacobsen, 2010). Most of those interviewees perceived the importance and value of sharing their

data for the good of others and themselves, but they expressed disagreement with some specific NSF requirements. To attain competitive advantages, EE researchers must publish papers within specific timeframe before their data became public. Three interviewees commented they needed more time to process, analyze, and document the data and to write and publish papers based on them. One pointed out NSF's data-sharing policy had limited her project team's ability to publish papers before their data were public:

> There is a lot of still unpublished work with the project, and the data's public right now… There's a time limitation where we had to share the data before a certain time. But it was the same [time] limitation for all sizes of projects… And since our project had so much data, I think that maybe the time limit should have been extended just to allow the students to publish, because after all of the data was organized there was no time to publish. (S2)

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 of an activity, between components of the activity, and between different but interconnected activities (Engeström, 1987). The interviewee above (S2) reported a contradiction between her activities of data curation and writing articles. To resolve this contradiction, NSF may allow EE researchers to adjust the timeframe they can withhold data based on the project size and the amount of experimental data collected. This may provide larger projects with sufficient time to process, analyze, and curate their data and better protect their ability to publish papers. As suggested by Fecher *et al.* (2015), funding agencies may provide EE researchers or their institutions with financial compensation for data management, documentation, and curation.

Besides the time limit to withhold data, the types of data required to share by NSF may need reconsideration. One computational researcher perceived little value in sharing and reusing only the raw experimental data in NEEShub (now Data Depot):

> … the NSF requirement, which says that you have to upload the raw data, personally to me doesn't make a lot of sense, because I don't think that anybody would ever download the raw data and go through the entire analysis that anybody did in their PhD just to check whether it's correct or not. What people are looking for is data already analyzed to some extent… I think it should be a requirement to upload the analyzed data. (S5)

To validate the simulation models they built, computational researchers need to reuse other researchers' experimental data to run models and compare the results. However, neither the raw experimental data preserved in NEEShub (now Data Depot) nor the analyzed data shared in publications or reports were in a state or format that could meet the requirement for reusing the data to validate models. A contradiction existed between the objective of this activity and the tools mediating it. The data and documentation did not have sufficient common structure or standardization, as a boundary object (Star and Griesemer, 1989), for successful reuse across communities (cf. Zimmerman, 2008). Since the minimum data-sharing requirement from NSF was to upload and share raw experimental data with documentation (e.g., reports) (Pejša and Hacker, 2013), EE researchers might not bother to share analyzed data. To improve the usability of data in Data Depot and other repositories, funding agencies and journals may consider requiring or encouraging EE researchers to share analyzed data in a more reusable state, one that can help facilitate more successful sharing and reuse across communities, to better support the model validation activity of computational researchers. Greater cognizance of the translation of

knowledge analyzed data and related documentation can support, via other policy and educational initiatives suggested above, should also help better support these researchers' activities.

One can group the motivations that may stimulate researchers to share data as intrinsic and extrinsic motivations (Ryan and Deci, 2000). Intrinsic motivations are internal and self-determined in finding gratification or joy in the activities one performs. On the other hand, extrinsic motivations are externally induced through external rewards or punishments. Although numerous data-sharing incidents identified in the current study were spurred by extrinsic motivations (e.g., requirements from funding agencies, industry partners, and advisors / PIs), one interviewee still brought up a data-sharing incident associated with an intrinsic motivation, intellectual generosity. This corresponds to previous findings that both extrinsic and intrinsic (e.g., altruistic) motivations have an influence on employees' knowledge-sharing intentions (Lin, 2007). Kim and Stanton (2016) showed both normative pressure at a discipline level and scholarly altruism had significant positive relationships with STEM researchers' data sharing. However, the current study did not find any community or disciplinary norms regarding data sharing in EE. EE researchers may not bother to share data without the requirements from funding agencies. To motivate data sharing by EE researchers, NHERI may promote the norms of altruism and reciprocity to the EE community, and establish an honor system to encourage contributions to the community beyond their personal gain.

5.2 Data ownership

This study found EE researchers—including those playing the key role as PI or project lead—were uncertain of data ownership and their perceptions varied. This vagueness and confusion in data ownership may hinder EE researchers' intentions towards sharing data even though they may have positive views of data sharing, as they are unsure of who has the final say. As indicated above, computational data (e.g., simulation models, programming code), analyzed data, internal documentation, and the data produced by non-NSF funded projects are not required to be shared, and open to researchers' judgment. EE researchers perceive computational data as having competitive advantages and value for others to reuse. Without clarification and definition of data ownership, EE researchers may be more inclined to withhold computational data of value for others to reuse and the data not required by funding agencies. Research institutions and funding agencies should consider specifying data ownership in the grant requirements or asking applicants to indicate data ownership in their data management plans and reports; and providing researchers, especially junior researchers, with training to increase their awareness of data ownership.

6 Conclusion

This study examined the data sharing practices of the EE community based on qualitative semi-structured interviews, uncovering the types and states of data EE researchers share, with whom, and under what conditions; and their perceptions of data ownership. Based on the findings, there are clear implications and suggestions for funding agencies, NHERI, and research institutions regarding data management and curation, such as allowing researchers to adjust the timeframe they can withhold data based on their project size and the amount of experimental data generated; expanding the types and states of data required to share; defining data ownership in the grant requirements; and providing EE researchers with training on data ownership, sharing, citation, and curation. The data sharing factors identified in this study can provide different stakeholders of the EE community—including but not limited to funding agencies, repositories,

databases, journal publishers, data curators, and data users—with new insights into the enablers and motivations for data sharing.

This study is limited in that most interviewees were postdoctoral researchers and doctoral students. More interviews should be conducted with researchers holding other academic or research positions (e.g., professors, PIs, lab managers, curators) to gain different perspectives. Future research includes developing and implementing a survey of EE and other natural hazards engineering researchers regarding their data practices.

Acknowledgement

The authors would like to express their gratitude to Dr. Besiki Stvilia, Dr. Roberta Brody, and Dr. Kwong-Bor Ng for their helpful suggestions, and to the reviewers for useful feedback.

References

Allen, D., Karanasios, S. and Slavova, M. (2011), "Working with activity theory: Context, technology, and information behavior", *Journal of the American Society for Information Science and Technology,* Vol. 62 No. 4, pp. 776-788.

Birnholtz, J. P. and Bietz, M. J. (2003), "Data at work: Supporting sharing in science and engineering", in *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work*, ACM, New York, NY, pp. 339-348.

Blee, K. M. and Taylor, V. (2002), "Semi-structured interviewing in social movement research", in B. Klandermans & S. Staggenborg (Eds.), *Methods of Social Movement Research*, University of Minnesota Press, Minneapolis, MN, pp. 92-117.

Blumenthal, D., Campbell, E. G., Anderson, M. S., Causino, N. and Louis, K. S. (1997), "Withholding research results in academic life science", *Journal of the American Medical Association,* Vol. 277 No. 15, pp. 1224-1228.

Borgman, C. L. (2012), "The conundrum of sharing research data", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 6, pp. 1059-1078.

Borgman, C. L., Wallis, J. C. and Enyedy, N. (2007), "Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries", *International Journal on Digital Libraries*, Vol. 7 No. 1-2, pp. 17-30.

Bradley, E. H., Curry, L. A. and Devers, K. J. (2007), "Qualitative data analysis for health services research: Developing taxonomy, themes and theory", *Health Services Research*, Vol. 42 No. 4, pp. 1758-1772.

Campbell, E. G., Clarridge, B. R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A. and Blumenthal, D. (2002), "Data withholding in academic genetics", *Journal of the American Medical Association*, Vol. 287 No. 4, pp. 473-480.

Cragin, M. H., Palmer, C. L. and Carlson, J. R. (2010), "Data sharing, small science and institutional repositories", *Philosophical Transactions of the Royal Society*, No. 368, pp. 4023-4038.

DesignSafe-CI (n.d.a), "Data publication guidelines: Guidance and best practices for publishing data", available at: https://www.designsafe-ci.org/rw/user-guides/data-publication-guidelines/ (accessed 20 February 2019).

DesignSafe-CI (n.d.b), "Experimental facilities", available at: https://www.designsafe-ci.org/facilities/experimental/ (accessed 20 February 2019).

Engeström, Y. (1987), *Learning by Expanding: An Activity-Theoretical Approach to Developmental Research*, Orienta-Konsultit Oy, Helsinki.

Faniel, I. M. and Jacobsen, T. E. (2010), "Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data", *Computer Supported Cooperative Work,* Vol. 19 No. 3, pp. 355-375.

Faniel, I. M. and Zimmerman, A. (2011), "Beyond the data deluge: A research agenda for large-scale data sharing and re-use", *International Journal of Digital Curation,* Vol. 6 No. 1, pp. 58-69.

Fecher, B., Friesike, S. and Hebing, M. (2015), "What drives academic data sharing?", *PLoS ONE,* Vol. 10 No. 2, e0118053.

Kaptelinin, V. and Nardi, B. (2012), "Activity Theory in HCI: Fundamentals and reflections", *Synthesis Lectures Human-Centered Informatics*, Vol. 5 No. 1, pp. 1-105.

Kim, Y. and Stanton, J. M. (2016), "Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis", *Journal of the Association for Information Science and Technology*, Vol. 67 No. 4, pp. 776-799.

King, G. (2007), "An introduction to the Dataverse Network as an infrastructure for data sharing", *Sociological Methods & Research,* Vol. 36 No. 2, pp. 173-199.

Kowalczyk, S. and Shankar, K. (2011), "Data sharing in the sciences", *Annual Review of Information Science and Technology,* Vol. 45, pp. 247-294.

Lave, J. and Wenger, E. (1991), *Situated Learning: Legitimate Peripheral Participation*, Cambridge University Press Cambridge, UK.

Lin, H. F. (2007), "Effects of extrinsic and intrinsic motivation on employee knowledge sharing intentions", *Journal of Information Science*, Vol. 33 No. 2, pp. 135-149.

National Endowment for the Humanities (2019*),* "Data Management Plans for NEH Office of Digital Humanities Proposals and Awards", available at https://www.neh.gov/sites/default/files/inline-files/Data%20Management%20Plans%2C%202019.pdf (accessed 20 February 2019).

National Institutes of Health (2003), "NIH data sharing policy and implementation guidance", available at http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm (accessed 20 February 2019).

National Science Board (2005), "Long-lived digital data collections: Enabling research and education in the 21st century", available at http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf (accessed 20 February 2019).

National Science Foundation (2011), "NSF data management plan requirements", available at http://www.nsf.gov/eng/general/dmp.jsp (accessed 20 February 2019).

Park, H., You, S. and Wolfram, D. (2018), "Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields", *Journal of the Association for Information Science and Technology,* Vol. 69 No. 11, pp. 1346-1354.

Pejša, S., Dyke, S. J. and Hacker, T. (2014), "Building infrastructure for preservation and publication of earthquake engineering research data", *International Journal of Digital Curation*, Vol. 9, No. 2, pp. 83-97.

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

Rathje, E. M., Dawson, C., Padgett, J. E., Pinelli, J., Stanzione, D., Adair, A., … Mosqueda, G. (2017), "DesignSafe: New cyberinfrastructure for natural hazards engineering", *Natural Hazards Review*, Vol. 18 No. 3.

Roos, A. (2012), Activity theory as a theoretical framework in the study of information practices in molecular medicine. *Information Research*, Vol. 17 No. 3, available at http://www.informationr.net/ir/17-3/paper526.html (accessed 22 May 2019).

Ryan, R. M. and Deci, E. L. (2000), "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being", *American Psychologist*, Vol. 55 No. 1, pp. 68-78.

Star, S. L. and Griesemer, J. R. (1989), "Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39", *Social Studies of Science,* Vol. 19 No. 3, pp. 387-420.

Tenopir, C., Allard, S., Douglass, K. L., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M. and Frame, M. (2011), "Data sharing by scientists: Practices and perceptions", *PLoS ONE,* Vol. 6 No. 6, e21101.

Tenopir, C., Christian, L., Allard, S. and Borycz, J. (2018), "Research data sharing: Practices and attitudes of geophysicists", *Earth and Space Science,* Vol. 5, pp. 891-902.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D. and Dorsett, K. (2015), "Changes in data sharing and data reuse practices and perceptions among scientists worldwide", *PLoS ONE,* Vol. 10 No. 8, e0134826.

Van House, N. A. (2003), "Science and technology studies and information studies", *Annual Review of Information Science and Technology,* Vol. *38*, pp. 3-86.

Wallis, J. C., Rolando, E. and Borgman, C. L. (2013), "If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology", *PLoS ONE,* Vol. 8 No. 7, e67332.

Witt, M., Carlson, J., Brandt, D. S. and Cragin, M. H. (2009), "Constructing data curation profiles", *The International Journal of Digital Curation,* Vol. 4 No. 3, pp. 93-103.

Wu, S. (2014), *Exploring the data work organization of the Gene Ontology*. Ph.D. Thesis. Florida State University. Available at http://purl.flvc.org/fsu/fd/FSU_migr_etd-9267 (accessed 23 May 2019).

Wu, S., Worrall, A. and Stvilia, B. (2016), "Exploring data practices of the earthquake engineering community", In D. Fenske & J. Greenberg (Co-Chairs), *iConference 2016 Proceedings*, iSchools, Champaign, IL. Available at https://doi.org/10.9776/16187 (accessed 23 May 2019).

Zimmerman, A. S. (2008), "New knowledge from old data: The role of standards in the sharing and reuse of ecological data", *Science, Technology, & Human Values,* Vol. 33 No. 5, pp. 631-652.