Title
Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators

Author
Mayernik, Matthew S

Publication Date
2011

DOI
10.2139/ssrn.2042653

UNIVERSITY OF CALIFORNIA

Los Angeles

Metadata Realities for Cyberinfrastructure:

Data Authors as Metadata Creators

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Information Studies

by

Matthew Stephen Mayernik

2011

The dissertation of Matthew Stephen Mayernik is approved.

_____
Gregory H. Leazer

_____
Ramesh Srinivasan

_____
Sharon J. Traweek

_____
Christine L. Borgman, Committee Chair

University of California, Los Angeles

2011

TABLE OF CONTENTS

FULL TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I have many people to thank for their contributions to this work. First and foremost is my advisor Christine Borgman. Chris introduced me to the world of cyberinfrastructure and data practices research, and has been an incredible source of help, support, expertise, and encouragement during my time at UCLA. Her amazing ability to synthesize complex issues was invaluable from the beginning stages of my dissertation work to the end. I could not have asked for a better example of academic leadership, scholarly excellence, and student mentoring. I am also indebted to my dissertation committee: Greg Leazer, Ramesh Srinivasan, and Sharon Traweek. They continually pushed me to read more widely, think more deeply, and write more clearly. My dissertation would not have taken shape as it has without their ability to give insightful feedback, particularly on the short notice that I often required.

Many thanks also to the many fellow students I have had the opportunity to work with as a member of the CENS data practices team: Dave Fearon, Andrew Lau, Alberto Pepe, Katie Shilton, Jillian Wallis, and Laura Wynholds. Most of the ideas in my dissertation grew out of my work within the data practices team, and I cannot thank them enough for their many discussions and close readings of my/our papers. Extra special thanks to Jillian Wallis, whose imprint can be seen throughout this dissertation. Jillian shared her office, her interviews, and her field sites. She spearheaded my first CENS project, the CENS Deployment Center, as well as our participation in development of the CENS Annual Reporting system. She also co-developed and administered the CENS Metadata Registry user tests reported in Chapter 7. Thanks also to Michael Wartenbe and

Amelia Acker for coming to my practice presentations and providing valuable outside perspectives on my work.

I also want to thank the fine folks from the Monitoring, Modeling, & Memory (MMM) project, particularly the Metadata Friction group: Paul Edwards, Archer Batcheller, Geof Bowker, and Chris Borgman. The MMM meetings and phone calls gave me many opportunities to develop ideas, and helped me to situate my own work within larger scholarly and institutional contexts.

My ultimate appreciation and gratitude go to the members of CENS, the Center for Embedded Networked Sensing. I cannot thank enough the scientists and engineers who brought me in to their labs and out to their field sites, who sat down for my interviews, and who participated in our user tests. I also want to thank the CENS leadership and administrative staff, particularly Deborah Estrin, the Director of CENS, for creating an environment that enabled our social scientific projects to flourish, and Jeff Goldman, the Administrative and Program Development Director, for allowing us to piggy-back our Metadata Registry on the CENS Annual Reporting system. I also want to acknowledge the UCLA Academic Technology Services (ATS), who provided technical development and support for both the CENS Deployment Center project and the CENS Metadata Registry, without which my projects would not have been possible.

Finally, I thank my family for supporting me throughout my many years of schooling, and for always being excited by and interested in everything that I do. And to my wife Becca, who I met at the end of my first term at UCLA. We haven't been apart since, and I can't imagine it any other way…

VITA

1981	Born, Lewistown, Montana, United States of America.
2003	B.S. Engineering & Applied Science, Caltech, Pasadena, California.
2006	Reference Desk Assistant, Science and Engineering Library, University of California, Los Angeles (UCLA).
2007	Master of Library and Information Science, UCLA.
2009, 2010	Special Reader, IS 260: Information Structures, Department of Information Studies, UCLA.
2006-2011	Graduate Student Researcher, Center for Embedded Networked Sensing, UCLA.

PUBLICATIONS

Edwards, P.N., **Mayernik, M.S.**, Batcheller, A., Borgman, C.L., & Bowker, G.C. (in press). Science friction: data, metadata, and collaboration in the interdisciplinary sciences. *Social Studies of Science.*

Mayernik, M.S., Batcheller, A.L., & Borgman, C.L. (2011). How institutional factors influence the creation of scientific metadata. In *Proceedings of the 2011 iConference* (iConference '11). New York, NY: ACM (pg. 417-425).

Mayernik, M. (2010). The distributions of MARC fields in bibliographic records: a power law analysis. *Library Resources & Technical Services*, 54(1): 40-54.

Mayernik, M.S. (2010). Metadata tensions: A case study of library principles vs. everyday scientific data practices. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1-2.

Mayernik, M.S. (2010). Metadata realities for cyberinfrastructure: data authors as metadata creators. *iConference 2010 Proceedings* (pp. 148-153). Urbana-Champaign, IL: University of Illinois at Urbana-Champaign.

Pepe, A., **Mayernik, M.S.**, Borgman, C.L., & Van de Sompel, H. (2010). From artifacts to aggregations: modeling scientific life cycles on the semantic web. *Journal of the American Society for Information Science and Technology*, 63(3): 567-582.

Wallis, J.C., **Mayernik, M.S.**, Borgman, C.L., & Pepe, A. (2010). Digital libraries for scientific data discovery and reuse: from vision to practical reality. In *Proceedings of the 10th annual joint conference on Digital libraries* (JCDL '10). New York, NY: ACM.

Mayernik, M.S., Wallis, J.C., Pepe, A., & Borgman, C.L. (2008). Whose data do you trust? Integrity issues in the preservation of scientific data. *iConference 2008*. Feb. 28-Mar. 1, Los Angeles, CA.

Wallis, J.C., Borgman, C.L., **Mayernik, M.S.**, & Pepe, A. (2008). Moving archival practices upstream: an exploration of the life cycle of ecological sensing data in collaborative field research. *International Journal of Digital Curation*, 3(1).
URL: http://www.ijdc.net/ijdc/article/view/67/67

Wallis, J.C., Pepe, A., **Mayernik, M.S.**, & Borgman, C.L. (2008). An exploration of the life cycle of eScience collaboratory data. *iConference 2008*. Feb. 28-Mar. 1, Los Angeles, CA.

Borgman, C.L., Wallis, J.C., **Mayernik, M.S.**, & Pepe, A. (2007). Drowning in data: digital library architecture to support scientists' use of embedded sensor networks. *JCDL '07: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, Vancouver, BC*. ACM.

Mayernik, M. (2007). The prevalence of additional electronic features in pure e-journals. *The Journal of Electronic Publishing*, 10(3).
URL: http://hdl.handle.net/2027/spo.3336451.0010.307

Mayernik, M.S., Wallis, J.C., Borgman, C.L., & Pepe, A. (2007). Adding context to content: the CENS Deployment Center. in Andrew Grove (Ed.), *Proceedings of the 70th ASIS&T Annual Meeting, vol. 44, 2007* (pg. 691-698). Silver Spring, Md.: American Society for Information Science and Technology.

Pepe, A., Borgman, C.L., Wallis, J.C., & **Mayernik, M.** (2007). Knitting a fabric of sensor data and literature. in *Information Processing in Sensor Networks. 2007*. Cambridge, MA: ACM/IEEE.

Wallis, J.C., Borgman, C.L., **Mayernik, M.S.**, Pepe, A., Ramanathan, N., & Hansen, M. (2007). Know thy sensor: CENS as a case study of the relationship between data integrity, metadata, and data interpretation. *11th European Conference on Research and Advanced Technology for Digital Libraries, 2007*.

ABSTRACT OF THE DISSERTATION

Metadata Realities for Cyberinfrastructure:

Data Authors as Metadata Creators

by

Matthew Stephen Mayernik

Doctor of Philosophy in Information Studies

University of California, Los Angeles, 2011

Professor Christine L. Borgman, Chair

As digital data creation technologies become more prevalent, data and metadata management are necessary to make data available, usable, sharable, and storable. Researchers in many scientific settings, however, have little experience or expertise in data and metadata management. In this dissertation, I explore the everyday data and metadata management practices of researchers through a multi-sited ethnographic study of metadata creation by researchers in the Center for Embedded Networked Sensing (CENS). In studying metadata practices, I focused on the ways that researchers document, describe, annotate, organize, and manage their data, both for their own use and the use of researchers outside of their project. This study illustrates how researchers

within CENS rarely create documentation that is not directly tied to their own use of their data, and correspondingly, they rarely share data with users from outside of their immediate projects. From these observations, I develop a metadata typology that includes six components, including metadata for: *data identity*, *data characteristics*, *data quality*, *data collection equipment*, *data collection methods*, and *data analysis methods*. I use a framework of accountability to discuss the ways that metadata practices fit within social research settings. Metadata are situated in regimes of mutual accountability in which researchers learn what is important to document, what counts as sufficient documentation, and how documentation practices are to be accounted for in social research settings. Researchers work within social ontologies in which "metadata-for-data-sharing" have very low visibility. As a consequence, when asked to create metadata descriptions of the data for a shared CENS metadata registry, researchers lack specific data users, and thus describe their data for members of their most likely "imagined public:" other researchers with shared research interests and methods. I argue that the cyberinfrastructure vision of wide-spread data sharing is fundamentally mis-aligned with the realities of the day-to-day metadata practices of researchers in small-scale field sciences.

1. INTRODUCTION

In the spring of 2011, the Southern California coast experienced two major fish die-offs. In two separate incidents, hundreds of thousands of fish swam into protected harbors, became trapped, and died en masse. According to local news reports, around 180 tons of fish died during the events. Marine life die-offs on this large of scale are rare, but have been increasing in frequency in the Southern California region. The usual cause of such events is overwhelming blooms of algae or other microorganisms that metabolize all of the oxygen in the water and, at the same time, release neurotoxins that poison other marine species (Glibert, et al., 2005; Anderson, et al., 2006). One of the harbors affected by this particular event had experienced fish die-offs before. In response to the previous fish die-offs, a team of aquatic biologists from the Center for Embedded Networked Sensing (CENS), an interdisciplinary science and technology research center dedicated to developing sensing technologies for scientific and social applications, had installed newly developed sensing equipment. With the help of the new sensing system, the CENS aquatic biologists were studying why algae growths occur on large-scales in order to be able to potentially predict when such events occur.

The two 2011 events, however, were unusual because in neither case was there clear evidence of large algae growths prior to the fish deaths. In looking at the measurements taken by their sensors during the die-off event, CENS aquatic biologists

were able to determine that the fish themselves depleted the oxygen in the water after they swam into the harbor. Any toxins from pre-existing microorganisms may have had a small effect on the fish, but were not the primary cause of the fish deaths. This finding enabled the question about the cause of this event to shift. Typically the main question around fish die-off events is "what caused massive algae growth?" In this case, however, the main question became "what caused these fish to swim into the harbor?" As of this writing, proposed answers to this question ranged from ocean weather to whales and dolphins.

Without the presence of the CENS sensors, unraveling these fish die-off events would likely have been much more difficult. That these events had been captured by sensors at all is an extremely unusual and unique opportunity for the aquatic biology team. As the lead biologist noted in one of the many media interviews he conducted after the event, the multiple sensors installed in the harbor captured unprecedented measurements of a fish die-off taking place. As his team analyzes these measurements into the future, he stated, his team will be able to document this type of event as well as any fish die-off has been documented anywhere in the world.

The monitoring of this fish die-off event is an example of how digital technologies, such as sensor networks in the environmental sciences, social networking tools in the social sciences, and the digitization of cultural artifacts in the humanities, allow researchers to produce digital data with volumes and complexities much greater than were possible in the past. Digital technologies, and the data that they produce, offer tremendous opportunity for researchers in every academic discipline to ask questions that

were previously impossible to study. As the fish die-off example illustrates, some digital technologies enable researchers to study very local phenomena in great detail. Other digital technologies enable the integration of many diverse data streams from numerous small-scale projects in order to conduct synthesis and longitudinal studies of large-scale environmental phenomena such as climate change and species shifts. Similarly, the digital integration of historical texts, photos, and maps from libraries and archives all over the world can enable comparative humanities projects that would otherwise be difficult or impossible to conduct. But while the possibilities of digital data are exciting, they also present tremendous challenges: how to best organize and manage data, how to make data discoverable and accessible to diverse user communities, and how to store and preserve data over the long term.

In the United States, government agencies have heavily promoted and invested in "cyberinfrastructure," a suite of technologies that includes high performance computing, grid and cloud infrastructures, and virtual communication systems, as a means to address the challenges and opportunities of digital data. The motivation for cyberinfrastructure for digital data is to leverage web-based data collection, transmission, and storage techniques to develop "multidisciplinary, well-curated federated collections of scientific data" (Atkins, et al., pg. 7) that become central to everyday scientific practices. The vision of cyberinfrastructure is explicitly predicated on the notion that data collected for one project will be re-used and re-purposed by researchers other projects. An influential 2007 National Science Foundation (NSF) report listed data accessibility and use as one of the guiding principles for NSF investments made between 2006 and 2010: "Science and

engineering data generated with NSF funding will be readily accessible and easily usable, and will be appropriately, responsibly and reliably preserved" (*Cyberinfrastructure Vision for 21st Century Discovery*, 2007, pg. 26).

This vision has yet to become a reality in most scientific settings. The fish die-off events described above took place towards the end of the period in which I was conducting the research detailed in this dissertation: a study of data and metadata management in CENS. I undertook my study motivated by an apparent contradiction between these influential reports issued by government agencies that outline the importance of making data available for re-use by other researchers and previous studies of data management in CENS. Previous studies have illustrated how CENS researchers rarely, if ever, shared their data with researchers from outside of their immediate project teams (Borgman, Wallis, & Enyedy, 2007; Wallis, et al., 2010). When such data sharing did occur, it involved direct person-to-person contact and communication. Cyberinfrastructure systems are intended to facilitate data sharing, with shared data repositories mediating data sharing and use. CENS researchers, particularly those in the environmental sciences, are not aware of any such repositories that are applicable to their data.

In an interview I conducted about four months before the fish die-off events, I asked the lead CENS aquatic biologist about long-term plans for the sensor data that they were collecting in the soon-to-be afflicted harbor. The following passage, which gives his response, acutely illustrates the tension between the vision and the reality of large-scale data sharing systems in his field:

Author: Okay. So I'll just kind of turn to long term plans for the data you guys are collecting. Is there a long term plan? Will they be kept locally? Is there a place that you might submit them?

CENS Aquatic Biologist: I'd love to have a place where we could submit them. At some point, they, depending on how funding goes, they may go into some sort of a national repository. There is an oceanographic data repository. So they have not necessarily shown an interest in these kinds of things. Again, a lot of this is fairly specialized but... And we do and we'll keep track of this, we, not as a part of the CENS program but as a part of the [local institute], we have a long-term monitoring system. It's a monthly sampling that takes place out in the [ocean, off the coast]. And we do have essentially our own repository for that information and [our harbor sensor data] will probably be folded into that because that's the most likely place that a modeler or somebody who in the future might want the information would go to look for it. Long term data sets, for the longest time, long term data sets were poo-poohed as not important, not interesting. Now, of course, they're vitally important because of issues of global climate change. So the country and the world is all of a sudden taken a real liking to long term data sets and therefore, some care in maintaining them.

Author: Of course which is where our research is also coming out of.

CENS A.B.: Exactly. But it is still a lot of that is, it's still ad hoc. I'd hate to say it.

Author: So the... You said the [local institute]. That's like a home built repository?

CENS A.B.: It is just basically our own dataset, that it's not, it hasn't even... The web part of it hasn't been kept up in these years just because funding has

> fluctuated, but it is... It's something that we will keep obviously, and I do think will again gain use in the future because somebody will mine it.

As this passage illustrates, fluctuating funding for data management and a dearth of appropriate data repositories lead to ad hoc approaches to long-term data sets, not the standardized and institutionalized approaches desired and required by cyberinfrastructure systems.

My study is also directly motivated by another finding from earlier studies of CENS data management, in particular, CENS scientists' negligible use of metadata standards. Data, whether temperature readings, census records, or digital photos of paintings, must be supplemented by information about the data itself in order to be useful beyond local settings. This "information about data" is commonly referred to as "metadata." Metadata, the documentation, descriptions, and annotations that researchers create and use to manage, discover, access, use, share, and preserve data resources, are a key component of data storage, sharing, and preservation systems (*Cyberinfrastructure Vision for 21st Century Discovery*, 2007; Lynch, 2008). Cyberinfrastructure systems rely on metadata that are created and stored in standardized formats. Formalized metadata products, such as the Dublin Core Metadata Schema and the Ecological Metadata Language, increase the precision with which data from disparate sources can be combined by establishing common ways in which data are to be described and categorized (Edwards, et al., in press). As Borgman, Wallis, & Enyedy (2007) noted, however, CENS researchers rarely use established metadata standards in their own work. "Also paradoxical is [CENS] researchers' awareness of extant metadata standards for

reconciling, managing, and sharing their data, but their lack of use of such standards. The present dilemma is that few of the participating scientists see a need or ability to use others' data, so they do not request data, they have no need to share their own data, they have no need for data standards, and no standardized data are available" (Borgman, Wallis, & Enyedy, 2007, pg. 27)

If CENS researchers are not documenting their data using standardized metadata forms, how are they documenting their data? This question kick-started my study. As I illustrate in this dissertation, the answer is that CENS researchers document their data in a wide variety of ways. Instead of using formalized metadata products, they create metadata via informal and pragmatic metadata processes, such as one-off Excel spreadsheets and Word documents, which act as lubricants in disjointed, imprecise data practices (Edwards, et al., in press). The following passage picks up from the above passage in which the CENS aquatic biologist indicated that they have no repository to which they are likely to submit their data outside of their local team's database. After asking about long-term plans for data, I asked about long-term plans for metadata related to equipment calibrations. Researchers calibrate sensing equipment by measuring known values in order to establish that equipment are working correctly. Records of calibration are a key type of metadata in most sensor-based research, as they document when, and to what degree, sensor readings might need to be corrected.

> **Author:** Right. And how would you, or would you be concerned about folding in like calibration files, things like that into a database like [the database at the local institute] for long-term preservation?

> **CENS A.B.:** I think, yeah, for the basic oceanographic information what is important is to document how it has been massaged, how it has been processed. Beyond that, there may well be new insights on to how to do that data processing in the future and they would take the information from the state that it was in at that point to the next stage if they are going to do some further processing of it. But I think the big thing there is to document the process that the information has gone, the data has gone through to get to its present state. I think that would be the best thing and that typically is done with datasets so that somebody knows who's fiddled with it beforehand. Bad data just get removed, typically. If you know they're bad then there's no percentage in keeping them. You know, you're basically just running the risk that somebody might use it or think that it's useful.

This passage illustrates how researchers prioritize metadata that focus on "process." In this case, the aquatic biologist describes the importance of knowing how data were "massaged" and "processed," and that such processes are often targeted towards ensuring that data are of high quality. Data, metadata, equipment, people, institutions, and physical objects such as water or soil are inextricably intertwined in academic research. Metadata descriptions of data that gloss over research processes are of little use to future data users because they remove the interconnections that exist among all of the constituents of a given data collection situation.

The two interview passages provided above also indicate how creating and collecting data and metadata for future data users is an even more significant challenge when there is no clear notion of who the future data users might be. Researchers are unlikely to prioritize documentation forms that facilitate long-term data availability so that "somebody" might re-use their data in the future. In this sense, researchers lack what

Christopher Kelty (2008) has called a "recursive public" for data sharing. A "recursive public" is a "public that is vitally concerned with the material and practical maintenance and modification of the technical, legal, practical, and conceptual means of its own existence as a public; it is a collective independent of other forms of constituted power and is capable of speaking to existing forms of power through the production of actually existing alternatives" (pg. 3). Free software, the focus of Kelty's study, has become a recursive public by being organized around building the internet-based free software that makes the community possible. CENS researchers, lacking a recursive public for data sharing, have what could be called an "imagined public" for data sharing, in that they can suggest possible future uses and users of their data. Beyond immediate collaborators, however, the future users that researchers suggest are broad groups, such as "government," "public health," "climatology," or "modelers," as the aquatic biologist suggested in the first interview passage quoted above (Borgman, Wallis, & Enyedy, 2007).

A recursive public for data sharing would require researchers to commit to making cyberinfrastructure technologies and ideologies central to the organization and practice of their scientific activities. As Kelty noted, "the commitment to becoming a recursive public, however, raises unprecedented issues about the nature of quality, reliability, and finality of scientific data and results – questions that will reverberate throughout the sciences as a result" (Kelty, 2008, pg. 304). Metadata are one means to ensure quality, reliability, and finality of data. Because information or data specialists with professional training in data and metadata management are not yet common in

scientific research settings, cyberinfrastructure community data sharing systems must rely on "data authors," that is, researchers who collect original data, to provide documentation for submitted data sets.

The two findings from previous studies of data management in CENS that I have pointed out thus far go hand in hand: 1) researchers rarely share data outside their immediate teams, and 2) do not use standardized forms of metadata. Without standardized forms of metadata, it is difficult to share data with someone outside an immediate team because of the necessary work involved in documenting and communicating critical details about data person-to-person. From the other direction, if little data sharing is taking place, researchers have little incentive to adopt formal metadata standards, which are often quite complex to learn and implement. A colleague of mine, Jillian Wallis, has called this back-and-forth the "vicious cycle of data sharing" (personal communication): without standardized metadata, nobody shares data, and without data sharing, nobody uses metadata standards.

In this dissertation, I examine the metadata half of this "vicious cycle" by studying how data authors create metadata for their everyday work and for a shared community metadata registry. I analyze situations in which working scientists, as data authors, do and do not create metadata, what they understand metadata to be, what problems they encounter related to metadata, and their work practices in performing metadata creation tasks.

Why is my study of data authors as metadata creators important? New technologies, including cyberinfrastructure technologies, do not bring about

transformative changes on their own. Instead, new technologies must fit into existing sets of social institutions (Agre, 1998). Gray, et al., (2005) state that one of the key technical advances crucial for scientific analysis of large data sets is the development of "extensive metadata and metadata standards that will make it easy to discover what data exists, make it easy for people and programs to understand the data, and make it easy to track data lineage" (n.p.). I argue that far from being only a technical problem, metadata creation, use, and management are inextricable from the disciplinary, individual, and institutionalized day-to-day work practices of scientists, technicians, and technology developers.

Through an ethnographic study of the everyday metadata practices of CENS researchers, I illustrate how metadata practices exist in systems of social accountability. Researchers are accountable for metadata that are, and are not, created during everyday research tasks. Researchers are obligated to take responsibility for metadata activities that are commensurate with their research role and expertise. In addition, researchers must be able to account for any inconsistency or incompleteness in their metadata processes and products. When asked to create metadata descriptions of the data for a shared CENS metadata registry, however, the lack of a recursive public for data re-use precludes any similar form of accountability. Without a clear future data user for whom to describe data, researchers describe their data for members of their most likely "imagined public:" other researchers with shared research interests and methods.

This dissertation begins with an introduction to the importance of data and metadata in cyberinfrastructure developments. To motivate my study of data authors as

metadata creators, I then discuss human considerations of the metadata creation process. I then use sociological analysis to illustrate how my study of everyday metadata creation draws from established theoretical and philosophical frameworks in sociology, anthropology, and information studies. This study provides guidance for the development of metadata collection policies, processes, and technological systems for future cyberinfrastructure projects.

1.1 Cyberinfrastructure and eScience

In this section, I delve more fully into the developments surrounding "cyberinfrastructure." Advances in high-performance computing, distributed data systems, and virtual communication systems, collectively called "cyberinfrastructure," are facilitating new kinds of data-intensive science. The notion of "cyberinfrastructure" has existed for less than a decade. The *National Science Foundation Blue-Ribbon Advisory Panel on Cyber-infrastructure, Revolutionizing Science and Engineering through Cyber-infrastructure* (Atkins, et al., 2003) report, commonly known as the "Atkins report," jumpstarted the study and development of cyberinfrastructure. (The first reference to "cyberinfrastructure" in the Web of Science database is Borgman (2002), who referred to a draft version of the Atkins report). The Atkins report outlined the potential capabilities and benefits of the use of cyberinfrastructure, and recommended that the U.S. National Science Foundation undertake a sustained period of investment in this area.

So what exactly is "cyberinfrastructure?" The common notion of "infrastructure" refers to the system of roads, highways, power grids, buildings and water systems that

constitute the physical components of Western industrial society (Friedlander, 2008). "Cyberinfrastructure" refers to infrastructure based upon distributed computer, information, and communication technologies. As the Atkins report states, "if infrastructure is required for an industrial economy, then we could say that cyberinfrastructure is required for a knowledge economy" (Atkins, et al., 2003, pg. 5).

The Atkins report lays out a grand transformative vision. In it, cyberinfrastructure technologies are expected to facilitate new kinds of science by serving "individuals, teams and organizations in ways that revolutionize *what they can do*, *how they do it*, and *who participates*" (Atkins, et al., 2003, pg. 17). Cyberinfrastructure is envisioned to have a significant role in future science and engineering research, "cyberinfrastructure will become as fundamental and important as an enabler for the enterprise [of science and engineering research] as laboratories and instrumentation, as fundamental as classroom instruction, and as fundamental as the system of conferences and journals for dissemination of research outcomes" (pg. A-1). For example, cyberinfrastructure technologies for biomedicine allow a "third way" of conducting research that mixes the two conventional kinds of research: individual investigations and large team collaborations (Buetow, 2005).

Metadata are critical to the success of cyberinfrastructure systems. The Atkins report, taking a technology-centric view of metadata, defines metadata as "machine-readable and interpretable descriptions of the data itself" (2003, pg. 43). Metadata repositories, the report states, are necessary to "institutionalize community data holdings" by providing "tutorials and documents on data format, quality control, interchange

formatting, and translation, as well as tools for data preparation, fusion, data mining, knowledge discovery, and visualization" (pg. 42). The report also recommends that the creation of discipline-specific metadata standards be a priority for future NSF funding.

Researchers in the United Kingdom (UK) spell out similar visions. There, the development of cyberinfrastructure falls under the label of "eInfrastructure." The term "eScience," and the broader "eResearch," are used to refer to both the new technologies being developed and the set of institutions in which they are to be embedded (Borgman, 2007). The goals of eScience initiatives in the UK parallel the cyberinfrastructure visions in the US: leverage new grid computing services and semantic web technologies as a "core set of middleware services that will allow scientists to set up secure, controlled environments for collaborative sharing of distributed resources for their research" (Hey and Trefethen, 2005, pg. 818).

In response to the Atkins report and related eScience work in the UK, the US National Science Foundation released another vision report, entitled *Cyberinfrastructure Vision for 21st Century Discovery* (2007). This report describes cyberinfrastructure as consisting of the following complimentary areas: "computing systems, data, information resources, networking, digitally enabled-sensors, instruments, virtual organizations, and observatories, along with an interoperable suite of software services and tools. This technology is complemented by the interdisciplinary teams of professionals that are responsible for its development, deployment and its use in transformative approaches to scientific and engineering discovery and learning" (pg. 1).

Metadata are again called out in this vision report as an essential component in cyberinfrastructure data systems. “Metadata summarize data content, context, structure, interrelationships, and provenance (information on history and origins). They add relevance and purpose to data, and enable the identification of similar data in different data collections” (pg. 22). The vision report declares that scientists who contribute to community data repositories should be collecting and depositing metadata as routinely as they are collecting and depositing data.

The expectations for cyberinfrastructure technologies span disciplinary boundaries. In this dissertation, I focus on cyberinfrastructures for scientific research, but much of the following discussion could equally apply to research in the humanities and social sciences. Digital humanities and data-intensive social science research face challenges in managing high volume and complex data resources similar to those described below, and likewise benefit from the ongoing development and application of cyberinfrastructure technologies (Unsworth, et al., 2006; King, 2011).

1.2 Data deluge

Cyberinfrastructure and eScience initiatives have co-evolved with the development of new methods of creating and collecting data. New technologies in arguably every kind of research are producing data at rates and complexities much higher than have been possible in the past. From prototypical “big science” projects like the CERN Large Hadron Collider, to the burgeoning use of digital sensing systems in the ecological and environmental sciences, to the mass digitization of cultural artifacts for remote study by humanists, data provide opportunities and challenges for new kinds of

research (*Long-Lived Digital Data Collections*, 2005). The struggle to scale conventional data management, analysis, and preservation techniques up to this "data deluge" has many facets and involves numerous stakeholders, including researchers, funding agencies, data curation institutions, and publishers (Hey & Trefethen, 2003; Borgman, 2007; Hanson, Sugden, & Alberts, 2011).

New computational techniques are being developed to address some of these technical challenges, including the need to transfer unprecedented volumes of data, query massive databases, and compile diverse data streams (Newman, Ellisman, & Orcutt, 2003; Bell, Hey, & Szalay, 2009). In addition, as the amount of data being collected and created increases, metadata takes on more importance. Metadata concerns arose early on, as first generation cyberinfrastructures began producing large volumes of data. Hey and Trefethen (2003) describe how metadata are vital for storage and preservation of data, providing information about data provenance and user access controls, and facilitating interoperability in federated databases and data archiving systems. They also state that metadata can help in dealing with the vast outpouring of data by describing the "interesting" features of data for potential users, though they do not address the question of what "interesting" means, and to whom.

The notion that the amount of data being created is outpacing researchers' abilities to manage them is not new to the 21st century. Rhetoric of "information overload" has been used to motivate new developments in information and data management techniques at least as far back as Paul Otlet and the European Documentation movement in the early 20th century (Day, 2009). In another historical

example, the development of information science as a professional discipline in the United States after World War II was heavily motivated by a desire to manage the floods of documents that emerged in the post-war period, including technical reports from U.S. government labs (Farkas-Conn, 1990) and seized Nazi Germany government documents (Richards, 1994). Concerns about digital data are certainly a more recent development, but extend back a couple of decades or more (Maddox, 1988; Gershon & Miller, 1993).

The current attention to "data deluge," while not representing a completely new phenomenon, stems from the recognition that research data offers the possibility for new questions. New information and communication technologies offer ways to make data available and usable from a distance (though these technologies present their own problems). Perhaps even more importantly, government granting agencies are recognizing the increasing value of the digital data being collected and created through publicly funded research.

1.3 Data sharing policies

Leveraging the data created through public funds is a chief goal of cyberinfrastructure and eScience initiatives. Policy recommendations that grant recipients be more active in sharing research data have appeared with increasing regularity since the mid-1980s (Borgman, 2007). The impediments, however, that stand in the way of data sharing are well documented. In 1985 the National Academy issued a report, entitled *Sharing Research Data*, which listed a number of disincentives to data sharing, including researchers who fear the possibility of conflicting reanalysis, potential breach of confidentiality, intellectual property issues relating to possible patentable or marketable

discoveries, and legal agreements that stipulate non-disclosure as a condition of data collection (Fienberg, Martin, & Straf, 1985). These disincentives still exist, and others have been identified since. From a more practical perspective, data sharing can require significant expenditures of time and energy to work data into a form that can be shared at all, and researchers often have no career incentive to contribute their data. Academic promotions and tenure decisions are based on producing publications, not producing and managing data sets (Borgman, 2007).

With increasing frequency, government funding agencies are implementing data management and data sharing requirements and mandates as stipulations of grant agreements. The *Long-Lived Digital Data Collections* report includes a recommendation to the NSF that research proposals should be required to include data management plans, stating "research proposals for activities that will generate digital data, especially long-lived data, should state such intentions in the proposal so that peer reviewers can evaluate a proposed data management plan" (2005, pg. 47). The NSF report *Cyberinfrastructure Vision for 21st Century Discovery* (2007) lays out similar plans to institute such a requirement: "NSF's actions will promote a change in culture such that the collection and deposition of all appropriate digital data and associated metadata become a matter of routine for investigators in all fields. This change will be encouraged through an NSF-wide requirement for data management plans in all proposals" (pg. 29). These reports stop short, however, of describing how these requirements will be implemented or enforced.

The US National Institute of Health (NIH) has already implemented such a requirement. As of October 1, 2003, any research application requesting more than $500,000 of direct costs in a single year from the NIH "are expected to include a plan for sharing final research data for research purposes, or state why data sharing is not possible" (U.S. NIH, 2003, n.p.). According to the NIH policy, grant applications can include requests for additional money to fund data sharing. The NIH enforcement of the data sharing requirement allows researchers considerable wiggle room, as the NIH policy simply states that NIH program staff will assess data sharing plans for adequacy and appropriateness. The plans themselves have no hard requirements, though the NIH suggests a number of components, including an "expected schedule for data sharing, the format of the final dataset, the documentation to be provided, whether or not any analytic tools also will be provided, whether or not a data-sharing agreement will be required and, if so, a brief description of such an agreement, and the mode of data sharing" (n.p.). The timeline for data sharing is also left open, but the NIH policy states that "timeliness" for data release and sharing is generally considered to be "no later than the acceptance for publication of the main findings from the final dataset" (n.p.). The NIH has also recently implemented an open access requirement for publication deposition. This requirement stipulates that investigators receiving funding from the NIH must deposit their final peer-reviewed manuscripts into the National Library of Medicine's PubMed Central no later than 12 months after the official date of publication (Suber, 2008).

Following the NIH's example, the National Science Foundation announced in May of 2010 that research grant applications would be required to include data

management plans beginning in 2011 (NSF, 2010a). The specifics of what data management plans should include remain an open question for the NSF, as variations in recommendations for data management plans are considerable across disciplinary directorates (NSF, 2011a; 2011b). Complicating the development of these requirements is the fact that best practices for data management are not well understood in many disciplines due to the high variability of research practices. The distributed mode of science made possible by new information technologies and infrastructures adds new kinds of variability.

1.4 Components of data management

I have outlined how data management is necessary to make data available, usable, sharable, and storable in the ways that cyberinfrastructure and eScience initiatives promote. The question of how to manage data created in distributed scientific projects is still an open one, but consensus is forming around the necessity of certain components (Arzberger, et al., 2004; Anderson, 2004; Duerr, et al., 2004; Gray, et al., 2005; Uhlir & Schröder, 2007; Borgman, 2007; Berman, 2008). These include:

- Reliable technology infrastructure for storing, transmitting, and analyzing data.
- Policies that promote data management and sharing incentives
- Metadata that can facilitate data management, discovery, and use
- Effective day-to-day data management practices

In this study, I look more in depth at the last two components, the metadata practices of scientists and how they are manifested in day-to-day data management

practices. In the next section I introduce the setting for my research, the Center for Embedded Networked Sensing (CENS), an interdisciplinary research center devoted to developing a key application for cyberinfrastructure technologies in multiple scientific disciplines: sensor networks.

1.5 Research setting

As outlined in the introductory section of this chapter, my dissertation research was based within the Center for Embedded Networked Sensing (CENS). CENS is a National Science Foundation Science and Technology Center based at UCLA with four partnering institutions in southern and central California: University of Southern California, Caltech, University of California, Riverside, and University of California, Merced. Over 300 faculty members, students, and research staff from a number of disciplines have been associated with CENS since the center's inception. The main focus of CENS is to develop sensing systems for real-world scientific and social applications through interdisciplinary collaborations between seismologists, terrestrial ecologists, aquatic biologists, environmental scientists, and computer scientists and engineers. Other members of the center come from urban planning, design and media arts, and information studies. CENS was founded in 2002 for an initial five years, and received renewal funding in 2007 for an additional five years.

CENS is dedicated to developing sensing technologies that can enable the exploration of new research problems, and the exploration of existing research problems in ways that were not previously possible. Sensor networks are not themselves cyberinfrastructure technologies, but they are regularly cited as contributors to the deluge

of digital data (see for example, Estrin, et al., 2003; *Cyberinfrastructure Vision for 21st Century Discovery*, 2007, pg. 11; *Riding the Wave*, 2010, pg. 14). CENS itself is not specifically implementing cyberinfrastructure technologies; rather, CENS is a test bed for the development of sensor networks that could in the future be key applications for cyberinfrastructure systems.

In the process of developing sensing technologies through collaborations between technologists and domain scientists, CENS is producing data that have considerable re-use value. CENS researchers are documenting unique events, such as the fish die-off described in the introductory section, as well as monitoring recurring phenomena such as plant growth and seismic events. Such data could potentially be re-used in synthesis and longitudinal studies of large-scale phenomena such as climate change, species shifts, and global seismic activity. As noted in the introduction, real-world observational data are cited in the *Long-Lived Digital Data Collections* report (2007) as being of higher long-term value than other kinds of data, because observations of a particular location at a particular time cannot be replicated in the strong sense. "Observational data, such as direct observations of ocean temperature on a specific date, the attitude of voters before an election, or photographs of a supernova are historical records that cannot be recollected. Thus, these observational data are usually archived indefinitely" (pg. 19)

Collecting and archiving data in perpetuity, however, is not a chief goal of CENS. CENS researchers collect many different kinds of data, including environmental parameters like temperature and relative humidity, physical parameters like chemical concentrations and pH, as well as many parameters related to the equipment themselves,

such as battery voltages, disk space, and wireless link quality (Borgman, et al., 2007). These data are collected for immediate use in answering particular research questions. CENS data vary widely in type and formats, including numerical time-series data, geospatial data, images, audio recordings, hand collected physical samples, among many others.

Data have become increasingly visible as a CENS research product over the lifetime of the center. In the first two to three years of CENS, researchers produced little data of lasting value, as the nascent sensing technologies were not reliable enough to generate much more than experimental data of dubious quality. As the sensing technologies became more robust around 2005-2006, years three and four of CENS, they became legitimate field-usable research tools. When this occurred, the amount and quality of the data being collected by application scientists increased (Wallis, et al., 2010). Around this time, the statistics research group within CENS developed a centralized web accessible database system called Sensorbase to facilitate direct steaming of data from sensors installed in field research sites to UCLA (Chang, et al., 2006). Sensorbase did not become the central repository for all CENS data, as it was envisioned, but has become the default data management and storage tool for a number of research teams. Additionally, with the increase in the amounts of data being collected, data quality and integrity in-and-of-themselves became research topics for both social science and technical researchers (Wallis, et al., 2007; Ni, et al., 2009).

As CENS has matured, it has been more proactive in making research products available. This has stemmed from internal needs, including the administrative need to

keep better track of the center's growth, as well as from external pressure from the NSF to increase the visibility of the center's research output. Our first effort in this direction was to make CENS' research publications available on the internet through the University of California eScholarship repository (Pepe, et al., 2007). The eScholarship repository, being OAI-PMH compatible, has been successful in increasing the visibility and utilization of CENS publications on the internet.

The sharing and curation of research data, however, were not a point of emphasis until CENS started receiving outside pressure from the NSF, the primary funder of the center. During the 2009 annual NSF site visit to CENS, the NSF site visitors asked specifically that CENS make its data more widely available. In response to this pressure, we began working within CENS to construct a metadata registry that enables CENS data to be discovered by potential users. We chose to focus on a metadata registry for several reasons. One reason is the diversity of research products that might be considered "data." CENS researchers collect images, audio recordings, physical samples, and numeric data in both digital and analog form, among other formats, and these resources are distributed around community, lab, and individual computer systems. Some CENS research groups have made their data available on lab websites, but large portions of these resources reside in protected computer systems, personal laptops, or in file cabinets or refrigerators. Collecting and integrating all of CENS' data into a single system would be prohibitively expensive and time consuming, even if researchers were willing to release them. The pressure that NSF is putting on CENS to provide data to interested users, while being significant enough to have engendered our response, has been informal pressure, not a

mandate. Such a mandate would require a clear definition of data, and clear policies for what is to be released, in what form, and when (Wallis, et al., 2010). Lacking such mandates, definitions, and policies, we are designing a Dublin Core based metadata registry that will enable potential data users to discover what CENS data exist, to determine whether those data may be useful, and to learn how to acquire data of interest. These metadata descriptions will be put on the CENS web site to make them accessible to web search engines.

CENS thus provides an ideal cyberinfrastructure data management case study. To be more specific, CENS, as a collaboratory, provides a case study of the potential settings of cyberinfrastructure implementations, not of cyberinfrastructure development itself. Collaboratories are now a common organizational framework for scientific research (Finholt, 2002). Collaboratories vary widely in mission and scope. Some are organized around shared data collections, while others like CENS (a distributed research center) are organized around other goals (Bos, et al., 2007). Thus, as data sharing requirements and mandates become more common, whether formalized into policy or issued informally (as was the case with CENS), researchers in many situations will find themselves facing challenges related to managing, sharing, and preserving data that they have never had to address before.

My study thus derives from our work in developing a centralized metadata repository for CENS. In completing this dissertation, I conducted a multi-sited ethnography of the metadata creation practices of CENS researchers, approaching this topic in two ways. First, I conducted an ethnographic study of metadata creation in the

day-to-day activities of CENS researchers, and second, I studied how CENS researchers create metadata for our new CENS metadata registry. Comparing metadata creation practices across these two related sets of activities illuminates challenges in creating communal data and metadata repositories for cyberinfrastructure research communities.

Before I jump into the details of my study, however, I need to provide more background on the notion of "metadata" itself. What are "metadata" and why are they an important topic of study, particularly in relation to cyberinfrastructure? The next chapter discusses "metadata" in more detail, illustrating how notions of metadata are highly variable and contingent on the institutional setting. I also outline why a study of metadata creation by data authors can help to inform metadata efforts for cyberinfrastructure development.

2. BACKGROUND – INTRODUCTION TO METADATA

The term "metadata" literally means "data about data." The term dates from the early 1970s, but its use exploded in the 1990s during the dramatic increase in the numbers and importance of digital information resources (Greenberg, 2005). Keeping track of web pages and other digital information resources with traditional library cataloging practices proved to be difficult, despite the best efforts of library professionals, because of the malleable nature of internet-based materials. Web pages: 1) can continually change (though not all do), and 2) differ considerably from one to the next. These characteristics challenged two of the core tenets of cataloging practice, namely, that information resources can and should be cataloged only once, and that catalog descriptions should be highly standardized from resource to resource. As internet boomed, many in the library and information community turned to new approaches for documenting and describing digital information resources. "Metadata" became the accepted term to refer to such documentation and descriptions, and over time has even come to encapsulate the cataloging practices to which it was initially in contrast. Common examples of "metadata" typically include library catalog records, the HTML headers of web pages, hardware and software user manuals, and user-created "tags" on internet-based social networking tools.

In my study, I use "metadata" to refer to *documentation, descriptions, and annotations created and used to manage, discover, access, use, share, and preserve informational resources.* This definition emphasizes the physicality of metadata. Metadata are documentary forms, not abstract concepts or information that exists without material form (Blanchette, 2011). My definition also focuses on the functional role of metadata in work practices. Metadata are created in relation to specific work tasks; they serve one or more functions within particular social and institutional settings. My definition draws most heavily from Coyle (2010). Coyle outlines how metadata are constructed, constructive, and actionable:

- Metadata are constructed: They are an artificial creation not found in nature.
- Metadata are constructive: They are created for a purpose, activity, or to solve a problem.
- Metadata are actionable: They are intended to be useful in some way. (pg. 6)

My definition of metadata, along with Coyle's, is just one of many that exist. The term "metadata" has highly situational understandings. In this chapter, I examine what some of the different understandings of metadata are, and how those understandings vary based on institutional settings. The diversity of definitions and notions of metadata that I discuss in the next few sections illuminates a fundamental tension in the vision of cyberinfrastructure laid out in Chapter 1. Cyberinfrastructure data systems cannot succeed without standardized forms of metadata. Metadata standardization involves not only a common operationalization of data description elements and representation schemes, standardization also assumes a common conception of what metadata are, and

what forms they will take. As I illustrate below, however, conceptions of metadata are highly fluid and multiple.

2.1 Notions of metadata

As the term 'metadata' has spread, it has been defined and redefined in numerous ways. Many notions of metadata have moved past the most common definition, the literal "data about data," to more nuanced and pragmatic discussions of requirements and functions. Greenberg (2005), for example, stated that the term metadata "addresses data attributes that describe, provide context, indicate the quality, or document other object (or data) characteristics" (pg. 20).

As Coyle's definition suggests, discussions of metadata are often heavily tied to particular settings, people, or types of resources (Campbell, 2005; Greenberg, 2009). In this section, I contrast notions of metadata in information institutions and in scientific research settings to illustrate how understandings and implementations of metadata can range broadly depending on the institutional context.

2.1.1 Information institutions and metadata

A primary institutional responsibility for information institutions such as libraries and archives is to make resources discoverable, usable, and accessible for their patrons. The description and representation of information resources is a central activity in fulfilling this responsibility. Information professionals create descriptive representations, such as library catalog records and archival finding aids, which serve as surrogates for information resources in discovery, organization, and access systems. "Metadata" is often used as a blanket term for descriptions and representations of information resources,

despite the extensive history of library cataloging work. Content standards, such as the Anglo-American Cataloging Rules (AACR), and format or encoding standards, such as the Machine Readable Cataloging (MARC) format, are commonly characterized as "library related metadata standards" (Intner, Lazinger, & Weihs, 2006).

Notions of "metadata" in current information institutional settings descend from over a century of tradition and practice in organizing, describing, and providing access to library materials through catalog records and other similar descriptive surrogates (Taylor and Joudrey, 2009). These traditions and practices are heavily based on strong organization and description principles. Svenonius (2000), in a book widely read in graduate library and information studies programs, describes a set of general principles that in her view govern the design of bibliographic and other descriptive systems (pg. 68):

- *user convenience*: descriptions should be made with the user in mind, including using vocabulary according with common usage
- *accuracy of representation*: descriptions should be based on the way an information entity describes itself, and should faithfully portray the entity described
- *sufficiency and necessity*: descriptions should be sufficient to achieve stated objectives and should not include elements not required for this purpose
- *standardization*: descriptions should be standardized to the extent possible
- *integration:* descriptions for all types of materials should be based on a common set of rules, to the extent possible

These principles, Svenonius states, offer directives for how information systems and the languages they use should be designed. Notions of metadata in information institutions, both new and old, reflect this principled approach.

Gilliland (2008) takes a broad but principled view of metadata, stating that it can be thought of as "the sum total of what one can say about any *information object* at any level of aggregation" (n.p., italics in original). Later in her paper, she more pragmatically defines metadata as "the value-added information that [information professionals] create to arrange, describe, track, and otherwise enhance access to information objects and the physical collections related to those objects" (n.p.). She breaks the term "metadata" down into five types: *administrative, descriptive, preservation, use,* and *technical*. Examples of these types reflect back on her broad view of metadata and illustrate the uses of metadata: cataloging records, selection criteria for digitization, finding aids, annotations by creators and users, documentation of physical condition of resources, hardware and software documentation, authentication and security data such as encryption keys, passwords, circulation records, and search logs, among many others. As this list illustrates, Gilliland considers most traditional information institutional resource description practices to fall under the general term "metadata." These notions of metadata are motivated by what she considers to be the three main features of information objects: content (the intrinsic aspects of the object), context (extrinsic aspects), and structure (intrinsic and extrinsic aspects of the object regarding associations with other objects). Gilliland's discussion does not explicitly call out Svenonius's principles, but in discussing the roles of metadata

"in an environment where a user can gain unmediated access to information objects over a network," she gives the following list. Metadata:

- certifies the authenticity and degree of completeness of the content;
- establishes and documents the context of the content;
- identifies and exploits the structural relationships that exist within and between information objects;
- provides a range of intellectual access points for an increasingly diverse range of users; and
- provides some of the information that an information professional might have provided in a traditional, in person reference or research setting. (n.p.)

Overlap with Svenonius's principles can be seen in this list. The first two points in this list address the *accuracy of representation* and *sufficiency and necessity* principles, the third point addresses the *standardization* and *integration* principles, and the last two points address the *user convenience* principles.

The development of metadata for digital libraries followed this principled tradition as well, through the development of standards such as the Metadata Encoding and Transmission Standard (METS, 2009; McDonough, 2006) and the Dublin Core Metadata Initiative (DCMI, 2009). This principled approach to metadata design comes out clearly when looking closely at the development of the Dublin Core metadata schema. The Dublin Core was developed as a minimal set of elements for use in describing and discovering digital objects (Sugimoto, Baker, & Weibel, 2002). These goals themselves reflect Svenonius's principles of *standardization* and *integration*. The

development of the Dublin Core metadata schema followed a set of assumptions that further reflect Svenonius's other principles. As Weibel (1995, n.p.) describes, these assumptions were:

1. *Intrinsicality:* The Dublin Core concentrates on describing intrinsic properties of the object. Intrinsic data refer to the properties of the work that could be discovered by having the work in hand, such as its intellectual content and physical form.
2. *Extensibility:* Extension mechanisms will allow the inclusion of intrinsic data for objects that cannot be adequately described by a small set of elements. Extensibility is important because users may wish to add extra descriptive material for site-specific purposes or specialized fields.
3. *Syntax Independence:* Syntactic bindings are avoided because it is too early to propose formal definitions and because the Dublin Core is intended to be eventually used in a range of disciplines and application programs.
4. *Optionality:* All elements in the Dublin Core are optional.
5. *Repeatability:* All elements in the Dublin Core are repeatable.
6. *Modifiability:* Each element in the Dublin Core has a definition that is intended to be self-explanatory. However, it is also necessary that the definitions of the elements satisfy the needs of different communities. This goal is accomplished by allowing each element to be modified by an optional qualifier.

Assumption #1 clearly overlaps with the *accuracy of representation* principle, assumptions #2 and #6 overlap with the *user convenience* principle, and assumptions # 4

and #5 overlap with the principle of *sufficiency and necessity*. The lack of syntactic structure, illustrated in assumption #3 above, is a notable deviation from the principled approach.

The success of the Dublin Core schema is subject to debate. Though it has become the de facto standard for sharing and harvesting data through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), it has been criticized as being too flexible and incomplete, and too oriented towards human readability. For example, Lagoze, Lynch, and Daniel (1996) discuss the metadata requirements for digital documents in relation to the Dublin Core metadata set. They describe how multiple "classes" of metadata might be necessary in different situations, and point out that *descriptive* metadata, the focus of the Dublin Core, is often insufficient for "real work applications." Other metadata types that might be required in a given setting, they argue, include *terms and conditions*, *administrative data*, *content ratings*, *provenance*, *linkage or relationship data*, and *structural data*. Greenberg (2005) shows how these categories map to similar typologies developed by Greenberg (2001) and Caplan (2003). The development of "qualifiers" for Dublin Core metadata schema and the creation of "application profiles" in other schemas are part of a movement to increase the structure and specification of metadata (Howarth, 2005). Human readability is important for many reasons, but this feature can hamstring attempts to automate metadata creation and sharing, discussed further below.

In information institutional settings, research data have been treated with the same kinds of principled approaches. The second edition of the *Anglo-American Cataloging*

Rules (AACR2), released in 1979 and still the main library cataloging standard in use today, included a chapter, Chapter 9, dedicated to "machine-readable data files" (Dodd, 1982). This chapter followed the principles and practices developed for cataloging books, serials, music, etc., including specifying rules for recording titles, physical descriptions, notes, standard numbers, and other standard kinds of information. Chapter 9 still exists in the most recent updates to AACR2, but has been re-named to "electronic resources" in order to encompass not only data files, but also programs and other resources available in electronic form (AACR2, 2005).

2.1.2 Metadata for scientific data

In scientific research, particularly in disciplines such as environmental science and ecology where small scale projects are the norm, metadata practices are much less principled and standardized. Metadata definitions and schemas abound in most disciplines, with the degree of standardization varying with the uniformity of research methods, data types, and cultures of data sharing (Bowker, 2000; Griffiths, 2009). Astronomers, for example, have made substantial progress in developing community data and metadata standards (Hanisch, 2006), as have seismologists (Ahern, 2002), while habitat ecology has been less successful in this endeavor despite considerable effort (Borgman, et al., 2007; Millerand & Bowker, 2009).

Discussions of metadata types and functions are common in many scientific fields. Michener, et al. (1997), for example, define metadata as "representing the higher level information or instructions that describe the content, context, quality, structure, and accessibility of a specific data set." They provide a thorough description of the varied

kinds of metadata that should be used to describe ecological data (pg. 330). They outline five classes of metadata descriptors: *data set descriptors*, *research origin descriptors*, *data set status and accessibility*, *data structural descriptors*, and *supplemental descriptors*. Similarly, Ellison, et al. (2006) discuss how metadata that describe the structure and content of ecological data sets must be supplemented by documentation of the analytical processing by which the data were derived, or "process metadata." Gray, et al. (2005) focus on the kinds of metadata necessary for the analysis of large data sets, regardless of the discipline. They consider metadata to consist of "descriptive information about data that explains the measured attributes, their names, units, precision, accuracy, data layout and ideally a great deal more. Most importantly, metadata includes the data lineage that describes how the data was measured, acquired or computed" (n.p.). In a similar article, Singh, et al., (2003) outline a typology of metadata in distributed data management systems, with their typology broken up into five levels: physical, domain-independent, domain-specific, virtual organization, and user metadata. This typology might best be characterized as ranging from most technical (physical metadata) to personal (user metadata), with institutional levels in between (domain-independent, domain-specific, and virtual organization metadata). Lawrence, et al. (2009) take a different approach, outlining five classes of metadata: archive, browse, character, discovery, and extra. This typology is notable in that the final class - "extra" - consists of what the other articles in this paragraph consider to be most important: discipline-specific metadata, such as descriptions of equipment and data collection practices, and metadata

about the data derivation and computation. As we can see then, notions of metadata for scientific data differ considerably depending on the specific setting.

Looking at metadata from a functional point of view, we can see similar variance amongst the sciences. Helly, Staudigel, and Koppers (2003), discussing research and data sharing in the earth sciences, state that the two main functions of metadata, in their view, are "discovering the existence of data by searching a metadata catalogue or its equivalent…[and] documentary information describing the content, context, quality, structure, accessibility and so on of a specific data set." Michener (2006), in outlining metadata functions for ecological data, considers the "three levels of increasing metadata functionality that may be easily categorized from a scientific perspective" to be: (1) support data discovery; (2) facilitate acquisition, comprehension and utilization of data by humans; and (3) enable automated data discovery, ingestion, processing and analysis. (pg. 4)

In a paper titled "Understanding Metadata," the National Information Standards Organization (NISO) outlines both types and functions of metadata. They define metadata as "structured information that describes, explains, locates, and otherwise makes it easier to retrieve and use an information resource" (NISO, 2004, pg. 1). They state that there are three main types of metadata: *descriptive*, *structural*, and *administrative* (which includes *rights management* and *preservation* metadata as subsets), but later in the paper make a point to emphasize the importance of "technical metadata" without giving a precise definition or outlining how it relates back to any of the previous three categories. Functionally, NISO describes metadata as facilitating the discovery of

relevant information, as well as helping to organize electronic resources, facilitating interoperability and legacy resource integration, providing digital identification, and supporting archiving and preservation.

2.1.3 Comparing metadata notions

Comparing the notions of metadata in the two institutional settings discussed so far, information institutions and scientific research, reveals important differences. The approach in information institutions has been to engage in very principle-based development of metadata schemas and systems. Standardization across institutional boundaries has been very important, as reflected in cross-cutting standards such as the AACR2, MARC, EAD, METS, and the Dublin Core. In discussions of metadata for scientific data, on the other hand, metadata typologies and functionalities are rarely viewed in the same kind of principle-based way. Metadata typologies vary considerably from one project to the next, and are often very customized to particular kinds of data or resources.

Burnett, Ng, and Park (1999) provide a similar comparison between what they call the "bibliographic" description tradition that derives from library science and the "data management" metadata tradition that they state derives from computer science. In their view, the chief difference between the two traditions is the emphasis in the "bibliographic" description approach on the development of "rules and standards for the operationalization of metadata, or the rules for data modeling" (pg. 1211). In contrast, they characterize the "data management" approach as favoring more loose definitions of

what metadata descriptions can or should include, usually customized to an individual project or setting.

Standardization always requires distillation of complex situations and processes, and must be negotiated among stakeholders with differing interests (Bowker & Star, 2000). Many scientific projects may not require, or benefit from, the kinds of metadata standardization that information institutions have developed for bibliographic and archival description. In these cases, customization and looseness of metadata descriptions are to be expected. The largely ad-hoc approach to metadata has resulted, however, in a multiplicity of metadata standards and schemas in almost every domain. Gorman (2006) goes so far to state that the loose approach to metadata development and implementation is based on a failed utopian dream of a "third way" of description (with the bibliographic description approach and the free-text "Google search" approach being the first two ways), as illustrated by the current metadata cacophony. He argues that metadata projects will have to move toward the more structured and complex approaches used in the library and archival community if they are to be successful over the long term. But Gorman's argument might gloss over the important differences in institutional settings and goals, and the particular situational activities in which metadata play a role. I return to this issue in Section 4 below, in a discussion of social studies of workplaces and scientific research settings.

2.1.4 Metadata as a fluid and multiple concept

As the diversity of the definitions, functions, and roles given above illustrates, metadata is not a definite and singular concept. Rather, it is a fluid, multiple, and

fractional concept (Law, 2004). Metadata is "fluid" in that file naming conventions, catalog records, data descriptions in repositories, user tags on YouTube, personal Excel spreadsheets, emails, and html tags can all be called "metadata." Metadata, as a concept, is also characterized by "multiplicity" in that it is enacted differently in different social settings and situations, from Dublin Core records created by information professionals to "process metadata" created by scientists to document their analysis techniques. These enactments can overlap, as when domain scientists and information managers work together to create metadata for a data repository. They can come into conflict, as when local description schemes are incompatible with agreed upon description standards. Or they may overlap in fractional ways, for example, if "process metadata" created by scientists get incorporated into Dublin Core records created and managed by information professionals.

It is useful to use Edwards' (2010) "metadata friction" metaphor to conceptualize challenges and problems that arise in the process of creating, handling, managing, and preserving metadata products. In this conceptualization, metadata include formal standardized metadata products and informal metadata processes (Edwards, et al., in press). Metadata products, exemplified by records in standard formats such as Dublin Core, help increase the precision of research interactions by facilitating interoperability, machine readability, and resource discoverability. Informal metadata processes, such as personal emails and face-to-face discussions, lubricate the research process by smoothing the communication of data and metadata. I use the term "metadata practices" to encompass both metadata products and processes. The notion of "practices" emphasizes

an interactional view, focusing on how people interact with each other and with their surroundings, and the "consequences and interpretations of [their] actions for themselves and for others" (Dourish, 2004, pg. 28).

2.1.5 Metadata in/and ontology

Metadata reflect an individual's ontology. That is to say, metadata reflect what is visible and important about the thing being documented to the individual(s) responsible for the documentary act. The relation between metadata and ontology is often implicit in discussion of cyberinfrastructure data systems, but is not often explicated. In this section, I detail how metadata fit within ontologies, both in the technical and philosophical sense of the term "ontology."

Ontology, in the philosophical sense, refers to the study of what exists in reality and the general features and relations of whatever that might be (Hofweber, 2009). The term "ontology," however, has been adopted by computer science in recent years to refer to an engineering artifact which maps terms, relationships, and meanings in a specific knowledge domain. In the rest of this dissertation, I call these artifacts "technical ontologies" to avoid confusion with the philosophical sense of ontology. A technical ontology is "constituted by a specific *vocabulary* used to describe a certain reality, plus a set of explicit assumptions regarding the *intended meaning* of the vocabulary words" (Guarino, 1998, pg. 4, italics in original). Technical ontologies closely resemble classification systems, in that they relate concepts to terms and definitions, provide maps of relationships among terms, and support information retrieval, among other functions (Soergel, 1999).

Metadata and technical ontologies are often discussed together, as both are intended to be used in integrating information (see for example, *Cyberinfrastructure Vision for 21st Century Discovery*, 2007, pg. 22 and 28). According to Doerr, Hunter, and Lagoze (2003), the key difference between them is that metadata are typically created, edited, and used by humans, while technical ontologies are created as formal models to be used by machines. Because of this, they state, for technical ontologies "higher levels of complexity are tolerable and the design should be motivated more by completeness and logical correctness than human comprehension" (pg. 2). This is a useful, but not hard, distinction. Metadata are often generated and used by machines, for example when digital cameras create photo documentation that indicates the size, format, and resolution of an image, and similarly, technical ontologies are not exclusive to machine use. The mass of YouTube user tags, for example, are used by both humans and machines, and could be considered both as metadata and as a technical ontology.

A more useful distinction between metadata and technical ontologies is that technical ontologies provide categories, terms, and types of relationships from which particular metadata descriptions can be created. In discussing technical ontologies, Srinivasan, Pepe, and Rodriguez (2009) note, "Ontologies are often created for multiple purposes. First, they allow information objects to be classified, providing users with basic metadata with which they can be associated. Second, their structure describes not just the information in the database but the semantic relationship presumed between categories. The ontology, in other words, presents the worldview of the information system, allowing information to be presented within a set of classifications and categories presented

alongside one another" (pg. 610). If technical ontologies present the worldview of the information system, metadata map that worldview to particular objects (or documents) within that world.

Moving now to the relation of metadata and ontology (in the philosophical sense of ontology), metadata reflect the ontologies of the individuals who create them. As human beings living in societies, our ontologies are social ontologies. We know the constituents of our world, and how to relate to those constituents, as members of social systems. Our ways of knowing and acting within our world are dependent on our position within multiple and overlapping social systems (Weissman, 2000). Socially defined entities, such as "data" or "metadata" have "deontic powers" (Searle, 2006), meaning that once they are recognized to exist within a social system they entail social obligations, responsibilities, and duties. Furthermore, "to recognize something as a right, duty, obligation, requirement and so on is to recognize a reason for action" (Searle, 2006, pg. 19). "Data" are recognized to be central to the ontology of scientific work. In an editorial in *Science*, Hanson, Sugden, & Alberts (2011) go as far as saying, “Science is driven by data” (pg. 649). This statement glosses over the practical work required to ensure its truth. Data can only drive scientific work if they verifiably meet socially acceptable criteria for what “data” are. Metadata practices are one means through which “data” become socially acceptable. What counts as socially acceptable data and metadata varies depending on the social situations in which data and metadata are used, and reflects the social ontology of the individual(s) involved.

Before metadata can be used, however, they must be created. Research roles and career trajectories contribute to the development of knowledge about (and practices for) creating metadata. The next section discusses metadata creation techniques, as well as the different groups of individuals who play a part in the metadata creation process.

2.2 Creating metadata

Above, I outlined the important role of metadata in cyberinfrastructure data systems, as well as the ways that metadata are understood and enacted multiple social settings. The social settings in which metadata are created have a large impact on what form metadata take. In libraries and archives, for example, the creation of metadata is an institutionalized task. Catalogers, archivists, and, increasingly, professionals with titles like "metadata librarian" (Choudhury, 2008; Delserone, 2008) are assigned responsibility for creating metadata. The ways that metadata creation are, or will be, institutionalized in scientific cyberinfrastructure data systems are less clear. In this section, I discuss metadata creation processes and roles in relation to scientific data.

2.2.1 How are metadata created?

Metadata can be created through both automated and manual processes. Both of these methods present challenges. Greenberg (2004) describes how automated metadata creation techniques typically follow one of two approaches, extraction or harvesting. In metadata extraction, "an algorithm automatically extracts metadata from a resource's content" (pg. 62). Common applications of the extraction approach include automatic abstract generation for publications and summary displays of web pages given by web-search systems. Metadata harvesting, on the other hand, involves compiling metadata

automatically from distributed resources, such as collecting HTML headers from web sites. As Greenberg notes, "the 'harvesting process' relies on the metadata produced by humans or by full or semi-automatic processes supported by software" (pg. 63).

Automated metadata extraction and harvesting methods primarily exist for text-based documents. Many of these techniques do not directly extend to creating metadata for scientific data, as a considerable proportion of scientific data is not text-based. This is a significant impediment for eScience and cyberinfrastructure data collection, management, and sharing systems to overcome. As Hey and Trefethen note, "in many instances, the sheer volume of data will dictate that this [metadata annotation] process be automated" (2005, pg. 818). "Workflow" systems are one approach to generate metadata automatically for scientific data. A workflow is a precise step-by-step description of a scientific procedure that acts as a script for the coordination of research tasks. Tasks might be computational processes, such as "running a program, submitting a query to a database, submitting a job to a compute cloud or grid, or invoking a service over the Web to use a remote resource. Data output from one task is consumed by subsequent tasks according to a predefined graph topology that 'orchestrates' the flow of data" (Goble and De Roure, 2009, pg. 138). The more structured the workflow, the easier it is to automate the creation of metadata, particularly to record provenance information about how data have been derived or changed over time (Gray, et al., 2005; Chin & Lansing, 2004). Outside of highly instrumented laboratory settings, however, workflows are very difficult to implement. Automatic techniques for recording provenance require considerable customization to the particulars of the data creation instrumentation and processes, which

can vary widely from setting to setting, and in some research settings, from case to case (Miles, et al., 2007).

The challenges that stand in the way of automatic metadata generation for scientific data suggest that metadata creation in cyberinfrastructure projects will depend in part or in whole on manual efforts. The next section outlines how different groups have different roles and responsibilities regarding the metadata creation process.

2.2.2 Responsibility for creating metadata

The responsibility for creating metadata falls on different individuals depending on the institutional setting. The National Science Board *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century* report outlines four main actors who play important roles in the data collection and curation process:

- *Data authors*: the scientists, educators, students, and others involved in research that produces digital data.
- *Data managers*: the organizations and data scientists responsible for database operation and maintenance.
- *Data scientists*: the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection.
- *Data users*: the larger scientific and education communities, including their representative professional and scientific communities. (2005, pg. 25-28)

Metadata responsibilities are spread across these groups. The report assigns responsibility for metadata creation to data authors. In turn, data managers are responsible for participating in the creation of community metadata standards and ensuring that these standards are enforced. Data scientists are assigned no specific responsibility in relation to metadata, but the report does mention that data scientists are in some cases called upon to cross-check and combine metadata to ensure accuracy.

Following the release of the *Long-Lived Digital Data* report, Swan and Brown (2008) conducted a study of data management practices in the United Kingdom research community. They lay out their own typology of important data management roles in the following manner:

- *Data creators or data authors:* researchers with domain expertise who produce data…
- *Data scientists:* people who work where the research is carried out – or, in the case of data centre personnel, in close collaboration with the creators of the data – and conduct all or a number of the [data author, data manager, and data user] functions…
- *Data managers:* people who are computer scientists, information technologists or information scientists and who take responsibility for computing facilities, storage, continuing access and preservation of data…
- *Data librarians:* people originating from the library community, trained and specialising in the curation, preservation and archiving of data… (pg. 8)

They describe how the boundaries between these roles may overlap, with certain individuals taking on more than one role. The responsibility for metadata is less clear cut in their discussion, due to the fuzzy nature of the boundaries between the groups.

NISO (2004) emphasizes the cooperative effort between researchers, technical staff, and information professionals that is often necessary to create metadata. As they describe, many configurations between these groups are possible, but typically technical staff who digitized or created the digital object create the administrative and structural metadata, while it is best if the originator of the resource provides the descriptive metadata, particularly in the case of scientific data sets. Wayne (2005) notes that in practice metadata often are incomplete, ambiguous, or corrupt. As a potential remedy, she suggests that the division of metadata labor emphasized by NISO be instituted as policy. She outlines how responsibilities for different parts of the metadata production process should be assigned to different groups. Managers, technicians, scientists and field staff, analysts, information technology/system managers, and data stewards would be responsible for disparate metadata tasks, from the coordination of metadata collection (managers), to managing metadata records (system managers), to maintaining and distributing data and metadata (data stewards). Creating this kind of metadata workflow and consolidating the efforts of multiple contributors in the way Wayne recommends is sure to require considerable effort and sustained emphasis, but can mitigate the problems associated with asking scientists or information professionals to create metadata without contributions from members of the other group (Currier, et al., 2004).

The fact that data scientist, manager, or librarian positions are far from ubiquitous in research settings confounds these attempts to formalize metadata responsibilities (Palmer, et. al, 2007). Individuals working as data scientists, managers, or librarians often reach their positions through "accidental" career choices (Pryor & Donnelly, 2009). Where these positions exist, they are usually in institutions dedicated to sustained data management and curation, such as the Long Term Ecological Research (LTER) program, which employs dedicated data managers to facilitate their mission of long term stewardship of ecological data (Karasti & Baker, 2008; Karasti, Baker, & Halkola, 2006). The cost of employing people in these positions can be prohibitive. Lord and McDonald (2003) found that staff salaries constituted between 69% and 82% of the total budget of large data archives.

In practice, data creators are often expected to create metadata for their data without training or help. This can help keep down costs and leverage their firsthand knowledge about the data (Jones, et al., 2006; Whitlock, 2011), but requires them to transfer their knowledge in an unfamiliar fashion, as data authors often have little experience in creating structured metadata using formal schemas. Scientists may understand that metadata and metadata standards are important, but be unaware of existing standards in their research area (Cragin, et al., 2010). Additionally, the time, energy, and attention involved in creating, collecting, assembling, checking, and/or understanding metadata can be significant (Edwards, 2010). For working research scientists, these metadata costs are added on top of their primary scientific work, which typically focuses on using data, not describing them for potential future users. Therefore,

there can be great difficulty in getting data creators to create metadata (Jones, et al., 2006). In some cases, researchers will refrain from sharing data because it takes too much effort to produce the data and associated documentation necessary for its use (Campbell, et al., 2002).

Cultural norms of data sharing in different scientific communities are intimately tied to the contexts of data production, namely, the social practices, relationships, and material conditions in which data and metadata production takes place (Vertesi & Dourish, 2011). Few studies have investigated how working scientists approach metadata creation tasks. Most relevant are a number of studies conducted by Greenberg, who examined web resource authors as metadata creators. Greenberg's studies showed that authors of web resources consider metadata to be valuable for resource discovery. Tools that simplify and streamline the metadata process can enable resource authors with little formal training to create metadata equal in quality to metadata created by information specialists (Greenberg, et al., 2002; Greenberg, et al., 2003). Resource authors in her study believed that they would benefit from collaborating with metadata experts such as catalogers, (Greenberg & Robertson, 2002), but Crystal and Greenberg (2005) showed that effective information system design can mitigate some of the difficulties resource authors may encounter while creating metadata, making collaboration with catalogers less necessary. Greenberg's studies are useful in indicating how resource authors can serve as effective metadata creators, but they have limited applicability to my study of data authors for two reasons. First, they focused on text-based web resources. Research data, which largely are not text-based, challenge the metadata creation process in ways

that other text-based digital resources do not. Second, Greenberg, et al. did not study how the metadata creation process fit into the everyday work practices of the resource authors involved in their experiments. Their studies presume that metadata are going to be created, and illustrate how resource authors can be effective in performing that task. As described above, however, for working scientists metadata is often at best an afterthought as they perform their primary research. It is not safe to presume that data authors, as working scientists, will have the time, resources, or the inclination to create metadata at all. In an off-shoot of Greenberg's work, White (2010) performed an interview study with evolutionary biologists about their data organizational work. She notes that scientists understand the importance of performing data organizational and metadata work, and do so for a number of reasons: "to help data sharing, …to change the way people think about their field, and to assist in collaborative work environments for data collection or storage" (pg. 170).

2.2.3 Research question 1 – everyday metadata creation

In what situations, then, do working scientists create metadata? As I note above, metadata have highly situated understandings in scientific settings. Conceptions of the role and functions of metadata vary from discipline to discipline, and not according to any clear characteristics. The first challenge for my study, then, is to understand how researchers in my particular research setting understand metadata. This involves looking not only asking researchers about "metadata" as a concept, but studying what metadata scientists actually create. I know from previous studies that CENS researchers rarely use standardized forms of metadata. I also know that CENS researchers share their data

amongst their project teams, and re-use their own data in multiple ways (Borgman, et al., 2007). Thus, researchers must have some forms of documentation that they use on a day-to-day basis. My first research question is centered around understanding what those forms of day-to-day documentation look like, and how they are created.

- Research Question 1 - How and where are metadata created, by whom, and for what purpose in everyday research?

Related questions focus on researchers' language around the topic of "metadata." What does the term "metadata" itself mean to scientists in distributed projects? What other language do researchers use in reference to data description activities? Linguistic formulations reveal important characteristics about the ontologies of individuals and communities (Searle, 2006), and are important to the successful development of work practices. In the next chapter, I outline theoretical and methodological perspectives in support of a study of the day-to-day research practices of working scientists in order to develop Research Question 1 further.

3. THEORETICAL AND METHODOLOGICAL PERSPECTIVES

Edwards, et al. (2007), state that "data are the product of 'working epistemologies' that are very often particular to disciplinary, geographic, or institutional locations" (pg. 32). Metadata are likewise products of 'working epistemologies', and are enacted in different ways in different situations. Viewing metadata from the point of view of the objects (physical or digital descriptions/representations) being created and the functions that they serve glosses over the practical work involved in their creation and use. Social interactions around data are an important aspect of data management, use, and sharing, and metadata cannot completely replace or support these interactions (Birnholtz & Bietz, 2003; Zimmerman, 2007). Studying this practical work requires an analytical frame that views metadata as a type of practical action, not as objects to be manipulated, shared, and used. Many theoretical frameworks are applicable to studies of information issues in distributed scientific collaboration, and increasingly information researchers are using frameworks from sociology and anthropology (Cronin, 2008). Information issues relating to scientific research can be studied from a number of perspectives, including that of place (lab and field, Kohler, 2002), people (individual and group, Roth & Bowen, 2001; Cragin & Shankar 2006), and discipline (application science and technology disciplines, Wallis, et al., 2008), among others.

In this chapter, I outline how sociological, anthropological, and philosophical

studies of scientific work inform my study. The work surveyed here, while far from being a homogeneous set of ideas and perspectives, illustrates the fruitfulness of studying the contested, contingent, and constructed nature of scientific, engineering, and medical practices through ethnography and participant observation in everyday lab and field settings. I draw from these studies to provide theoretical and methodological foundations for my investigation of the everyday metadata creation practices of researchers in distributed scientific research communities.

3.1 Science and Technology Studies

The sociological sub-discipline of science and technology studies (STS) provides a well-established tradition of investigating social issues in scientific research settings. STS methods and theories have been widely applied to studies of information related issues and phenomena. White and McCain (1998) showed that, as of 1995, prominent scholars in the sociology of science were among the most cited people in information science, including Kuhn (1962, 1970), Price (1965), and Merton (1973). Many more recent developments in STS are also relevant to information studies, including laboratory studies, actor-network theory, social studies of technology, social construction of technology, epistemic cultures, feminist STS, and workplace studies, among others. These developments are as diverse as they are similar, but a unifying theme among them is the emphasis on the diversity of scientific methods and work practices (Traweek, 1996). All of these STS developments will not be reviewed in depth here; Van House (2004) provides a comprehensive review of how the full range of work in STS applies to information science research. I will instead discuss important works in a few areas of

STS that I believe are most useful to my study: laboratory studies and actor-network theory.

"Laboratory studies" as a label refers to a body of research in which social scientists engaged in field studies of the actual work being done in scientific laboratories. These studies use "direct observation and discourse analysis to document the actual, messy work of science, in both its material practice and its sociality" (Van House, 2004, pg. 11). Laboratory studies stress the importance of "practice, tools, and technicians in the construction of knowledge, the interaction of the human and the nonhuman, and the role of embodied skills, as opposed to the sanitized reports of science as an intellectual, cognitive activity" (pg. 13). In one of the most widely known laboratory studies, Latour and Woolgar (1979) show how both formal and informal communication methods are used to share and evaluate information and judge personal reputations in the daily course of scientific research. As Latour and Woolgar describe, scientific research practices are organized around processes of "inscription," in which the objects of study, such as rat brains, are transformed into traces, graphs, and texts by laboratory machines and researchers. These inscription processes lead to the creation of "immutable mobiles" (Latour, 1987), particularly publications, that transport the laboratory activities into other settings. Graphs and publications are "immutable" in the sense that once created they exist as independent unchangeable entities in the world, and they are "mobiles" in the sense that they traverse across organizational boundaries via personal networks (sharing of pre-prints, conference presentations) and professionally sanctioned publication venues (journals).

Latour and Woolgar also describe how machines and research methods that become widely adopted to the point of "immutability" become "black-boxed," or accepted as non-problematic without contestation or investigation. This is not a foregone result, however, of scientific machines or methodologies. Traweek (1988) details how for high-energy physicists "inventing machines is part of discovering nature" (pg. 49). In high-energy physics, machines, such as particle detectors, are always sources of contestation: who gets to build and use them, how their accuracies and inaccuracies should be measured and reported, who gets to speak for their results, and who gets to evaluate those results, among other issues. In Traweek's account, the machines of high-energy physics are not only the producers of texts, but are themselves texts that are produced, read, and contested.

Actor-network theory (ANT) grew out of these laboratory studies, particularly those of Bruno Latour, Michel Callon, and John Law. ANT states that we need to "follow the actors" in the settings we are researching in order to understand the associations, both weak and strong, between them. "Actors," according to actor-network theory, can be human, machines, facts, laws, objects, and institutions (Bowker & Star 1996, Latour 1987). The distinction between social and technical factors, as well as the distinction between humans and non-humans, is considered to be artificial. According to ANT, information technology, texts, and humans need to be analyzed in the same terms, namely as actors with agency.

ANT is less a formal theory than a methodological perspective. Latour (2005) describes the three ANT "moves" that are necessary to study the distributed associations

between the diverse actors that constitute science and technology research settings:

1) *Localizing the Global:* In his first move, Latour eliminates the dichotomy between the micro and the macro. He posits that chains of association link places, times, and actors. Following these links step by step prevents social scientists from applying arbitrary "social" explanations for complex situations.

2) *Redistributing the Local:* This move asks the reverse question: "How is the local itself being generated?" (2005, pg. 192) Latour uses this question to analyze how "face-to-face" interactions, for example interactions that may appear to involve few actors, actually involve great (and varying) numbers of actors, each of which may act on different time scales (for example geological, academic, and daily time scales), and may displace other actors out of view (Callon, 1986). For example, a person using a map to navigate in a car is linked to an unending chain of other actors: the makers of the map, other drivers, new roads or buildings that are not on the map, road signs, police regulations, the car manufacturer, etc. (Latour, 1999, Chapter 2)

3) *Connecting Sites:* The third move, connecting sites, is, according to Latour, what we get if we practice the first two moves – *localizing the global* and *distributing the local* - together. Taking together the idea that chains of associations link places, times, and actors (first move), and the idea that each local interaction consists of a multitude of such chains (second

> move), we "end up with a superposition of various canals as entangled and varied as those that an anatomist would see if she could simultaneously color all the nerve, blood, lymph, and hormone pathways that keep organisms in existence" (Latour, 2005, pg. 220). Once these distributed and interconnected networks are made visible, Latour claims, we become able to look at the connections and mediators that constitute the associations between actors.

Pickering (1995) builds on the ANT view of "science in action" to propose understanding science as consisting of a "practical goal-oriented and goal-revising dialectic of resistance and accommodation" (pg. 22). Pickering retains the ANT flattening of the "human" and "non-human" distinction, but he critiques the atemporal nature of actor-network theory. Specifically, he states that ANT does not take into account the timelines of scientific work. He instead details a real-time and performative understanding of scientific research, which involves a "dance of agency" between humans and machines:

> "As active, intentional beings, scientists tentatively construct some new machine. They then adopt a passive role, monitoring the performance of the machine to see whatever capture of material agency it might effect. Symmetrically, this period of human passivity is the period in which material agency actively manifests itself. Does the machine perform as intended? Has an intended capture of agency been effected? Typically the answer is no, in which case the response is another reversal of roles: human agency is once more active in a revision of modeling

vectors, followed by another bout of human passivity and material performance, and so on" (pg. 21-22).

As this quote demonstrates, the dance of agency consists of alternating cycles of human action, passivity, and reaction to emerging material agency. In Pickering's terms, these cycles consist of alternating "resistances" and "accommodation" by both humans and the machines being used or developed. Resistance denotes "the failure to achieve an intended capture of agency in practice", and accommodation is then "an active human strategy of response to resistance, which can include revisions to goals and intentions, as well as to the material form of the machine in question and to the human frame of gestures and social relations around it" (pg. 22).

As scientific work practices face obstacles in this dance of agency between humans and machines, articulation work is necessary (Strauss, et al., 1985; Strauss, 1988). Articulation work is the orderly accomplishment of coordination in order to change the state of a common field of work (Schmidt & Simone, 1996). Researchers organize their work in ways that enable them to align and coordinate their tasks, equipment, projects, and research goals within their social situations. Articulation work is necessary to create "do-able" research problems (Fujimura, 1987), develop infrastructures (Star, 1999), and to move across boundaries of "scientific and technical production work" to bring multiple social worlds into alignment (Suchman, 2002). An important addition to the idea of articulation work is the notion that not all actors are present and available to be organized and articulated all the time. Clarke (2005, pg. 85)

emphasizes how silences can be important indicators for unspoken assumptions, missing voices, and misaligned goals within a given situation.

The work in STS I have outlined thus far applies to my study of the metadata practices of scientists in many ways. In the following list, I pull out some of the key ideas that inform my study:

- There is not one "science" or "scientific method," but many "sciences" and "scientific methods."
- The production of texts in different forms is a prime goal in scientific research.
- "Black-boxing" of research results, methods, and machinery can occur, but in some situations does not occur.
- Research settings and activities are constituted by both human and non-human agency.
- These agencies are linked in extended chains of associations.
- Scientific practices are not atemporal, but are performed in real-time.
- Articulation work is central to coordination and development of work practices.

I follow Frohmann (2004) in applying these ideas to information issues. Frohmann analyzed the abstract idea of "information" in the intellectual culture of information studies through the lens of STS research. He critiqued the view that the "informativeness" of a document is centered in what happens in the mind of someone who understands it. Instead, he focuses on scientific practices with documents:

> "A discourse of practice replaces the question, what is the role of the scientific journal article in communicating the information scientists need in order to

advance scientific knowledge? With another: What is the role of the journal article in stabilizing scientific phenomena and contributing to the maintenance and development of research programs?" (pg. 101)

Frohmann's question directly extends to my study by replacing "journal article" with "metadata." In doing so, we get: What is the role of metadata in stabilizing scientific phenomena and contributing to the maintenance and development of research programs? In the next section, I extend Frohmann's question by using ideas from ethnomethodology.

3.2 Ethnomethodology

Ethnomethodology is an approach to studying the ways in which people make sense of their lives, and the events and interactions that take place therein (Garfinkel, 1967). In the ethnomethodological view, social situations are reflexively constituted as an ongoing sense-making activity by the individuals involved. Social facts are treated as accomplishments. As Pollner (1974) states, "[w]here others might see 'things', 'givens' or 'facts of life', the ethnomethodologist sees (or attempts to see) process: the process through which the perceivedly stable features of socially organized environments are continually created and sustained" (pg. 27). In particular, ethnomethodology seeks to understand how people account for the ongoing success of their social worlds in light of the inescapable indexicality of ordinary language and actions, as exemplified by the words "this," "you," and the phrase "I am here now" (Garfinkel, 1967, pg. 4; Agre, 1997, pg. 230).

Garfinkel (1974) coined the term "ethnomethodology" to describe his observations of juror deliberations in jury rooms. He found that jurors applied common-

sense knowledge of how evidence should be evaluated. Jurors had adequate "methods" for determining what was relevant, true and false, and what was conjecture, even though they had little formal knowledge of "legal methods." Garfinkel developed the term "ethnomethodology" as an analog to other "ethno" terms, such as "ethnobotany" or "ethnophysiology," which refer to the common-sense knowledge available to a member of society about particular subjects like plants or the human body. Ethnomethodology, then, in its initial conception, referred to the study of people's own common-sense methods and processes of understanding, creating, and sustaining social situations.

Early ethnomethodological studies illustrated how social interactions are determined not by rules or norms, but are continually enacted according to the particular occurrences of the immediate situation. Even in simple interactions, such as daily morning greetings and responses, there exist a "mass of unstated conditions which are, in various ways, tacitly oriented to by social participants" (Heritage, 1984, pg. 128). A greeting is typically tendered to initiate some kind of interaction with the expectation that it will be returned. This expectation in no way determines the subsequent response, which can take the form of a return greeting, a look of confusion, or nothing at all, among many other possibilities. Deviations from the normal response, such as a dirty look or a non-response, are interpreted as motivated by "'special', if presently undisclosed, motives" (Heritage, 1984, pg. 99). In his well-known "breaching" experiments, Garfinkel (1967) and his students disrupted ordinary scenes in order to bring background expectancies into plain view. In one exercise, students pretended to be strangers in their own family homes. In another, students asked family members to explain in detail the commonplace

expressions they used in conversation. These experiments, and the "bewilderment" that they elicited in their subjects, illustrated how everyday social interactions and conversations are constituted by continual repair and sense-making processes.

In another early study, Garfinkel (1967, Chapter 6) and his students examined the records kept by the Outpatient Psychiatric Clinic at the UCLA Medical Center. They found that many of the records were "bad," in that they were missing important details of the treatment and processing of individual patients. Moreover, they found that hospital and clinic administrators were aware that these kinds of deficient records existed. Garfinkel describes how this clinic was not unique; in fact, reporting systems were uniformly "bad" in this same way across most clinics. In explanation, Garfinkel illustrates how this phenomena of "bad" clinic records was not something that clinic personnel were "getting away with," but that instead, "the records consist of procedures and consequences of clinical activities as a medico-legal enterprise" (pg. 198). In light of this medico-legal responsibility, clinic records and folders were assembled such that they were portrayed as "having been in accord with expectations of sanctionable performances by clinicians and patients" (pg. 199). As Garfinkel describes, records were created with the expectation that readers would have knowledge of 1) the person to whom the record refers, 2) the person(s) who contributed to creating the record, 3) the organization and operating procedures of the clinic at the time the record was created, 4) a "mutual history" with other patients and clinic members, and 5) procedures for reading the record itself. Without this broad set of knowledge, readers would not be able to understand the place of the record within the clinic.

Research in ethnomethodology has ranged widely since Garfinkel's initial studies. Multiple schools of ethnomethodology developed over time, most prominently "conversation analysis" (Heritage, 1984) and "ethnomethodological studies of work" (Lynch, 1993), as well as the closely related work in "interaction analysis" (Jordan & Henderson, 1995). These schools differ in significant ways, but the unifying principle among them is the emphasis on describing "how members manage to produce and recognize contextually relevant structures of social action" (Lynch, 1993, pg. 30). In the rest of this section, I focus on how ethnomethodological research applies to my study, specifically focusing on ethnomethodological studies of work.

Ethnomethodological studies of work, also occasionally labeled "workplace studies" (though as Van House, 2004, indicates, this label also includes non-ethnomethodological studies), focus on the everyday settings in which people perform their activities. As Lynch describes, ethnomethodological studies of work take "work" to encompass activities and settings outside of purely occupational endeavors. These studies investigate "the work-specific competencies through which musicians make music together, or lawyers conduct legal arguments, in and as collaboratively produced and coordinated actions" (Lynch, 1993, pg. 114). Examples of ethnomethodological studies of work include studies of scientific discovery (Garfinkel, Lynch, & Livingston, 1981; Roth, 2009), photocopy machine repairmen (Orr, 1986), and trucking accidents (Baccus, 1986).

The ethnomethodological view has proven to be a useful approach to studying information related issues and settings. Lucy Suchman (1987), in one of the most

influential workplace studies to information-related research, developed the concept of "situated action" through a study of photocopier interface development and use. As Suchman describes, situated actions depend "in essential ways upon … material and social circumstances" (pg. 50), and are "tied…not to individual predispositions or conventional rules but to local interactions contingent on the actor's particular circumstances" (pg. 28). Suchman demonstrated the kinds of problems users encountered when performing photocopying tasks. These problems resulted from imbalances in the communication between the humans and the machine. The imbalances occurred because "the organization of human activities…is not imposed from the outside by objective structures. Instead, it is a local kind of organization, an organization continually made afresh by the situated, collaborative work of particular individuals" (Agre, 1990, pg. 372). The photocopier was not able to continually adjust its activities as a human would when breakdowns in action occurred. Whereas humans negotiate shared meanings in an emerging situation, the photocopier could not deviate from its pre-planned sets of actions, leading to user confusion and ineffectiveness in completing the specified task. Suchman's study illustrates a "movement from definitions of key concepts to investigations of the production of the activities glossed by such concepts" (Lynch, 1993, pg. 201). In her case, she did not seek to define what "planning" is, rather, she investigated how plans serve to help people orient themselves in the midst of situated actions.

Suchman's study, and her use of ethnomethodology, has been widely influential for work in Computer Supported Cooperative Work (CSCW), Human-Computer Interaction (HCI), and information systems design and evaluation (Dourish & Button,

1998; Suchman, 2007). A well-established research program in Britain, for example, calls for "ethnomethodologically informed ethnography" to be an integral part of the information system design process (Crabtree, et al., 2000; Randall, Harper, & Roucefield, 2007; Sharrock & Randall, 2004). Information system design, they argue, should include study of the "distinctiveness, the specificity, of the setting" in which such systems are to be used (Randall, Harper, & Roucefield, 2007, pg. 110). Ethnomethodology provides a lens through which information system designers can understand the "mundane and practical ways in which people make sense of what they do" (pg. 118) while using information systems as part of everyday activities. Similarly, Srinivasan (2007) uses an ethnomethodological approach to study how cultural and community ontologies, or knowledge structures, can be integrated into the design of information systems. His work reflects the ethnomethodological emphasis on the local achievement of social knowledge, by showing how information systems for particular communities are benefited by an active engagement with community members' endogenous knowledge structures.

Lynch (1993) describes how ethnomethodological studies of work should begin by taking up one or more 'epistopics,' like "discovery," "observation," "interpretation," or in the cases described above, "planning," "design," and "ontology." My study focuses on the epistopic of "description," specifically in the context of data description. In ethnomethodological study, "descriptions" are treated as actions. As Heritage (1984) states, "no description is strictly *compelled* by the state of affairs it describes. Any description is thus inherently *selective* in relation to the state of affairs it depicts.

…[C]hoices which underly any description...are all sources of clues concerning how the description is to be interpreted" (pg. 150-151, italics in original).

Summarizing the important ideas from my discussion of ethnomethodology, metadata descriptions can be viewed in the following manners:

1. Metadata are always indexical and selective.
2. Metadata are enacted according to the particular occurrences of immediate situations in a manner adequate for enabling immediate research tasks.
3. Metadata encompass negotiated shared meanings. They are created with the expectation that reader or users of the descriptions will have knowledge of how to read and interpret them.
4. Metadata are "accountable" by their creators. Scientists or other researchers are able to give "accounts" for why metadata descriptions are or are not created for their data, as well as "accounts" for the selectivity of those descriptions.

These points illustrate how analysis of the metadata creation process should view metadata description as a kind of action situated in social settings. As Heritage (1984) describes, viewing metadata description as "unavoidably an *action* which maintains, transforms, or more generally, *elaborates* its context of occurrence and, hence, as unavoidably a *temporally situated phase of a socially organized activity*" (pg. 156, italics in original), points to the need for a better understanding of what actions are involved in metadata creation and use, and how they are enacted within the scientific research processes we study.

3.3 Research questions

Looking back to the conclusion of Chapter 2, I formulated my first research question: How and where are metadata created, by whom, and for what purpose in everyday research? Applying my discussions of STS and ethnomethodological studies of work to this question adds more nuance to my investigation. Ethnomethodological studies suggest that I seek to identify the routine practices of researchers when creating metadata for their own data. How do researchers account for their metadata routines and the fact that metadata are inevitably selective? In turn, STS suggest that I attend to the roles that data collection equipment, physical objects and phenomena, and literatures play as actors in metadata processes. How are the relationships between these disparate actors articulated and organized? What silences appear in scientists' metadata practices? When do scientists create metadata and when do they not create metadata in the course of their day to day research? These questions allow me to orient myself to the role of metadata within particular practices and situations.

In the next section, I move to my second research question. Drawing on studies of social systems and communities of practice, in addition to STS and ethnomethodology, I consider social aspects of metadata practices, specifically: how scientists know and learn how to create metadata, and how personal and professional identities are at play in metadata practices.

3.3.1 Research question 2 – the social nature of metadata creation

STS and ethnomethodology studies of scientific work, while coming from distinct intellectual lineages, have similar focus on the everyday research practices of scientists

and engineers. Ribes (2006) illustrates how these two approaches complement each other when applied to studies of cyberinfrastructures and other forms of distributed scientific collaboration:

> "From the perspective of ethnomethodology ANT provides a 'gloss', skipping over vast swaths in the practical details of doing. However, ANT studies tend to cover more ground, can tie together vastly heterogeneous methods and are able to focus on longer term consequences than ethnomethodological studies" (pg. 56).

Actors, in the ANT sense, build contexts as ongoing situated actions. By "contexts," I refer to lab settings and activities, field excursions, professional careers, data organization methods, and an unending chain of others. In this section, I outline how widening our view to understand these contexts as embedded in social systems brings about additional research questions for my study.

Members of cyberinfrastructure projects are constituents of numerous social systems and communities, including universities, professional societies, academic units, networks of funders, legal regimes, student cohorts, research teams, and of course the project itself. Such social systems have many properties (Weissman, 2000). They can be loosely or tightly bound, and can be aggregated or nested in hierarchies. Social systems can also be dependent on each other and in some cases in competition or indifferent to each other. Scientists and researchers know and achieve their ongoing circumstances through their everyday situational activities within these systems, with knowledge being distributed amongst social settings (Hutchins, 1996). Button and Sharrock (1998), for example, used the ethnomethodological approach to study the multiple systems in

collaborative engineering workplaces. They found that many of the problems that engineers encountered in their work were not related to technical or engineering details of their tasks, but instead arose "from the intricacies, complexities and intractability of organizing projects" (pg. 91). Furthermore, these organizational problems were understood by the engineers to contribute to the outcomes of their engineering projects: "It is a routine, quite taken-for-granted feature of the work of the engineers we studied that the success and failure of their engineering ventures, and the nature of the problems which will contribute to such outcomes, will frequently depend on the way the work is organized" (pg. 97). In this example, we see the engineers working within multiple overlapping systems, the company, the engineering profession, and the workplace organization, among others. The dependencies between these systems – the engineering work is intimately tied to the workplace organization, which produces the company output – contribute to the accounts that engineers give of their work.

My research subjects are members of the systems encompassed by academic and university settings. Academic researchers have particular career trajectories, whether they are ecologists, computer scientists, seismologists, or physicists. As Traweek (1988) describes, students initially learn formalized core knowledge from textbooks and journal articles, and take lab classes to learn classic experiments and methodologies. As they continue in their schooling, they learn how to explore new research areas and are given more responsibility in conducting and overseeing their research. They are also expected to contribute their own ideas to the research process. More advanced students begin to rely more on interpersonal communication as the prime means of learning. At this point

in their career, the ability to seek out and talk to the correct people about a specific issue is very important, both in learning and in promoting their own research and ideas. Successful researchers move on to positions in academic or research institutions, now as researchers with the knowledge and means to design, conduct, and produce their own independent research projects.

Within this career trajectory, academic researchers take many unique paths (Turnbull, 2007). Researchers bring knowledge from their personal lives and social interactions to bear on research projects. Additionally, researchers often are involved in multiple projects simultaneously, applying knowledge and practices from one project to another. Becoming part of academic scientific communities also involves learning how to navigate through institutionalized networks of administrators, funders, and public interest groups. In all these ways, researchers must learn to deal with incommensurable knowledge structures that exist in multiple social systems, both from inside the Western academic tradition and outside (Turnbull, 2009).

Along these career paths, researchers learn as "legitimate peripheral participants" (Lave & Wenger, 1991). Individuals initially participate in social and research activities by performing simple but essential tasks, such as cleaning beakers, preparing specimens, and carrying field equipment. Through this, they learn daily practices and become more integrated into the research communities. In participating in these essential ways, students become part of the social circles within larger systems and communities, and over time learn to perform more complicated tasks, like setting up and running equipment, and conducting the experiments themselves. As students move to the center

of their individual research communities, they become more accountable for their work. Wenger (1998) describes how "accountability" in this sense refers to learning about "what matters and what does not, what is important and why it is important, what to do and not to do, what to pay attention to and what to ignore, what to talk about and what to leave unsaid, what to justify and what to take for granted, what to display and what to withhold, when actions and artifacts are good enough and when they need improvement or refinement" (pg. 81). Accountability exists at both individual and organizational levels (Yakel, 2001), and always exists in a social system. As Wenger describes, research practices exist within a regime of "mutual accountability," wherein interpretations of work competency are negotiated among the members of a community. These negotiations are central to the learning experience, and enable individual researchers to understand in what situations the boundaries between "competent" and "incompetent" practices can be pushed. "Being able to make distinctions between reified standards and competent engagement in practice is an important aspect of becoming an experienced member" (Wenger, 1998, pg. 82). Understanding when, how, and why particular ideas or techniques can be applied to boundary cases is fundamental to the development of expert knowledge (Day, 2005).

Viewing legitimate peripheral participation as having a primary role in the learning process "decenters" the unit of analysis from internal individual cognitive learning approaches to social learning practices that are co-constituted by individual identity, knowing, and social membership. Trace (2007), using an ethnomethodological approach, illustrates how document work – learning how to create and use documents – is

often “hidden work” within educational settings. Documents are at the center of educational tasks, but are themselves rarely the focus of the educational process. Shankar (2002; 2007; 2009) studied data management and recordkeeping in an academic laboratory. As she describes, junior scientists learned proper recordkeeping and documentation through legitimate peripheral participation: “becoming an active research scientist requires that the individual mesh his/her personal ways of working with the modes of work demanded of his profession – work that is rich, embodied, often tacit, and as such often anxiety-producing” (Shankar, 2009, pg. 163). Shankar’s study focused on paper notebooks, emphasizing the ongoing importance of paper records even as digital technologies grow in ubiquity. Her study was also focused on largely individual work of scientists at the beginning of their careers, in contrast to the highly digital and collaborative CENS environment. Researchers who utilize new information and communication technologies run into additional complications in the process of meshing personal and individual modes of work, as they might find themselves in the middle of the “traditional” professional work practices of (Lamb & Davidson, 2005).

This discussion leads to my next main research question, which centers on the social nature of the metadata creation process. Researchers learn through social interactions and participation when, how, and why - as well as when not, how not, and why not - to create, use, and manage metadata:

- Research Question 2 - How are metadata creation tasks learned and parceled out in research groups?

Important sub-questions relate to how team members make sense of metadata descriptions created by others in their team. What counts as an adequate metadata description, and in what situations?

Finally, how are the expectations and norms of different social systems and communities reflected in metadata practices? In the next section, I develop my third and final research question, which brings together the multiple themes in my discussion so far to resolve the major tension around metadata in cyberinfrastructure data systems.

3.3.2 Research question 3 – moving from local to global

Putting research questions 1 and 2 together - viewing metadata creation as both inescapably tied to local practices and inescapably embedded within larger systems – I arrive at my next main research question:

- RQ3- How do local metadata practices translate to the creation of metadata for shared community repositories?

As I described in Section 2.2, cyberinfrastructure data systems rely on working scientists to create metadata in standardized ways. The day-to-day metadata practices of working scientists, however, are situated within research cultures that may or may not be based around data sharing. What are the practical problems for scientists in creating metadata for a shared repository? Heading off practical problems before they appear will significantly improve the chances that cyberinfrastructure systems are successful. In investigating Research Question 3, I can reflect on how the day-to-day metadata creation practices of scientists compare with the ways in which they create metadata for a shared repository.

3.4 Summary

In this chapter, I review literature in STS, ethnomethodology, and social systems and communities of practice theory as background for my study. CENS researchers are members of multiple social and institutional systems, and develop research practices, including metadata practices, within those systems. The literature I surveyed in this chapter indicates that metadata practices exist within actor-networks. Both humans and non-humans impact metadata practices. Additionally, metadata practices will be inevitably selective, with researchers creating forms of documentation that allow them to be accountable to members of their own communities of practice. In the next chapter, I outline my research methods, drawing on this theoretical background.

4. METHODS

To study these questions, I conducted a multi-sited ethnography using a combination of participant observation and semi-structured interviews. My research subjects were four CENS research groups: 1) a soil ecology group, 2) an aquatic biology group, 3) an environmental science group 4) a seismic sensing group. In studying these groups together, I conducted a multi-sited ethnography. Multi-sited ethnography "moves out from the single sites and local situations of conventional ethnographic research designs to examine the circulation of cultural meanings, objects, and identities in diffuse time-space" (Marcus, 1995, pg. 96). Despite all these groups being part of a single research center, CENS, I can characterize this study as multi-sited because these groups are all based at different institutions, and all groups perform research activities in different real-world field settings. As an additional characteristic of multi-sited research, I use diverse sources of data, including ethnographic observation, interviews, documents, photos, and web sites, allowing me to study practices, discourses, and situations (Clarke, 2005).

I selected these groups for my study because they represent the four main scientific application areas for CENS sensing technologies. All four groups are also involved in science and technology research concurrently; they are using developmental CENS sensing technologies to conduct scientific investigations. They provide

comparable cases of small-scale field-based scientific research, while crossing disciplinary and institutional boundaries. The first three groups range across the field-based life sciences – ecology, biology, and environmental science – while the seismic sensing group provides an interesting counterpoint. The life sciences are markedly diverse in data types, and correspondingly have very low levels of standardization in data and metadata formats and standards (Bowker, 2000). Seismology, while also being a field-based science, has well established data formats (Ahern, 2002), providing a contrasting case for my study. I examined how researchers in these groups create, organize, manage, use, share, and archive metadata in the course of their research activities in both lab and field settings.

4.1 Ethnographic study of metadata creation in day-to-day practice

This section lays out my ethnographic methodology for the study of the everyday metadata creation practices of scientific researchers. I used multi-faceted data collection methods: participant observation, semi-structured interviews, and collecting supplementary documents, including published papers, data sets, documents, web sites, and emails created and used by my research subjects. These data sources follow those typically used in qualitative field studies: direct experience, social action, talk, and supplementary data (archival records, physical traces, and photographic data) (Lofland, et al. 2006).

The first data collection method, participant observation, consisted of my direct experiences and observations of the social action in the everyday research settings of my subjects. Participant observational methods allow for a “many-sided and situationally

appropriate relationship with a human association in its natural setting for the purpose of developing a social scientific understanding of that association" (Lofland, et al. 2006, pg. 17). My participant observation took the form of trips to field settings in which CENS researchers installed sensing equipment, collected data by hand and via sensors, and in some cases removed equipment from a field site. I also conducted observation in lab settings, visiting the desks and offices of individual researchers, and observing as researchers analyzed samples via lab machinery and computers. As the term "participant observation" suggests, I took part in some of the activities I observed, with my participation ranging from digging holes and carrying equipment to recording sensor values in a notebook for my subjects. To record my observations, I took notes at the time events were occurring when possible, but when my own participation in activities precluded immediate note-taking I wrote my field notes down after events took place. My observations were either recorded in a field notebook and transcribed into a computer, or directly recorded via a computer. My field notes consist of accounts of the lab and field activities of CENS researchers, my discussions with CENS researchers and their discussions amongst each other, and my own thoughts and ideas relating to the activities in which I engaged. In addition to narratives of actions and discussions, I tried to note "member terms," the language that researchers in these four CENS research groups use to describe their own work, their categories, and meanings (Emerson, Fretz, & Shaw, 1995). When possible, I also took pictures of the research settings and activities in which my participant observation took place. After downloading the pictures off of my camera, I noted in a separate document where the pictures were taken, and what they are showing.

My participant observation consisted of 16 trips to lab or field settings, encompassing approximately 200 hours of observations. These trips ranged from two-hour excursions to a lab or field site to observe a particular activity, to a three-week trip to Peru to take part in installations of CENS seismic sensing stations. Most trips were one or two days in length or shorter. My field work in Peru took place in the summer of 2008, as did my initial field trips with the other three teams. The rest of my field work took place in 2010 and 2011. I created an individual document of field notes for each of these trips. My compiled field notes total 88 pages of single-spaced typed text, and I took over 125 pictures for use in my study, including all of those shown below.

Building from my participant observations, I conducted semi-structured interviews with 14 CENS researchers that focused on specific aspects of their data and metadata practices. Interviewees were drawn from each research team in my study, and my interview population was built via snowball sampling by asking each interviewee for names of other individuals involved in their research. I developed a distinct interview protocol for each interview. I carried some interview questions over from interviewee to interviewee, but I also developed new questions for each interviewee. Appendix I and II shows examples of two distinct interview protocols. I audio-recorded my interviews and either transcribed them myself or had them transcribed by a paid transcription service. In order to correct errors and ensure transcript accuracy, I proof-read transcripts produced by the paid service by listening to the audio-recording while reading the written transcript. My interviews averaged 43 minutes in length, with a range of 25-84 minutes. Transcription totaled 152 text pages.

As supplementary data, I collected a variety of documents that relate to my subjects' research. First, I collected published papers relating to the projects included in my study. To collect published papers, I examined the web sites of projects and individual researchers for links or citations to publications. Through this method, I collected 33 total published papers: ten each from the aquatic biology and soil ecology projects, eight from the seismology project, and five from the environmental science project. Each published paper I collected was authored or co-authored by at least one researcher who was included in my observation or interview samples, and was directly related to the projects included in my study. All collected papers were published between 2006 and 2010. In addition to published papers, I collected 28 documents that were produced by my research subjects in the course of their day-to-day activities, an average of seven documents per case study. These included data files in multiple formats, equipment outputs, word documents, lists of equipment, diagrams, and screenshots of software interfaces. I was also able to access project web sites, wikis, and email lists where they were available.

Finally, as a member of CENS myself since 2006, my study is informed by regular interaction with CENS researchers, both those directly included in my study and those that are not. In early 2007 I was assigned a desk in the CENS facility, a large space built out from an existing math and science building on the UCLA campus. I used my desk in the CENS facility as my regular workspace for the duration of my time in CENS. The facility serves as a lab and workspace for graduate students, research staff, and high school and undergraduate interns. Most of the students who work in the CENS facility on

a daily basis are computer science students, as CENS students from application science disciplines typically have workspaces in the laboratories of their advisors. Of the four CENS projects I focus on in my study, only the seismic team had members working in the CENS facility on a daily basis. The other three projects in my study were based in other locations either on the UCLA campus or in a partner institution (as were some other parts of the seismic team). Thus, while the CENS facility enabled me to participate in informal gatherings and discussions, the population based within the CENS facility largely differed from my study population. In addition to the daily work in the CENS facility, I took part in numerous formal CENS gatherings, such as research reviews and retreats, weekly CENS research seminars and coffee hours, and other research presentation events, which brought together researchers from all CENS projects.

The particular topics of my field notes, interviews, and supplementary document collection were continually adjusted and updated during the course of the study through "theoretical sampling" of emerging themes and issues that arise (Cerwonka & Malkki, 2007). By theoretical sampling, I refer to the process of jointly collecting, coding, and analyzing my data, and using this process to decide what data to collect next and where to find them. Theoretical sampling allowed me to develop theoretical considerations as they emerged (Glaser & Strauss, 1967; Clarke, 2005). As recommended by ANT, this involved following the links of associations to actors outside my immediate study, such as investigating the manufacturers of my subjects' data collection equipment. Drawing from ethnomethodology, theoretical sampling involved investigating how my research participants adapt their metadata practices to new students, problems, and field sites.

To analyze my field notes, interview transcripts, and supplementary documents, I developed a set of analytical codes in order to identify and develop emerging themes. My initial codebook was based on my research questions (given in Sections 3.3 – 3.5), and my codes were refined iteratively, as I developed themes and produced preliminary analyses in the form of conference papers (Mayernik, 2010; Mayernik, Batcheller, & Borgman, 2011) and analytical memos (Glaser & Strauss, 1967; Emerson, Fretz, & Shaw, 1995). I used the NVivo 9 qualitative data analysis software to apply the analytical codes to my data, and to identify particular passages of interest. In analyzing direct interview quotes, I employed what Alasuutari (1995) dubbed the "factist" and "specimen" perspectives. As Talja (1999) developed further, the "factist" perspective uses interview quotes "to find out about the actual behavior or attitudes of the participants," whereas in the "specimen" perspective interview quotes "are not descriptions of the object of research; they *are* the object of research" (Talja, 1999, pg. 471-472, italics in original). In other words, I use interview quotes in two ways, 1) as evidence of the practices and phenomena that the quotes are describing, and 2) as evidence of the ways that language can reveal norms, debates, and alternate viewpoints of the worlds in which my interviewees are a part.

Combining participant observation, interviews, and document analysis of personal documents and formal publications allows for a multi-faceted view of CENS researchers' metadata practices. In the next section, I describe the outcomes of my ethnographic work, providing a series of vignettes of lab and field work that span my four case studies. These

provide a picture of day-to-day metadata practices, and illustrate how these practices exist in social research settings.

4.2 Study of metadata creation for a community data/metadata repository

In this section, I describe our method for studying how CENS researchers approached the task of contributing to a community metadata repository. How do researchers' methods for creating metadata in their routine lab and field metadata creation practices compare with how they create metadata using an unfamiliar set of description fields? I investigated this comparison using two methods: 1) observing my subjects as they created descriptions for the community repository, and 2) interviewing them about their experiences in performing this task and about the results of the task, that is, the metadata descriptions they have created.

The CENS metadata repository has been implemented as a module in the CENS online annual report submission system. CENS must submit an annual report to the NSF every year. CENS researchers report on research progress, publications produced, research partnerships, and, using the CENS metadata repository module, data sets. The user tests for my study were conducted using a prototype of the 2011 version of the metadata repository. We used a prototype because the official version was under development, and thus not available for use. The prototype was designed to mimic the functionality of the official version as closely as possible. Figure 4.1 and 4.2 show portions of the interface for the official 2011 CENS Metadata Repository and the prototype used for my user tests respectively.

Figure 4.1 - Screenshot of the metadata submission form that was included in the official 2011 CENS annual reporting system. Six metadata elements are shown.

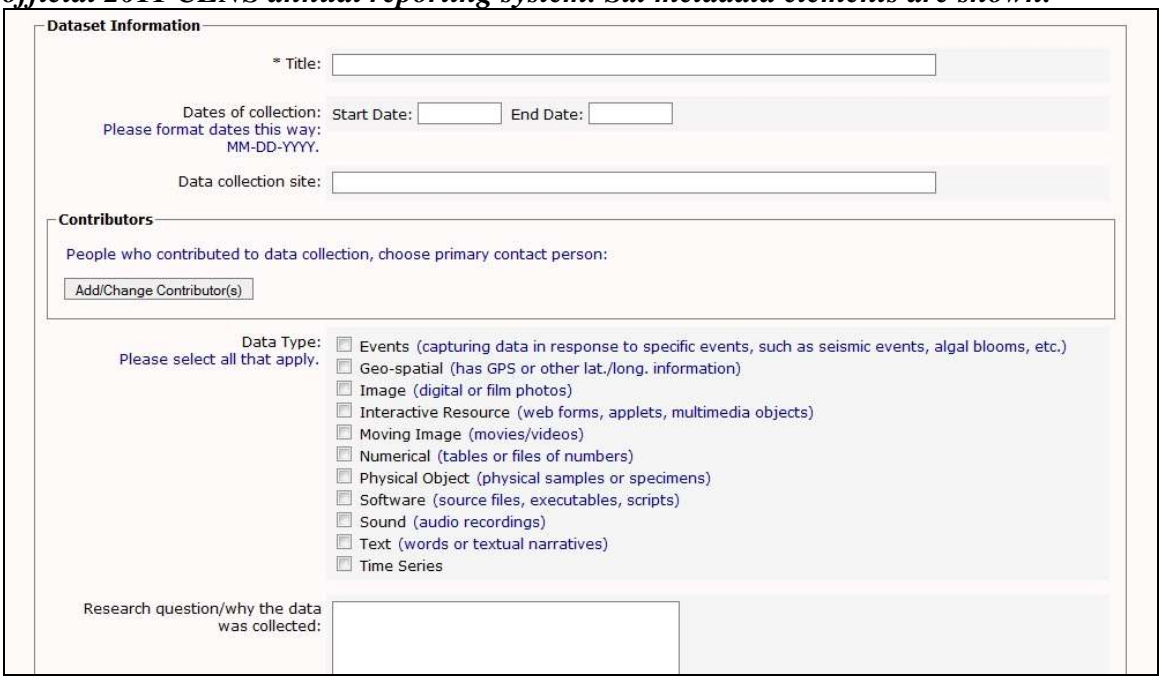
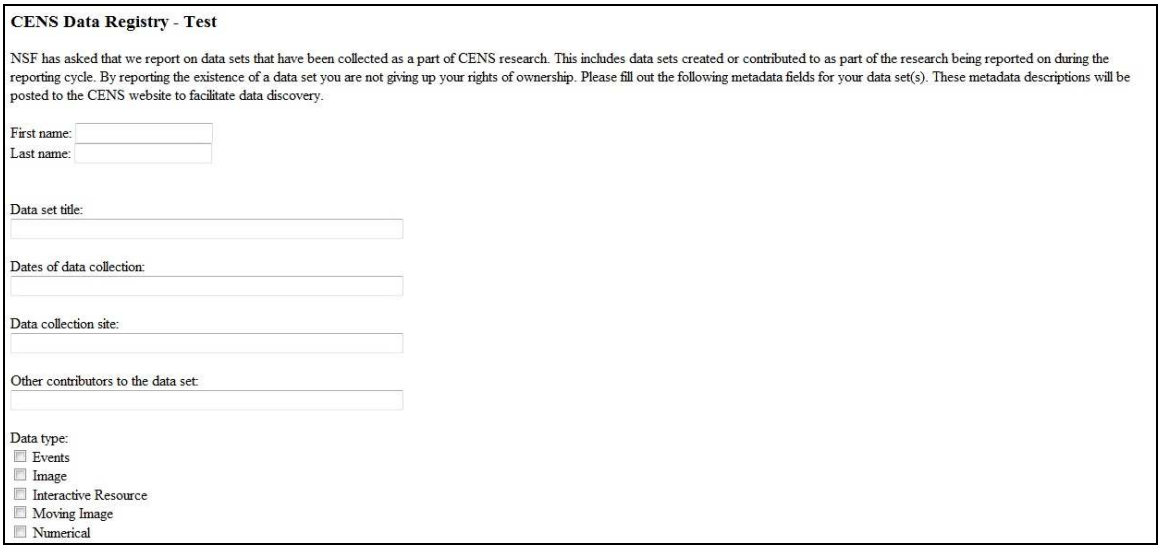


Figure 4.2 - Screenshot of the prototype metadata submission form used in this study. Five metadata elements are shown.



In these user tests, we asked CENS researchers to describe their data using the metadata fields listed in Table 4.1. These metadata fields were based on the Dublin Core

metadata set (DCMI, 2009), but were customized to use terminology more familiar to CENS researchers. Table 4.1 illustrates the field terminology, and how our terms relate to the Dublin Core schema. All of these fields except two were presented to the user as free-text entries. The "Data Permission Level" field and the "Data type" field were presented as pre-defined checklists. The options for the "Data Permission Level" field were taken from the Creative Commons list of open source licensing options (Creative Commons, 2011). The option list for the "data type" field were taken from the list of data types given in the DCMI type vocabulary (DCMI, 2010), with the following customization. We dropped the Dublin Core type "dataset" because the term "dataset" was too generic of a descriptor for our task, and we added "numerical" as an option because we knew from our previous work in CENS that a lot of CENS data were numerical in form.

Table 4.1 - Metadata Fields Used in the CENS Metadata Registry

Data description fields	Dublin Core elements*
1. Dataset Title	title
2. Dates of data collection	date
3. Data collection site	coverage
4. Primary contact person	creator
5. Other contributors	contributor
6. Data type	type
7. Research question (why collected)	description
8. Variables collected	description
9. Process and equipment used for collection	description
10. Data format	format
11. Data sharing permission level	rights
12. Funding source	source
13. Keywords	subject
14. Location of the data (URL)	identifier

*The Dublin Core elements "language" and "publisher" were not used. The "relation" field was not used in these tests, but was used in the official 2011 CENS Annual Report metadata repository to collect related publications.

Testers were solicited by contacting individuals who had submitted a CENS Annual Report in 2010. Twelve solicitations were sent out to members of my four case studies, and thirteen solicitations were sent out to members of other CENS projects. Other CENS projects were included in this portion of the study because the tests were also intended to give feedback to the CENS administration and the software development team on the CENS Annual Report system functionality. In total, eleven CENS researchers participated in the user tests. Six testers were from the projects in the ethnographic portion of my study: two from the seismology team, three from the aquatic biology team, and one from the soil ecology team. The other five testers were from other CENS projects, four from computer science and one from environmental engineering. Ten of the testers were students, with the eleventh tester being a research staff member. All tests except one took place in the testers' labs or offices, with the exception taking place at my desk in the CENS building. Tests took place between December, 2010, and April, 2011.

Nine of the tests were conducted with two test administrators present, myself and Jillian Wallis, both students in the UCLA Department of Information Studies and members of CENS. In the other two tests, I was the lone test administrator. Our test protocol is provided in Appendix III. We developed the protocol following the usability engineering lifecycle method (Mayhew, 1999). We described how our goal was to collect descriptions of CENS data, both to report to the NSF and to put on the CENS web site as a data discovery tool. We asked testers to create metadata for the main data that they

were using in their day-to-day research with those goals in mind. Specifically, we presented each tester with the title of the project that they submitted to the 2010 CENS Annual Report, and asked that they create metadata descriptions for data that were associated with that project. If users said that they were not working on that project any more, we asked them to describe the primary data they were using in their current project. We used a "talk-aloud" protocol, asking the testers to describe what they were thinking and writing as they completed the metadata descriptions. During the test, we observed and took notes of the researchers' activities and comments as they completed the task. All tests were also audio-recorded, and seven tests were video-recorded. Because not all tests were video-recorded (with some videos cut short by battery drain), the videos were not systematically analyzed. I used the videos as references if the audio-recording was unclear about a particular portion of a test (Suchman & Trigg, 1992). At the completion of the tests, we asked the tester to submit their response, upon which our form generated an output file with those responses. We collected these output files for our analysis. For one tester, the output file was collected from the official 2011 CENS annual report submission because he did not click the "Submit" button and thus we were not able to save the response file from his test. His responses to his official CENS Metadata Registry submission were checked against his test transcript for confirmation of his test responses.

After the testers completed the test, we performed targeted interviews about their experience in performing the task. Interview questions included asking the researchers why they chose to describe the particular data set that they chose, which metadata fields they felt were the most and least useful in describing their data, what additional fields

might be necessary, and what benefits (if any) they feel that they receive from creating this metadata, among other questions.

Thus, my analysis in this portion of the study focused on the testers' actual responses to our form, and the transcriptions of audio-recordings of the tests and post-test interviews. In the next section, I describe the results of these user tests.

5. RESULTS I – VIGNETTES OF FIELD STUDIES

In this chapter, I provide eight vignettes of my ethnographic case studies. These vignettes give descriptions of typical day-to-day research activities in seismology, aquatic biology, soil ecology, and environmental science. I also use these vignettes to illustrate typical metadata activities: what metadata people collect, when, how, and why data are documented. These narratives are composites. The activities and discussions depicted have been re-ordered, timelines have been changed, and individuals have in some cases been combined. The depicted events took place in some cases over a number of separate days. Names of people and places have been anonymized to protect identities. Exact quotes are provided in italics.

Following each vignette, I provide commentary on specific issues of particular interest: How do metadata creation and use fit into individual research practices and collaborative research settings? How are metadata practices learned within a team and within a discipline? Why are certain types of metadata give more priority than others? These commentaries contextualize the vignettes within larger disciplinary and institutional settings.

I also use the vignettes and commentaries to identify all that could be considered data and metadata within the research practices of my subjects. The vignettes illustrate data and metadata from the perspective of an observer within everyday research

situations and practices, whereas the commentaries provide my perspective on data and metadata as an outside observer trained as an information professional. Thus, in the commentaries I identify "metadata" using the notion of metadata that I introduced in Chapter 2: documentation, descriptions, and annotations created and used to manage, discover, access, use, share, and preserve informational resources, or as Greenberg (2005) states, "data attributes that describe, provide context, indicate the quality, or document other object (or data) characteristics" (pg. 20). I also try to note where the distinction between "data" and "metadata" is clear and where it is more ambiguous.

5.1 Seismology

Seismological sensing projects can be organized around active or passive field deployments. In active deployments, sensors are deployed to capture earth motions related to man-made perturbations, which are typically caused by detonating explosive charges. Researchers who use passive deployments, in contrast, leave sensors in place for long periods of time, with the goal of recording seismic events as they naturally occur. The CENS seismic deployments discussed here are passive deployments. In the first CENS seismic project, researchers deployed 50 radio-linked and 50 stand-alone seismic stations across Mexico for approximately two years, from 2005 to 2007. Following this initial deployment, the seismic sensing stations were moved to Southern Peru, where they were installed in 2008. In both projects, researchers installed seismic sensors and wireless communication equipment at approximately 5-km intervals from the Pacific Ocean into the continents, a range of several hundred kilometers. Over the course of the Mexico and Peru projects, 12-15 researchers have participated in different ways, including students,

staff, and faculty. CENS' innovation in relation to seismic sensing is the use of wireless radio communication systems that allow researchers to access sensing stations remotely, and allow sensor readings to be transferred from seismic sensors in the field back to lab databases.

5.1.1 Vignette 1 – Seismic deployment in Peru

Josh, Luke, and I prepare to leave the CENS apartment in Peru at 8:00 in the morning. Josh is a recent Ph.D. graduate in seismology, while Luke is a student from a partner institution. Christian, a Peruvian engineer who helps with the project (and is a critical source of local know-how), arrives, and we wait for the driver to arrive with his truck. The driver is also Peruvian. We drive about 20-30 minutes across town to get to a local geologic institution where the CENS team bases its Peruvian project. This institution is on top of a hill, the CENS team has a small building on the back of the lot where they keep all of their equipment. The building has six rooms filled with work benches, cables, boxes, tools, etc. They also have an office in the main building on the lot, which is just up the hill from their storage space. The office has a computer with internet access, and a big set of maps on the wall of the transect (See Figure 5.1). They also have antennas/radio on the roof, and a seismometer down in the basement of the office alongside vintage seismic equipment that the Peruvians still use.

Figure 5.1 – Seismic office in Peru

Josh, Luke, and Christian start loading equipment of various kinds into the truck. Luke is staying behind today to organize the equipment in the storage building, and prepare specific pieces to be deployed later in the week. Josh, Christian, and myself are going to perform a site installation. Once the truck is loaded, we drive across town to "site 33," which is a local elementary school. Each deployment site in the project has a designated number, between one and fifty, with the first site near the Pacific Ocean, and the 50th site located about 200 km inland. Josh and Christian go inside the school to check out the school grounds and get permission to do the installation. They come back, and we unload and start carrying the necessary gear inside.

The actual installation site is to be located in an area that is behind the school, between the school and an outer wall of the school premises. We find a good spot about

halfway down wall, and begin the installation. The first job is to dig two holes, one for the sensor and one for a metal box that holds other site electronics. First, we dig a hole for the metal box. The hole needs to be rectangular in shape and about two feet deep, deep enough to cover the box up to the lid, but still allowing the lid to open. About a foot away from the first hole, we dig another hole for the seismometer. The sensor hole is round, about 2-3 feet in diameter, and 2-3 feet deep (Josh says that the rough specification is 80cm deep). Christian and Josh do most of the digging for the box hole; I work on the round hole for the sensor. About halfway through the digging process, I notice that I have blisters on my hands, because I am not wearing any gloves. The digging is also going slowly because we need to dig around rocks that are not coming out of the ground easily. We dig the seismometer hole a bit bigger to avoid the rocks. I get tired pretty quickly, probably some combination of the elevation (8,000+ feet), heat, dryness, and my not being used to digging. Eventually we get the holes dug. Once we have the holes dug, Christian and Josh mix concrete in a bin. Christian puts cement in the bottom of the sensor hole and uses a cylindrical piece of cardboard to create a platform for the sensor at the bottom of the hole. The sensor will be placed on the platform. When done laying concrete, he places a plastic tube around the platform so that nothing will fall on the cement while it dries. The tube is about two feet high, and about one and a half feet in diameter. The sensor will be placed on the cement platform inside the plastic tube once the concrete dries.

We begin installing the rest of the equipment into the large metal box. This involves putting in the “black box,” a smallish radio unit with flash memory card for

local data storage, and the digitizer, which takes the signal from the seismometer and turns it into a digital signal, and the seismometer itself. (See Figure 5.2)

Figure 5.2 – Seismic station installation in Peru

When all the equipment is hooked up, Josh shows me how to log into the "black box" using the field laptop. This involves starting up Linux, navigating to the correct directory, and running a few commands. The process confuses me at first, because many of their file directories and scripts have "Mexico" in the file name. Josh says that the names of the software reflect how they are using much of the same electronic setup as the Mexico deployment. (A member of the team told me later that they attempted to change the directory names to "Peru" at one point, but the change caused the software to crash,

and so was aborted.) After logging into the "black box," we can see that the sensor is taking readings by running a particular command. The sensors work by measuring the movement of an internal mass in all three dimensions and translating those movements into acceleration measurements. These acceleration measurements are the data for this deployment.

Josh and Christian then go up on the roof to install the solar panel and radio antenna. Josh throws down the power cable from the roof to the car, because they need to hook the inverter to the car battery in order to get power for the drill. I have some confusion as to how to hook up the inverter, and am worried that I will cross the cables because the plus and minus pins on the inverter are very close together. Josh suggests that I use a piece a plastic to keep them apart, which I do, and eventually get the inverter turned on.

At that point Josh and Christian start drilling on the roof. After a period of time, Josh asks me to attach the solar panel cable to the sensor battery, wiring them together with a charge controller. This requires me to attach the plus and minus wires from the solar panel and the battery to the appropriate pins on the charge controller. Josh tosses down a voltmeter, and I test the leads from the solar panel. From the roof Josh asks me if I know how to use the voltmeter, because he has had somebody attach the wires backwards before. I check with him to make sure I am doing it correctly, and when sure that I am, I hook up the wires. While on the roof, Josh and Christian also adjust the solar panel and antenna mast to face the correct directions: the solar panel slightly tilted toward the north to maximize sun exposure, and the antenna pointed toward the next station in

the wireless routing line. Josh shouts down that they had a lot of trouble drilling into the concrete on the roof, and that he melted a drill bit in the process. When they come back down, Josh shows me the melted drill bit.

Josh then attaches the GPS unit to the digitizer. When he connects to the "black box" via the computer however, he says he is having trouble getting a fix from the GPS unit. He thinks the wall might be getting in the way. Earlier we had tried to run the GPS cable down from the roof, but it was not long enough to reach the metal box at the bottom of the wall. So Josh decides to leave the GPS unit in the metal box. He makes a note in his notebook that somebody will need to re-visit this site with a longer GPS cable so that the GPS unit can be moved to the roof.

When we are finished with the installation, we drive back to the Peruvian seismic institution where equipment is stored. During the trip, Josh, sitting in the passenger seat, types notes into the field laptop. We unload the equipment, and then go back to the apartment. In the evening, Josh sends an email to the CENS seismic mailing list with an outline of our day's work. His email consists of the notes he typed up while we were driving back from the field. Josh's email about today's work includes eight sentences about our work at site 33. The first of these sentences notes that we finished the installation of site 33. The second sentence notes that the next person to come down to Peru should bring longer GPS cables. The other six sentences in the email are dedicated to the sub-par drill bits.

The next morning, Josh, Luke, Kyle, Christian, and Katherine and I sit and talk for about two hours before leaving the apartment. Kyle, a CENS staff engineer, and

Katherine, a student from a partner institution, arrived in Peru the night before. Most of the talk is between Josh and Kyle. They run through the entire line, site by site, describing the current situation at each site. This discussion is mostly prompted by Kyle. (I thought that we were sitting there talking because the driver and truck had not yet shown up. But, in fact the truck was waiting for us outside the apartment the whole time.) Every once in a while Luke chips in, or they bring Christian into the discussion, but it is mostly Josh and Kyle. The do not refer to documents during this discussion; it is all out of their memory. For example, Josh would say, "Site 33 is at the elementary school, and the sensor, digitizer and box were put in yesterday by me, Matt, and Christian." They also go over the locations that are more important than others, with "importance" mostly decided by the wireless communication routes. During this discussion, Kyle describes his desire to have a standard log sheet for the seismometers, digitizers, and black boxes installed at each station, so that all of the associated serial numbers will be recorded in the field. He also wants pictures of the installations once they are finished.

5.1.2 Commentary on Vignette 1

This narrative illustrates the oral and informal nature of field work. Members of the deployment team build knowledge of the current state of field stations through being there and through talking. This knowledge is embodied knowledge. Researchers learn through physical experiences how holes should be dug, hardware should be wired, antennas should be harnessed, and sensors should be oriented. Knowledge is brought from past projects to new experiences. The three principals in this narrative, Josh, Luke, and Kyle, were all part of the CENS seismic Mexico project, and thus bring considerable

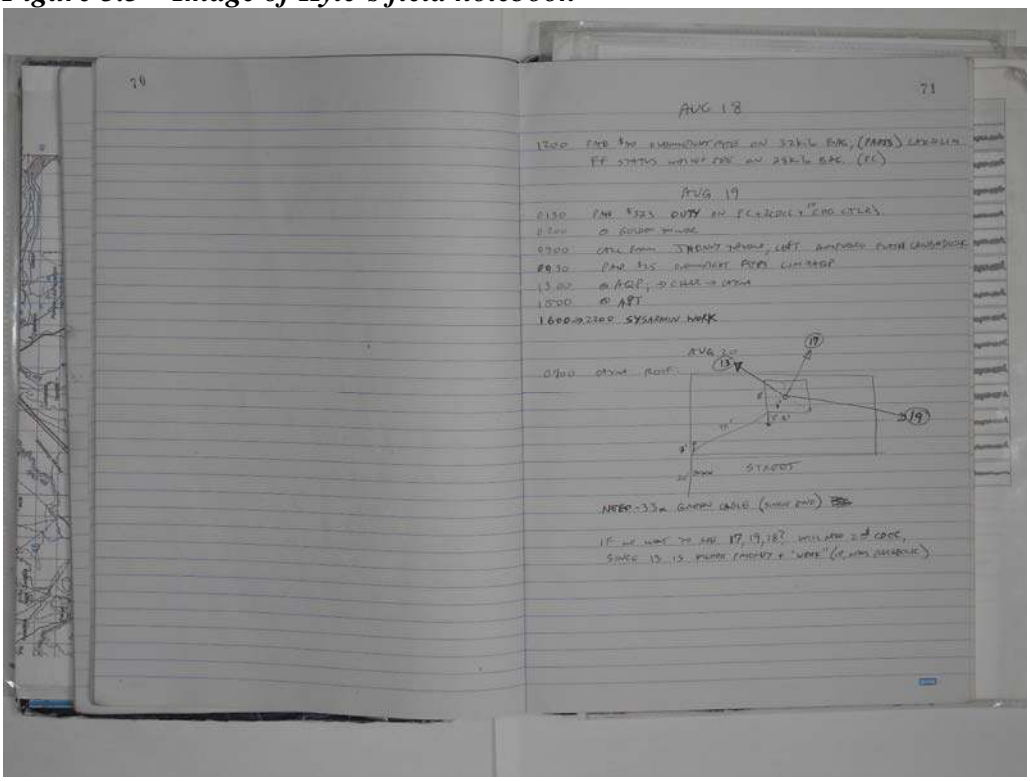
field experience to bear on their work in Peru. Josh and Luke's primary role in the Peru project, in fact, was to lead the on-the-ground field work during the initial sensor installations. Neither expected to use the Peru sensor data themselves. Luke had no intention using the Peru data because his main research interest laid elsewhere; Josh told me that he is low on the priority list for use of the Peru data because he primarily uses data from the Mexico deployment. He said that he might look for one particular type of seismic activity in the Peru data if no one else on the project does, but he will have to ask if anyone else is looking for it first.

As this vignette also notes, personnel changes are common among field teams. Students rotate in and out every few weeks, with varying degrees of overlap. During my three weeks in Peru, including Josh and Luke, I overlapped with five students. Two of these students were expecting to be primary users of the Peru data, in contrast to Josh and Luke. Bringing a new person up to speed, even somebody who has been in the field many times before, like Kyle, is an intensively oral process. The two hour morning discussion described towards the end of the vignette vividly exemplifies how talking is an integral part of team-based field work. Researchers talk about field sites, machines, other researchers, Peruvian politics, and money. They talk about future deployment plans, past activities, and problems encountered and solved or side-stepped. Talk is a central part of their work (Orr, 1996).

The daily email logs supplement this talk. They are written by whoever is nominally leading the field work at the moment. The email logs notify team members who are not in the field of salient issues as they arise, and outline plans for upcoming

work. They also often include pictures, maps, and diagrams. These email logs thus serve as orienting devices, allowing remote team members to keep in touch with the day-to-day deployment activities and routines, and to orient themselves to their surroundings when it is their turn to take part in the field work. The styles of the emails are highly personal. For example, Josh's daily log notes are typically brief, whereas Kyle is known for very structured and detailed notes. Kyle showed me how he writes his field notes into a paper notebook and then types them up into email form each evening (see Figure 5.3). Not everything noted in the notebook gets transferred to the daily emails. Diagrams written in field notes, such as the illustration of the antenna orientations shown in Figure 5.3, are rarely included in emails, though researchers occasionally create a digital version and sent it out as a separate document.

Figure 5.3 – Image of Kyle's field notebook



The utility of these emails varies from person to person on the team, with team members who are closer to the field work typically valuing them more. A PI on the project stated, "*frankly we don't do anything with those notes unless... They're used to jog the memories for about a month and [when] something needs to be fixed*." In contrast, Kyle, the primary field engineer on the project, uses the email notes as searchable archive of deployment information. He described how he uses precise syntax in his emails, specifying exact serial numbers for equipment and giving full names for field sites so that he can search his emails archive for mentions of specific field sites or pieces of equipment. Kyle indicated how he is occasionally frustrated that daily logs written by other team members do not use the same consistent syntax. He said that this makes it hard to search for activities related to particular field sites that took place while he was not in Peru.

Summing up this first vignette, Table 5.1 shows what data and metadata I identify that relate to the seismic field work. The data in this deployment is straightforward, motion readings from the sensors. Metadata, on the other hand, include a range of forms: maps, GPS readings, notebooks, and others. Note that I do list the numerous personal discussions between researchers as metadata. Personal discussions are critical to the continued success of the project, and the interplay between personal discussions and forms of documentation is a routine type of articulation work for field workers.

Table 5.1 – Data and Metadata from Vignette 1 and Commentary

Data

- Sensor readings of motion in three dimensions

Metadata

- Maps
- Site numbers assigned to each station
- Names of file directories and scripts
- GPS readings
- Notes in paper notebooks
- Notes typed into the field laptop
- Email to the CENS seismic mailing list outlining a day's work
- Logs of which seismometers, digitizers, and black boxes are installed at each station, including serial numbers
- Pictures of the installations
- Diagrams of field sites

5.1.3 Vignette 2 – Seismic data

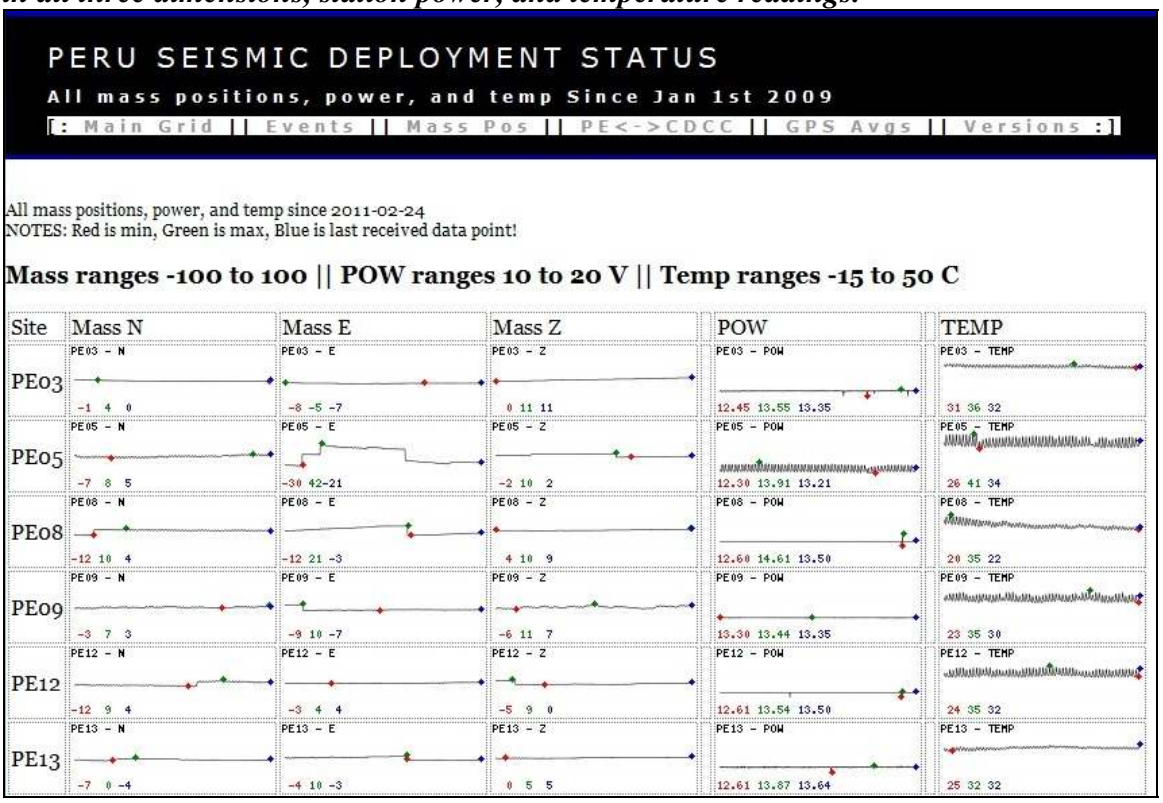
The CENS seismic project could not exist without the field work described in Vignette 1, but a large proportion of work on the project takes place in CENS universities, a continent away from the Peru field sites. I often run into members of the seismic project in the CENS facility at UCLA, particularly Jason, a computer scientist, and Kyle, the field engineer, both of whom have desks in the space. On one occasion I ask Jason to describe his work on the Peru project in detail for me. We have often talked informally about parts of his work, but I've asked if I can get the complete picture. I start by asking him about the how the data travel from the seismic sensor in Peru back to UCLA. First, he says, the sensor readings are grouped into hour-long data files. In Peru, data initially are stored on a compact flash card within the station electronics. If the wireless radio communication systems are working, the data files are then transmitted to

Internet hubs in Peru and back to UCLA using CENS point-to-point wireless routing protocols. The data files are initially stored and transmitted in a format created by the manufacturer of an off-the-shelf piece of hardware that digitizes the sensor's signals. If the wireless communication systems are not working, data files are stored on the flash card until manually downloaded by a member of the research team or until wireless communication is restored. Once arriving at UCLA, the data files are converted into the *Mini-SEED* format (Standard for the Exchange of Earthquake Data, 2009), a binary format that specifies a standardized structure for seismic data. The data are then sent to a partner institution for inclusion in the main project database, where they are held in the *Mini-SEED* format. At the partner institution, a new copy of the data is created in the *Seismic Analysis Code* (SAC, 2009) format, another binary file format for time-series seismic data. The pre-converted data in the native digitizer format and the *Mini-SEED* data are kept at UCLA in a local database. The second *Mini-SEED* copy and the converted *SAC* data are held at the partner institution for long-term storage. All subsequent analyses by researchers on the team use the converted *SAC* files. Jason's role in the project is to ensure that this software pipeline continues to transmit, convert, and deliver data without problems.

After this introduction, Jason shows me his newest creation. His computer screen is displaying a bunch of small graphs in a grid-like array. These are "sparklines," he says, which show how well the sensors in Peru are working. A common problem with these sensors, Jason describes, is that they can go "off center," meaning that their main sensing mass can shift around inside the sensor. Thus, sensors often need to be "re-centered" in

order to ensure that they are measuring motion correctly. The sparkline graphs show the deflection of the sensor masses in all three dimensions for the past 30 days, as well as the power availability at each station and the temperature at each station over the same time period. (see Figure 5.4)

Figure 5.4 – Sparkline graphs for six Peru stations showing (from left) mass deflection in all three dimensions, station power, and temperature readings.



The problem, Jason says, is that you can only know that you need to re-center the sensor by actually looking at how the sensor values change over time. So he created these sparkline graphs as a quick way to see which sensors need to be re-centered. He, or another team member, can then issue "re-centering" commands remotely via the internet, or if remote commands do not work, he can ask someone on the team to visit the site in person and diagnose the problem from there. Similarly, the sparklines that show the

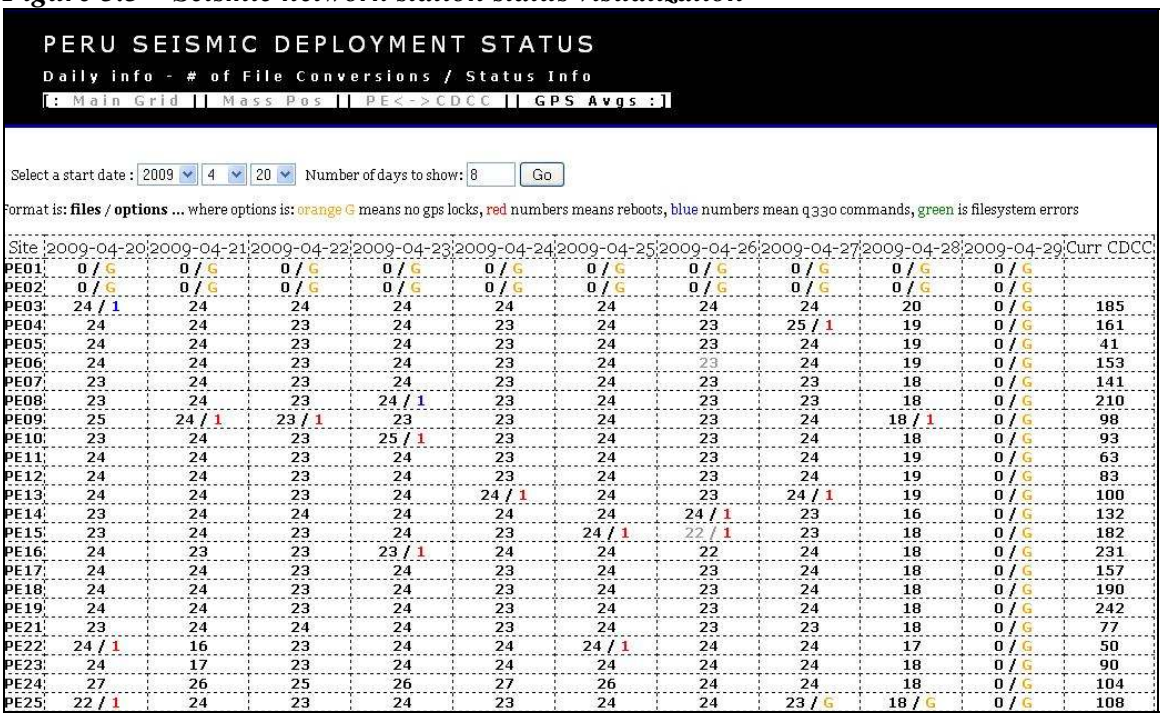
amount of power at each station are meant to identify problems with solar panels and batteries, so that any power-related problems can get prevented or fixed quickly.

Kyle, the field engineer, later expressed a similar sentiment, "*one of the things that we learned from our Mexico experience was that we didn't have sufficient tools in place to make it very easy to see how's it going today in terms of network health and the system and do I have sensors going bonkers or something get stolen, or a wide range of potential issues*." Thus, Jason developed scripts to create "network health" logs for each station and data packet. These logs include the disc space being used (and the disc space still available) at a station, wireless radio connection quality between stations, the route that data packets take through the wireless network as they get transferred back to UCLA, and any software errors that occur at a particular station. These network health logs are automatically generated by software, and are transmitted to UCLA with the sensor data packets. Once arriving at UCLA, the network health logs are separated from the sensor data, again automatically by software written by Jason, and stored in their own database.

The sparkline graphs, along with a few other similar visualization tools, leverage these logs. As Kyle described to me on another occasion, the visualizations are "*an indication …if there is anything else of interest, that I should know about, for example if there was a failure that day to get a GPS lock, if for some reason we have no metadata logs that should have come along, …system health logs, you know voltages and temperatures and boom or mass positions, and centering, and all that sort of stuff,…there's an indication of whether for some reason we didn't get that information. …[S]o this summary display page lets me see at a glance what's happening*."

Back at Jason's desk, he shows me the current version of his Peru deployment status summary web page (shown in Figure 5.5). With Jason, I co-supervised a summer intern who worked on the initial version of this and a few other similar pages, but Jason has since made a bunch of additions. The Peru deployment status summary web page, which is viewable by any member of the CENS seismic team (as are the sparkline graphs), shows how many data files have been transmitted back to UCLA for each station. If all is going well, each station should produce and transmit 24 data files each day, one for each hour. Thus, every "24" in the cells in Figure 5.5 indicates that a station is collecting and transmitting data files at the correct rate. If a cell is indicating that fewer than 24 files have been received for a given day, then it is likely that the station is experiencing a problem. Jason's most recent additions were to add the color-coded numbers after the slashes in the cells. These numbers show other system characteristics, including GPS locks, system reboots, and errors. This display allows the team to have a quick view of where and when problems are occurring in Peru.

Figure 5.5 – Seismic network station status visualization



PERU SEISMIC DEPLOYMENT STATUS

Daily info - # of File Conversions / Status Info

[: Main Grid || Mass Pos || PE<->CDCC || GPS Avgs :]

Select a start date : 2009 4 20 Number of days to show: 8 Go

format is: **files / options** ... where options is: orange G means no gps locks, red numbers means reboots, blue numbers mean q330 commands, green is filesystem errors

Site	2009-04-20	2009-04-21	2009-04-22	2009-04-23	2009-04-24	2009-04-25	2009-04-26	2009-04-27	2009-04-28	2009-04-29	Curr CDCC
PE01	0 / G	0 / G	0 / G	0 / G	0 / G	0 / G	0 / G	0 / G	0 / G	0 / G	
PE02	0 / G	0 / G	0 / G	0 / G	0 / G	0 / G	0 / G	0 / G	0 / G	0 / G	
PE03	24 / 1	24	24	24	24	24	24	24	20	0 / G	185
PE04	24	24	23	24	23	24	23	25 / 1	19	0 / G	161
PE05	24	24	23	24	23	24	23	24	19	0 / G	41
PE06	24	24	23	24	23	24	23	24	19	0 / G	153
PE07	23	24	23	24	23	24	23	23	18	0 / G	141
PE08	23	24	23	24 / 1	23	24	23	23	18	0 / G	210
PE09	25	24 / 1	23 / 1	23	23	24	23	24	18 / 1	0 / G	98
PE10	23	24	23	25 / 1	23	24	23	24	18	0 / G	93
PE11	24	24	23	24	23	24	23	24	19	0 / G	63
PE12	24	24	23	24	23	24	23	24	19	0 / G	83
PE13	24	24	23	24	24 / 1	24	23	24 / 1	19	0 / G	100
PE14	23	24	24	24	24	24	24 / 1	23	16	0 / G	132
PE15	23	24	23	24	23	24 / 1	22 / 1	23	18	0 / G	182
PE16	24	23	23	23 / 1	24	24	22	24	18	0 / G	231
PE17	24	24	23	24	23	24	23	24	18	0 / G	157
PE18	24	24	23	24	23	24	23	24	18	0 / G	190
PE19	24	24	23	24	23	24	23	24	18	0 / G	242
PE21	23	24	24	24	23	24	23	23	18	0 / G	77
PE22	24 / 1	16	23	24	24	24 / 1	24	24	17	0 / G	50
PE23	24	17	23	24	24	24	24	24	18	0 / G	90
PE24	27	26	25	26	27	26	24	24	18	0 / G	104
PE25	22 / 1	24	23	24	23	24	24	23 / G	18 / G	0 / G	108

Kyle, being the principal field engineer, is a main user of Jason's displays. Kyle described to me the benefits of being able to use these types of visualizations to see, for example, if a memory card is either too full or appears to be behaving erratically, "*when I see those errors, I have an opportunity to get out there in the next month, replace the card before we have a catastrophic failure, a data loss failure or something really really unpleasant*."

5.1.4 Commentary on Vignette 2

This narrative illustrates how metadata, whether metadata about data or technology, primarily serve immediate needs. This vignette describes a particular kind of metadata: automatically generated logs of data file transfer routes, disc space, station temperature, sensor centering, data conversions, and GPS readings. Members of the field team use this metadata to identify problems with an individual station or with the wireless

network, and to diagnose what needs to be done to eliminate the problem. As noted in this vignette, once the seismic data and metadata logs are transferred to UCLA, they are separated and stored in distinct databases through an automated process. The seismic data are converted to the *Mini-SEED* format for long-term storage. *SEED*, the Standard for the Exchange of Earthquake Data, is the accepted standard for storage and sharing of data within seismology (IRIS, 2010). A full *SEED* volume consists of two main parts: 1) the sensor readings, and 2) an associated header component. *Mini-SEED*, also called "data only" *SEED*, only includes the sensor readings. It does not include the associated header.

When I asked a Principal Investigator on the project about the differences between SEED and Mini-SEED, he responded:

"*The one is the subset of the other. …SEED is the more encompassing auxiliary information that goes around it. When I get a block of Mini-SEED data, which is the waveform data so, say 10 minutes of data in one block, it tells me the name of the station and the component that it is, the start time and the end time of that block. That's all it tells. …For example, there is no information on where the station is. That comes in additional information that may be added, so when we're well familiar with where the station is, we don't send all that information. So we just send the Mini-SEED blocks around and that's what the national archive wants too*."

The "national archive" for seismic data is IRIS, the Incorporated Research Institutions for Seismology (IRIS, 2011). The norm for seismology projects is to make data publicly available two years after the last sensor is removed from the ground. This is an accepted data embargo period within the seismology community, as multiple people

within the CENS seismology team told me. The data from the Mexico and Peru projects will be submitted to IRIS when this embargo period is complete. When I asked a Principal Investigator what kind of documentation is necessary to include when submitting data to IRIS, he described how they will have to create the full SEED header that is not included in their Mini-SEED data files:

"*Basically, what they would prefer to have is that the data, the Mini-SEED file; that's what they deal in and they love that. …But then you have to send them, what they call, data-less SEED volume. So it's like a complete SEED volume, it's just that there's no data in it. So that's … where they learn where the stations are, what are the properties of the station, which instruments are on there, so... We'll have to spend a little time creating those; we'll create those and send it off. I have a whole staff across the street that's very good at doing that, so I'll get some help on that*."

This quote is notable for three reasons, first in how it illustrates how the collocation of the seismic data with the metadata about instruments and locations will not happen until the project is completed and the data are to be submitted to a permanent data archive. Second, the "data-less SEED volumes" that document the data station and instrument properties will be created by "staff across the street," not the research team. "Staff across the street" refers to staff of a regional seismology data archive that the Principal Investigator also runs. And third, notably absent from this quote are the metadata logs discussed in the Vignette: the network health logs, the battery life logs, or the sensor centering logs. This information will not be submitted to IRIS as part of the data submission. Kyle describes the ambiguity that surrounds the long-term importance

of this metadata in the following passage. Note also how he discusses the sensor mass positioning readings (which are the measurement of how well sensor centering) as both "meta-information" and "data."

"*I don't know for sure what really happens with the data, whether that meta-information gets into a structured form somewhere associated with the actual seismology data so that somebody later can use that to validate or de-validate or whatever the right word is…invalidate their study and their conclusions. The UCLA stations, we do have extensive, minute by minute if not seconds, data about for example mass positioning [centering], it's very easy to take a glance at that data in a visual form and say, I wouldn't want to bet my dissertation or my reputation on data from this station, cause it looks really suspect. Um, and yet I don't believe we have a mechanism for that data to transfer as part of SEED or Mini-SEED to the um to the repository at IRIS. And this is a thing, where I just don't know and I haven't made it a priority to find out either. So this data may sort of um disappear at some point*."

Similarly, a student, when asked about what somebody else would need to know about the data collection process in order to understand and use the data, responded: "*You don't need to know anything about data collection except for some research you need to know what sensors, what hardware were used, in order to remove the instrument response. Because the sensor, it doesn't give you the pure ground motion. You use the ground motion plus the response of the instrument itself. So for some things you need to remove these instrument response in order to obtain the clean ground motion. So if you need this then you need to know about the hardware installation. But, like 90% of the*

research doesn't care about that. They just care about seismogram, basically location of the station and that the timing is proper, when you have it."

Thus, different forms of metadata have value to different people at different times. Metadata about the day-to-day functioning of the seismic network can be of critical importance at one juncture of a project, but be considered unimportant by the seismologists for long-term use and preservation of data. Similarly, the metadata necessary for long term preservation is not created in this case until the project is finished. Table 5.2 summarizes the types of metadata discussed in relation to the seismic data transfer and storage, and the network health logs. Note that I categorized the network health logs as "metadata." Researchers on the seismic team refer to these logs as metadata, but will occasionally call them "data" as well, as in "network health data."

Table 5.2 – Data and Metadata from Vignette 2 and Commentary

Data

- Sensor readings in digitizer format
- Sensor readings in Mini-SEED format
- Sensor readings in SAC format

Metadata

- Sparklines that show sensor and station characteristics: sensor mass position, power availability, temperature at each station
- Network health logs for each station and data packet which record: disc memory space, wireless radio connection quality between stations, data packets transmission routes, and software errors
- GPS readings
- *SEED* header components, also called a "data-less SEED volumes," which contain station properties such as locations and instruments details
- Sensor hardware specifications of the instrument response

5.2 Environmental science

In contrast to the CENS seismic projects, where sensors have been left in the same field site for in some cases longer than two years, the CENS environmental science projects discussed here have no regular sensor presence in any field site. Instead, these projects are oriented around opportunistic field trips, handheld sensors, the collection of physical samples, and laboratory analysis, all with the goal of measuring and tracking the movements of contaminants through an environment. CENS' contribution to research in this area focuses on detecting and tracking environmental contaminants, and developing methods for better and faster quantification of those contaminants.

5.2.1 Vignette 3 - Environmental science field work

I arrive at the Environmental Lab at 8am in the morning. When I get there, Julia and Rachel, both graduate students in the lab, are getting stuff together. Julia has a packing list, and is checking items off as they pack. They load everything into a green wagon, which is about three feet long and two feet wide with big wheels. They have packed an assortment of plastic boxes and bags with equipment, and a couple of coolers, one big and one small. The coolers are full of plastic bottles for collecting water samples. Each bottle is labeled. I find out later that these labels indicate a location where a sample will be taken. Julia also throws some rubber boots in the wagon. When we are almost ready to go, Julia calls Eric, another graduate student who is coming along, on her phone and tells him to meet us downstairs. Once finished packing, we go down to the parking lot just outside the building.

Eric is waiting outside. Julia has borrowed her roommate's truck; we put the gear in the back. Eric drives so Julia can sit in the back with Rachel, both of them being smaller than Eric and myself. We drive for about twenty minutes and turn into a parking lot on the beach. Next to the parking lot, we see our destination, a concrete storm drain that outputs into a water drainage area. After unloading the truck, we head down to the beach. Pulling the wagon across the sand is kind of tough, so I go behind and push from the back to help. We go down a small sand hill and into the storm drainage channel.

The storm drain is about 30 feet wide, with large concrete walls on either side rising up about 15 feet. There is not much water in the channel down by the beach, only a small trickle. We walk under over-passing streets as we head up the channel. We see a confluence of two other channels about 200 yards ahead of us, each with a stream of water in the center. Rachel drops a Dissolved Oxygen (DO) sensor into the water stream about 100 feet from the confluence, saying that it needs about 15 minutes to get conditioned before it is used. When we reach the confluence Julia and Rachel put on the rubber boots and walk across the east channel to a dry area between the confluence. One of them pulls the wagon across. The water is about a max of 2-3 inches high and about ten feet wide where they crossed. There are only two pairs of boots. Julia throws one of hers across for Eric and me to try on. Eric tries it on, but it is too small. I put it on, and can get my foot in there okay. So Julia throws the other one over and I walk across.

The general plan for today is to collect samples and take sensor readings at five locations. The team has eleven sampling locations picked out along these channels, but Rachel says that we do not have enough time to hit them all today. Rachel and Julia talk

about which locations to sample. After not too long, Rachel and Eric walk back down to where she had dropped the DO sensor to take readings and water samples. The water is not as wide down there, so Eric jumps across the stream. I stay up by the confluence with Julia as she pulls out some equipment.

When Eric and Rachel come back, Rachel gives Julia a bottle of water that she just collected from the stream. Rachel is going to do the data collection while Julia does some in-field analysis on the samples. Eric is to stay with Julia while she does the analysis (I'm not sure why), while I am to go with Rachel up into the channels to help her collect the water samples and serve as the sensor reading recorder. Eric gives me Rachel's notebook, which they use to record sensor readings. They show me what is in the notebook. There are a bunch of somewhat disorganized numbers and labels written down from past field trips. It takes me a few minutes (and questions) to parse out what Eric had recorded during the last trip. The primary phenomena that they record at each site are (with the units indicated in the notebook): dissolved oxygen (mg/L), pH, conductivity (milli-siemens), particulate material (in parts per thousand), temperature (degrees Celsius), stream depth and width (inches), and velocity (seconds per seven poles, the velocity units are discussed further below). The ways that they are written down in the notebook, however, are by abbreviations, most commonly the relevant units: "DO," "pH," "mS," "ppt," "°C," "vel." The team does not use formalized data collection forms, everything is written down however the data recorder decides to write it down.

Once I understand the notebook, Rachel and I set off into the left channel. We walk for a ways, passing under a few bridges. We reach the first site at which Rachel is

planning to sample, but she decides to go to the farthest one out and get this one on the way back. So we walk farther to the next site. When we get to the next site, we stop. Rachel says that they try to pick sample sites to be right below storm drains, which are usually under bridges. Rachel first puts the sensor into the water, which measures dissolved oxygen and temperature. She then gets out two plastic bottles and fills them with water from the stream. I ask her about the site name so I can note it in the notebook. This site is "W4." The sites in this channel are named W1-W5, the sites in the other channel are named E1-E5, and the sites below the confluence are named C1, etc. I assumed that the W, E, and C, stood for West, East, and Center, but when I ask Rachel, she says that the W and E stand for the streets that the channels follow, and the C stands for "confluence."

After Rachel is finished with filling the water bottles, she takes out a handheld sensor. The sensor measures pH, conductivity, particulate material, and temperature. Rachel reads me the values, which I write down in the notebook using the aforementioned abbreviations and units. She then takes out the tape measure and measures the width of the stream, as well as the depth of the stream at multiple points across the width. I write these down in the notebook, including the width interval between the depth measurements. To get the velocity reading, she measures about 10 yards downstream and makes a marker at that point. She gives me her watch, and drops a plastic cap into the stream to see how long it takes to get down to her marker. I time how long it takes, and write down the velocity in "seconds / 7 poles." At some point in the past they used an actual pole to measure the distance. The distance of "one pole" is about

46 inches, Rachel says, but they still make all of their velocity measurements in "7 poles." Similarly, when I look in the notebook at measurements from previous trips, I see depths and widths recorded in units of "Samantha's boots" and "Morgan's shoes."

We repeat this process at each site. We first go back to the site we passed earlier, and take all of the same measurements there. We then go back to the confluence, where Julia is analysis water samples with her field wet-lab. Rachel gives her a sample from the first site we went to. Rachel and I then go up the other channel, the E channel. Again we pass by one site that we will hit later. We walk up to the further location, collect samples, and take sensor readings. When finished, we walk back toward the confluence, and stop at the site we passed earlier. After finishing there, we walk back to the confluence.

When we get back to the confluence, Julia is still doing analysis on the second sample we gave her. She is using a drill as an ad-hoc mixer. She attached a thin rod to where the drill bit should go, and taped her vial onto the rod. She then uses duct tape to hold down the mixer trigger at a medium speed for 10 minutes. She uses various enzymes to try to detect certain types of bacteria. The enzymes create a reaction that generates light of a particular kind, and she has a piece of equipment that measures the light. She writes the light values down in her paper notebook. In the time we have been collecting four samples, she is halfway through analyzing her second sample. What she does not get done in the field, she will analyze back in the lab. Rachel says that as soon as they get back to the lab, they have to do the bacterial analysis and DNA extraction, because those must be done within six hours of the sample collection. Julia says that she is doing the in-lab analysis more as a proof that they can do in-field analysis than anything else. We wait

for about 30 minutes for Julia to finish analyzing her second sample, and when she is finished, we pack up and go back to the truck.

During the trip back, I asked what they do with the sensor readings that are collected in the notebooks. Julia says that they transcribe them into Excel spreadsheets. Eric says that they then "*try to find correlations between the high numbers and anything else*."

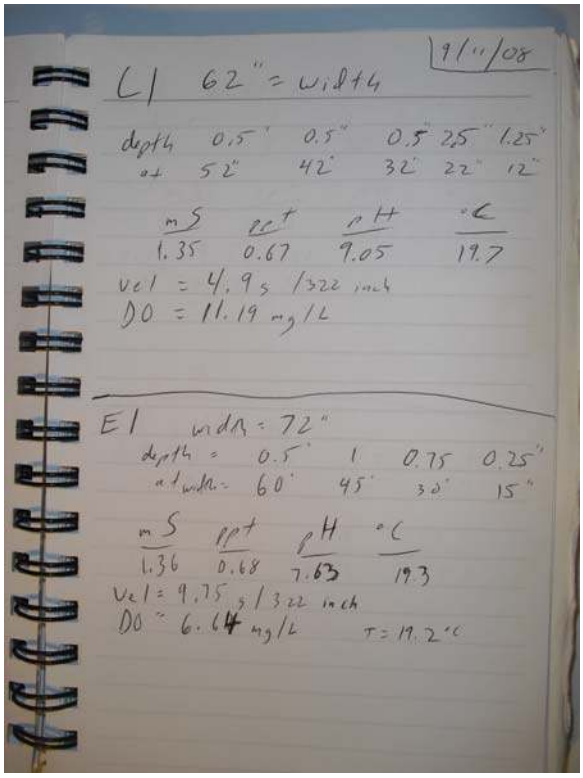
5.2.2 Commentary on Vignette 3

This vignette again illustrates the embodied and situated nature of field work. It also illustrates how multiple intermediaries exist between the phenomena of interest and what eventually is considered "data." In this lab, physical samples are often shared. Each person in the lab is responsible for performing a specific type of analysis. Physical samples, in this case water samples, are collected by one or two people (perhaps with the help of an ethnographer), and multiple kinds of analyses will be run on portions of the same sample. Team members thus have to rely on their lab-mates to collect the samples correctly. Because they work closely together and often collect samples for each other, the level of trust is high. Less immediate partnerships require even more trust. One student gets her samples from out-of-state partners. She said that she tells her partners approximately where to collect samples for her, like "on the edge of the field" or "near the center" of the field: "w*e'll come up with the document of 'this is what we would like, this will not happen, this is....' I give the procedural stuff that like it would take to get this for us. 'We would like you to note, blah blah blah, what location it was from, what depth it was, who collected it.' Anything that is relevant, necessary to know. We try to picture*

what they know to tell us. If they have questions, they'll usually contact us to be more specific or elaborate." But since she is not actually there to oversee the sample collection, she never knows for sure exactly where it was taken, "*So we do the best we can. But we're just never 100% certain.*" In her project, it is not possible to take many trips to the field site because of prohibitive travel costs, and so this reliance on remote assistance is necessary.

This trust in intermediaries carries over to recordings of sensor readings. The sensors used in this field narrative are all handheld, and the sensor values are all written down in notebooks to be transcribed to a computer later. I, in my first field trip with this team, was given the task of recording values from sensors in the main field notebook. On a subsequent field trip, I recorded the sensor values in my own field notebook because both of the people from the lab forgot to bring their own notebooks. One of the lab members took digital photos of my notebook afterwards to get the values. Figure 5.6 shows one of these pages from my notebook.

Figure 5.6 – Photo of my notebook from a field trip with environmental science researchers.



My practice for recording and documenting sensor readings and stream measurements emulated the practices of the environmental science team. The top half of the image shows the data collected at site "C1" on this particular date. Right next to the "C1" on the top line is the stream width measurement, 62 inches, underneath of which are the depth readings. The numbers immediately under the depth readings indicate at what distances across the stream the depth readings were taken. The stream measurements are followed by the readings from the sensor, "mS," "ppt," "pH," and "°C." Below the sensor readings are the stream velocity measurement (note the unit "s / 322 inch") and the dissolved oxygen sensor reading ("DO").

Labeling is also a salient theme of this vignette. Labeling is a key metadata process. Researchers pre-label the bottles to be used for sample collection. These labels then serve as the primary identifiers for data that are generated from those samples, and for sensors data that were taken at those locations. Insufficient labels can cause problems with long-held or shared samples. One student gave me an example of how some DNA samples did not have the year on the label. Because DNA samples can be stored almost indefinitely once processed, having long-held samples from more than one year ago is common, making the year of sample collection an essential part of a label. Similarly, the labeling of units can be highly situated, as illustrated by the use of "7 poles" as a unit of distance in the measurement of water velocity. One pole measures 46 inches. In Figure 5.6, my notebook shows that I recorded the velocity values on that particular trip with the unit of distance being "322 inch," which equals "7 poles" x 46 inches/pole. Once a unit like "7 poles" has been established, routines build up around it. Velocity values get recorded using the custom unit and work flows are created to convert values to the conventional velocity unit of meters per second.

A student on this project told me that to her, the data are "the numbers." As this vignette shows, getting from the field to "the numbers" involves intermediaries, metadata processes, and routinized action. Table 5.3 illustrates the types of data and metadata identified in this vignette. The environmental science team record numerous variables, from the physical and chemical sensor readings to the properties of the stream itself (width, depth, velocity), as well as performing field and laboratory analysis on physical samples. The types of metadata that they create and use during these processes also vary

widely. Categorizing some of the items in this table was not straightforward. For example, I categorized the width and depth of the stream as data, because these measurements are used to calculate contaminant flow. But, I categorized the "width interval between stream depth measurements" as metadata because they are similar to the units of measurement in that they indicate how the depth measurements were performed. This distinction, however, is somewhat ambiguous, and indicates the difficulty of assigning the category of "data" or "metadata" in some situations.

Table 5.3 – Data and Metadata from Vignette 3 and Commentary

Data

- Water samples
- Sensor readings written down in notebook
- Width of the stream
- Depth of the stream
- Stream velocity measurements
- Light measurements to detect bacteria, written down in notebook
- Bacterial analysis performed in lab
- DNA extraction performed in lab
- Excel spreadsheets containing measurements transcribed from notebook

Metadata

- Packing list
- Bottle labels
- Personal discussions about which locations to sample.
- Measurement units written in notebook
- Sampling site names
- Width interval between the stream depth measurements
- Sampling plans created for partners in remote field sites
- Digital photos of notebook that contains measurements

5.2.3 Vignette 4 - Environmental science lab work

I meet Claire in the main Environmental Lab to observe how she analyzes soil samples for a particular contaminant. When I arrive at the lab, Claire grabs Layla, an undergraduate assistant, and leads us out into the hall to go to another room. Claire realizes that she has forgotten the samples that she is going to analyze, and runs back into the Environmental Lab. She comes back into the hall with a few trays of tubes half-filled with liquid. The tubes were prepared before I arrived, and contain small amounts of soil diluted with a liquid that I find out later to be water and a type of acid. We walk down to what Claire called the "machine lab," which is just down the hall from the main Environmental Lab.

The machine lab, as the name suggests, has a bunch of machines in it. The one Claire is using today is the "Graphite Furnace," which she uses to do contaminant quantification. Claire says that that they keep the Graphite Furnace in the machine lab because there is no space for it in their own lab. The machine works by injecting small quantities of an aqueous sample into a small graphite cube, which is heated to high temperature. The machine then vaporizes the sample, and a light is shined through the vapor. The amount of light that is absorbed by the vapor is an indication of how much contaminant it contains.

To start, Claire shows us how to run the machine using a nearby desktop computer. She describes her process to us as she goes through the process. She initializes the Graphite Furnace machine to do six calibrations first, and then to run seven samples. From there, she inputs names for each of her samples. She makes sure that the sample

names she inputs into the software correspond to the sample names on the actual tube labels. We look at the tube labels for a minute. Each sample tube is labeled with the sample name and then "F1W1" or "F2W1," which indicate the methods for sample preparation, specifically "fraction 1 or 2" and "wash 1 or 2."

Around this point, Claire prepares the standard solutions for the machine calibrations. The standard solutions are purchased from chemical companies. The volume of solution the machine needs is quite small, about one pipette's worth. Claire pipettes the standard solutions into small plastic cups. She puts the standard solutions into the machine and, using the computer interface, starts the machine running. She says that it takes about five minutes to run each sample, and since there are six standards, one "blank" and five different standard concentrations, she can prepare her actual samples while the machine is running the standards calibrations.

She says that she calibrates the machine each time she runs it. So, if she runs the machine again tomorrow, she will do the full calibration again using this same process. She says that she is not sure that the re-calibrations are necessary, but she is also not sure how much the calibration variability affects her output values, so she does it anyway.

She then starts to prepare the actual samples. She is running seven samples today. She pulls sample tubes out of the trays seemingly at random, but she clearly knows which tubes she is grabbing. She glances at the labels as she pulls them out. She is analyzing the samples for a particular contaminant. They use four different sample preparation methods to extract all of the contaminant from the sample, and as Claire describes, each sample preparation method pulls the contaminant out of the sample progressively. In short, the

first method pulls the contaminant that is easiest to come off, then the second method pulls contaminant that is more tightly bound to the sample, and so on, until the fourth method pulls contaminant that is the most tightly bound to the sample. Today she is running just the samples extracted through the first method (as indicated by "F1," the "first fraction," on the tube labels).

She pipettes a small volume of each of the seven samples into small plastic cups, and puts the cups into the machine holding tray. The machine is now running the third or fourth standard solution. Claire mentions how she is sort of skipping a step with the calibrations today. She said that she probably should treat the calibration standards with acid before running them in the machine, because her actual samples are treated with acid in order to preserve them for longer periods of time. But, she says, she will probably run a "blank with acid" calibration standard later in the afternoon in order to get an idea of how the acid affects the readings.

As the machine analyzes the standard solutions, Claire showed us how she will output the data from the computer. She first sets up an output file, with the filename consisting of the contaminant and the date. For example, if she is analyzing samples for lead, then she would use the atomic name for lead, "Pb," and the date for the filename, as in "Pb_2-28-11." She sets the machine to output the results in a comma-separated text file so she can open it in Excel. But, she says, she does not actually use the machine output files because they are not in a form that is easy for her to use. Instead, she records the values that are important to her in her notebook. She says that she then transcribes the data from her notebook to Excel, and that she has an already existing Excel file to which

she will add today's values. But even though she does not use the text-file outputs from the machine, she writes the filename for today's analysis in her notebook anyway in case she needs it later. She also writes a brief description of today's work in her notebook.

The machine finally finishes with all six calibrations. Claire points to a graph of the calibration values that is being displayed on the computer interface, and says that the graph looks good. It also gives a calibration equation, but Claire does not write it down. She says that she does not actually use it, because she mostly just wants to know that the machine is in calibration, and at what point it starts going out of calibration. She points at the calibration graph again, and shows us how in this case the calibration starts becoming non-linear at about 50 ppm. Since most of their samples will be much lower than that, she says, they will all fall in the good range for the machine. Claire said that she deliberately has the machine calibrate to standards that are outside of its optimal range so she can get a good idea of what that optimal range is.

After the calibrations finish, the machine starts to analyze Claire's first sample. Layla, the undergrad assistant, says that she has to leave for class, but that she was glad to have seen the sample preparation and machine setup. She also says that she probably should have brought a notebook and written all the steps down. Claire laughs and says that there will be plenty of opportunities, and says that the seven samples we are running today are the first of about 150 samples in total. Claire also tells Layla that she has written up a "protocol" that describes the full analysis process, the computer and machine set-up, and the sample preparation steps, in a Word document. The protocol is posted on the lab wiki.

After Layla leaves, Claire says that it is nice having undergraduate assistants, but it is also hard sometimes because they have such a tight class schedule. She says that she tries to keep the undergraduates involved in the science part of the lab work, because they often spend much of their time washing equipment, which is important, but not that interesting.

When the first few samples are finished running, Claire points at a few numbers on the screen and says that these are the numbers she needs. She says that she is glad to see that they are not zero, because had they been zero the machine likely would not be working correctly. She writes these numbers down in her notebook.

When the machine finishes running all seven samples, Claire shows me the output file. She opens it in Excel. The file is about 100 lines long in total. The numbers Claire actually uses, which she wrote down in her notebook earlier, are about two thirds of the way down the file. The rest of the file is information that Claire says she does not use. She only needs one number per sample, so it is much easier to just write the values in her notebook and then transcribe it into Excel than it is to try to deal with this machine output file.

She shows me the folder in the computer where her machine outputs get saved. The folder has a bunch of output files, around 30 or 40, and Claire says that all the files are hers because she is the only one in her lab who uses this machine. She says that she never deletes the files off of the computer, she just leaves them there. She says she uses a jump drive to back the files up and to transfer them to her own computer because this computer is not connected to the internet.

5.2.4 Commentary on Vignette 4

Researchers need data in a particular form. This narrative illustrates how multiple steps of laboratory work that take hours and use complex computerized machinery may result in a researcher recording fewer than ten numbers in a notebook. It was much easier for Claire to record values in her notebook and then transcribe them into Excel than it was for her to try to manipulate the complex file that was output by the Graphite Furnace machine. The workings of the Graphite Furnace machine, being a purchased machine, not a CENS-developed machine, are particularly opaque to members of the Environmental Lab. Claire described how the Lab has to pay for maintenance on the machine from time to time, to a bill of about $4,000 each time.

Researchers may not learn much about their laboratory machines beyond what is necessary to continue their immediate work. For example, Claire had already generated more than 30 output files using the Graphite Furnace, but was still unsure about how long a calibration would last, or to what extent the calibration variation affected her output values. She used the calibration as an indicator that the machine would give her valid readings. She described how she used to send her samples to a chemistry lab on campus and pay for the samples to be analyzed. But according to Claire, her lab started having so many samples that it became more cost effective to buy this Graphite Furnace machine and learn how to run it themselves, even with the maintenance costs.

Claire described how her lab notebook is central to her lab work:

"*So in the notebook pretty much everything that I do in lab goes in my notebook. Whether it's like planning stages of experiment, or taking notes on papers I've read, or*

actual data I'm collecting from the experiments. So, everything that happens pertaining to my projects, I try to... note in my notebook. And then what gets pulled into an Excel file then is usually just the numbers, and the notes specifically pertaining to that number."

Her notebook thus contains a mixture of data and metadata. Her metadata processes include documenting experiment plans and literature, as well as metadata about the "numbers," which she considers to be her data. What gets transcribed into Excel, however, is only the metadata related to the numbers: "*So it will be all the numbers in [the Excel file] and then if a specific one needs a note or a general note, I usually put in a separate cell, like 'notes'*." She transcribes the other material recorded in her notebook if necessary. She said that she usually will not create Excel files to document the planning portion of her experiments unless they are particularly complex. Similarly, she will sometimes write up her literature notes in a text document if she wants to have it in a searchable format.

As Claire described to me, she will leave copies of her digital data files in the lab when she leaves, both the machine outputs and her Excel transcriptions, but she will not be leaving her notebooks. As she stated, "*We take our notebooks with us. So that we have our own notes, because it will be hard for anybody else to decipher what we did anyway [laugh]. …But I think most of us who pass through UCLA have all taken their stuff with them. … While it's good for me to have somebody else's notes, it's probably good for them to have their own notes too*."

Table 5.4 shows the data and metadata that manifest in this vignette. These categorizations illustrate the indistinct boundaries between data and metadata in this

particular research process. Researchers use lab notebooks to record both measurement values and notes about lab work. Similarly, the machine output files contain much more than just the measurement values, including details about the machine setup and calibrations that categorize as metadata.

Table 5.4 – Data and Metadata for Vignette 4 and Commentary

Data

- Contaminant concentration values written in the notebook
- Contaminant concentration value transcribed from the notebook into Excel
- Machine output files (concentration values that also get written in notebook)

Metadata

- Names of samples written on test tube label (F1, W1, etc.)
- Sample names input into the computer (same as tube labels, F1, W1, etc.)
- Description of the day's work in lab notebook, including list of standard solutions and samples analyzed, also a note of the file name of the machine output file
- Graph of the calibration curve being displayed on the computer screen
- Lab "protocol" that describes the full analysis process
- Lab wiki
- Notes pertaining to specific numbers that get transcribed into Excel
- Machine output files (calibration curves and equations, analysis parameters)

5.3 Aquatic biology

The CENS Aquatic Biology project uses a combination of continuous sensing, similar to the seismic project, and opportunistic data collection campaigns, similar to the Environmental Science project. The work discussed in the following two vignettes centers on a CENS study of harmful algal bloom growth in a southern California harbor. CENS biologists installed aquatic sensors at two locations in the harbor in 2008, with the goal of establishing a baseline characterization of water temperature, salinity, and dissolved oxygen content, among other parameters. As of 2011, these sensors have

remained in the field since their installation, with periodic removals for calibration or other maintenance. The team has also performed a number of shorter data collection campaigns and experiments, both pre-planned campaigns to investigate a specific question, and reactive campaigns to measure the onset of an algal bloom.

5.3.1 Vignette 5 - Aquatic biology field work

I arrive at the first harbor location, and catch Evelyn as she is going through the gate down to the dock. Evelyn is a new technician who has been on the job for about four months. I met her when I visited the aquatic biology teams' lab about a month earlier to observe a training exercise in which Maria, a senior graduate student, taught Evelyn how to calibrate one of the two kinds of sensors that they use. Today, Evelyn is visiting the harbor sites today with Samuel, her friend, to download sensor readings, clean the sensors, and perform sensor calibrations. She performs these maintenance trips bi-weekly, a schedule she has developed since she started working on the project.

Evelyn says that their first tasks are to download the readings that the WQM sensor has taken over the past two weeks, and then clean the sensor itself. WQM stands for Water Quality Monitor, and is an off-the-shelf sensor package. First Evelyn connects her computer to the WQM sensor using a cable. Once connected, she uses a computer program to stop the sensor from taking any more readings. After the sensor is stopped, they pull the sensor out of the water. The reason, Evelyn says, the sensor needs to be stopped before being removed from the water is that the sensor uses a pump to move water past the sensors, and the pump should not be activated while the sensor is not in the water. After the sensor is stopped, Evelyn starts downloading the readings from the

sensor. Downloading two weeks of readings from the WQM sensor takes quite a while (~10 minutes), so Evelyn and Samuel start cleaning the sensor while the download was ongoing. The sensors get quite dirty sitting in the water, which Evelyn calls "biofouling." There is quite a bit of algae on the sensor cage, the sensor tubes, and on top of the sensors. As Evelyn and Samuel clean the various nooks and crannies in the sensor cage, they pull out a bunch of small round gooey creatures. They tell me how a small crab ran out of the cage during one cleaning. They use various brushes, pads, and wipers to clean the sensors and cage. Figure 5.7 shows the WQM sensors before and after cleaning.

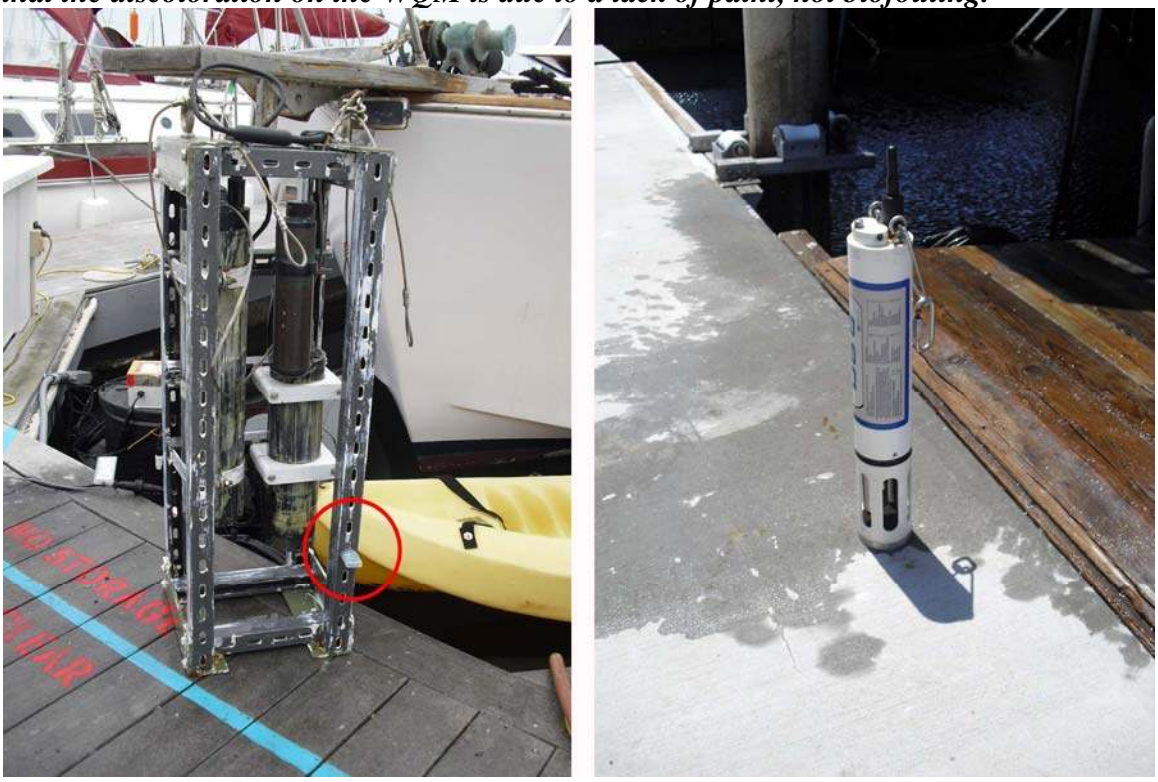
Figure 5.7 – WQM sensors before (left) and after (right) cleaning

Eventually the file of sensor readings finishes downloading. Evelyn opens the file to show me what it looks like. It has a header that lists the various sensor parameters, column headings with units, and columns of sensor readings (shown in Figure 5.8).

Evelyn then pushes another button that converts raw file of sensor readings into a .dat file. Evelyn opens that file to shows it to me as well. I notice that the header parameters are not present in this file; it just contains the column headings (with units), and the readings. I ask what happens to the header, and Evelyn says that since it always stays the same, Maria, the student who is the main user of the data, probably does not need it for every file. As Evelyn is setting up the next WQM data file, she shows me a few screens of the software interface. She shows me the screen where she sets up the parameters for the next run. These parameters will then be output in the header of the file that Evelyn downloads in her next trip.

Before they put the WQM back in the water, Evelyn attaches a light sensor to the outside of the sensor cage. She says that Maria gave her the light sensor to install. The light sensor is a self-contained little circuit board about an inch square, which contains a small light sensor in the middle. It is encased in a plastic covering. Evelyn attaches it on the WQM cage using zip-ties, at about the same height on the cage as the other sensors. She said that Maria told her to attach it at the same level as the other sensors. I take a picture of the “after cleaning” sensor, which also shows the light sensor, and then Evelyn and Samuel put the WQM back into the water.

Figure 5.8 – WQM with light sensor circled (left), and a Hydrolab sensor (right). Note that the discoloration on the WQM is due to a lack of paint, not biofouling.



We then move to the Hydrolab sensors. There are two Hydrolabs installed in this location, one near the surface (about 1.5 meter depth) and one deeper (about 3m depth). Hydrolabs are another brand of off-the-shelf aquatic sensors. Evelyn says that she has been having trouble connecting the computer to the shallow Hydrolab sensor, so her plan is to remove that one after she gets the data, and then move the deeper one to the shallower spot. She is also going to do the dissolved oxygen (DOX) calibration on both sensors if possible because Maria has been seeing discrepancies in the DOX values between the two sensors. Evelyn mentions that she is expecting an email from Maria with the barometric pressure at the harbor locations, because they need that value to do the DOX calibration. But when she checks her email (which she can access wirelessly on the

dock), she has nothing from Maria, so she looks the barometric pressure up online. She uses www.weatherforyou.com to get the value, which she writes in her notebook in both inches and millimeters.

Samuel asks Evelyn which sensor to pull up first, and Evelyn says the deeper one. This sensor has "Lazarus" written on it in large permanent marker letters. As Evelyn describes, at some point in the past the sensor had stopped working, but then started working again without obvious reason, hence the name "Lazarus." Samuel plugs the computer cable into the sensor, but Evelyn has trouble establishing a connection with the computer. Evelyn says that they commonly have trouble connecting to Hydrolab sensors. So they put the "Lazarus" sensor back in the water, and pull up the other sensor. Evelyn says that she is pretty sure that this sensor, which has "Steve" written on it, has a loose connection somewhere. She told the main lab technician that she would be bringing it in for repairs. But when they plug the computer cable into the "Steve" sensor, Evelyn is able to establish a connection. She then tries to download the file that contains the sensor readings from the past two weeks, but the program on her computer freezes when she tries to download the file. She re-connects and does it again, but the same problem happens. They try unplugging, re-plugging cords and connections of various kinds, but the same problem keeps occurring. Then she tries to download the file in a different way using the software, and it seems like it works because some file starts downloading. So we wait for this file to download. After it downloads, Evelyn opens it, but it has no sensor readings in it. It has the header information, which includes some details about the sampling run, but the readings themselves are not in the file. We puzzle over this for a

few minutes. Evelyn is not sure what this means, because she has not had this problem before. The readings should still be in the sensor, because otherwise the computer would have just downloaded a small file quickly, but that is not what happened.

Evelyn is not sure what to do, but decides to do the dissolved oxygen calibration on the sensor anyway. First they clean the sensor. Cleaning the Hydrolab sensors is not as onerous of a process as cleaning the WQM, because it is a smaller sensor with less crevices, but is still takes some time because there are a lot of delicate sensor parts that can be fouled up. After they clean it off, they perform the dissolved oxygen calibration. Evelyn pulls the Hydrolab sensor manual out of her backpack and opens it to the calibration section. She places the manual on the dock beside her for guidance as she goes through the calibration process, which involves dipping the sensor in a cup of purified water and following the calibration protocol on the computer. She writes the calibration values – the readings the sensor gives when in the purified water – in her notebook. When Evelyn and Samuel are finished with the calibration, they put the "Steve" sensor to the side.

Then they pull the "Lazarus" sensor back out of the water and try to connect again, but again it does not work. They try unplugging and re-plugging the connection about three times, but nothing happens. Then they power-cycle the sensor, that is, Samuel opens and closes the battery case. Samuel says that power-cycling has worked in the past when they had connection problems, and indeed after the sensor turns back on Evelyn is able to connect to it via the computer. But, now she gets an error when she tries to operate the software. She writes the error down in her notebook, then closes the error

dialog box and reconnects. This time she gets another error, but it is a different error. I asked what she was doing this time, and she said she was trying to open the file of sensor readings. Then she gets a third error, which is the same as the first but with a different location address for the error. Then a cascade of error boxes start appearing, so she shuts down the software. She is pretty distressed about what to do. So she decides to call Maria on her cell phone. Maria does not answer the call, so Evelyn leaves a message describing our problems. After she leaves the message, she says that she is angry that she forgot to take chlorophyll calibration readings before they cleaned the "Steve" sensor, which she was supposed to do.

Evelyn decides to clean the "Lazarus" sensor, so with Samuel's help, they clean it off. Evelyn decides to try to re-connect to "Steve," to see if she can at least do the chlorophyll calibration and restart the sensor to leave it in. She says that she wants to leave at least one sensor in the water. She says she can then at least tell the Principal Investigator that they are still collecting data, but if they have to take the sensors out, then she cannot tell him that anymore. She is able to reconnect to "Steve," and calibrates to the chlorophyll standards. There are two standards that slip over the chlorophyll sensor, which measures fluorescence from organic matter in the water. They usually do the calibration both before and after cleaning, but Evelyn forgot to do it before this time. They use the low standard first, then the high. Evelyn says that the clean and dirty calibration values are used like bookends. The clean value is used for the start of the data file, and the dirty value for the end. So the dirty values from today will be used for the sensor readings collected today, once they are downloaded off of the sensor. Evelyn gets

the calibration value from the computer while the sensor is reading the standard. She writes the values in her notebook. She then starts setting up the parameters for the next run.

Maria calls Evelyn back at this point. Evelyn talks to her for a few minutes. When she gets off the phone, she said that Maria says to just take both sensors out since they need to be recalibrated anyway. Evelyn was worried about having to take the sensors back to USC today because she only has the solid cup and cap for one sensor, and is not sure if the sensor can sit out in open air for the weekend. Maria told her to just put it in water when she gets home, and it will be fine. After Evelyn is off the phone, she shuts the "Steve" sensor down, and we start packing up. We bring both Hydrolab sensors back to their truck.

Evelyn says that she will let me know when she is planning to re-deploy the sensors, probably next week. Hopefully she will also know what the various problems were with the sensors and connection to the computer by then too.

5.3.2 Commentary on Vignette 5

Documentation of machines is a critical part of sensor-related research. Vignette 5 shows an example of how unreliable machines directly impact data collection process. In this narrative, Evelyn is forced to remove two of the three sensors from the field site because she is unable to connect to them with her computer. Without connecting, she cannot download the sensor readings currently on the machine, and she cannot perform calibrations or collect any calibration information. Calibrations are a critical step in the use of any sensor, as calibrations provide a ground-truth that the sensor is actually

collecting what it is supposed to be collecting. Calibrations show up throughout my Vignettes, from re-centering seismic sensors to running standard solutions in the Graphite Furnace machine, and are a key method of establishing and ensuring trust in machine functionalities. “Calibration,” in relation to sensor-based research, refers to the process of correcting systematic errors in sensor readings. Bychkovskiy, et al., (2003) outline how the term “calibration” has also often been used specifically “in reference to the procedure by which the raw outputs of sensors are mapped to standardized units. Traditional single-sensor calibration often relies on providing a specific stimulus with a known result, thus creating a direct mapping between sensor outputs and expected values” (pg. 302). As such, calibrations are tied into metadata processes, in that they are one means by which researchers ensure that the sensors are operating correctly.

In the case of the CENS aquatic biology project, calibrations are a particularly exacting process because the sensors that they use are actually multiple individual sensors packaged into one instrument. For example, the Hydrolab sensors have at least five individual sensors on them, each of which has a different calibration process. Calibrating the full suite of sensors on the sensor packages used in this project takes multiple hours. Thus, to save time while performing her routine bi-weekly maintenance and data collection field trips, Evelyn typically only performs a calibration on the chlorophyll sensor. The readings from the chlorophyll sensor is a light-based sensor, and is thus are directly affected “bio-fouling.” Algae and other organisms grow on the sensor while it is in the water, covering the entire surface of the sensor package, and thus impact the readings that a light-based sensor takes. This is why Evelyn and Samuel clean the sensors

as part of the maintenance. But the chlorophyll sensor must be calibrated both before and after the cleaning in order to measure the effects of the bio-fouling. Evelyn writes the calibration information in her notebook, and once back in her lab transcribes that information into an Excel file. When no calibration is performed, as this vignette illustrates, Evelyn still notes this in the Excel file, with a note saying, "Unable to download files. Sensor removed from field."

Trust in the machine thus involves knowing when it was and was not working properly. Calibrations are a main method through which researchers determine when a sensor was not working. The extent to which calibrations are documented varies. In the case of the chlorophyll sensor, the calibration documentation is very important, and, as the next vignette illustrates, a critical step if the resulting data are to be used at all. Table 5.5 summarizes the forms of data and metadata identified in this vignette. Sensor readings in multiple formats, and from multiple sensors, are the main form of data, while metadata include notes in notebooks, headers in data files, and sensor manuals, among others.

Table 5.5 – Data and Metadata in Vignette 5 and Commentary

Data

- Sensor readings downloaded from the aquatic sensors
- Sensor readings converted into a .DAT file
- Readings taken by the light sensor

Metadata

- Sensor download file header that lists the various sensor parameters and column headings with units
- Location where the light sensor was installed on the WQM sensor cage
- Barometric pressure at the harbor locations
- Label ("Lazarus") written on the sensor as an identifier
- Sensor manual from manufacturer
- Calibration values written in field notebook
- Excel file of calibrations and field activities relating to sensors

5.3.3 Vignette 6 - Aquatic biology lab

Maria, an advanced Ph.D. student in aquatic biology, shows me her data on her computer. We sit in her office, which is attached to the main Aquatic lab. The lab has a number of wings. As you enter, on the right and left are lab areas with machines, sinks, bottles, and coolers. In the rear of the lab is a common area with a number of office-like rooms attached, one of which is Maria's. When I arrived, Daniel, the faculty leader of the lab, was in the common area having a meeting with a number of other students I did not recognize. I am visiting the lab to ask Maria about her use of the sensor data being collected at the Harbor (described in Vignette 5), and any other data that she uses. Evelyn, the field technician, shares Maria's office and is listening in on our conversation.

Maria says that they have "lots and lots" of data that they are currently analyzing, and that the data are complicated to analyze because she has many factors to consider. As she describes, the CENS aquatic team has been collecting sensor data in the Harbor

continuously for a couple of years. They also collect physical water samples from the Harbor every two weeks. In addition, they have performed shorter experiments in which they have done more intensive data collection. Maria describes one two-week long experiment in which she collected sensor data continuously using a specially designed robotic winch. The winch allowed her to collect sensor readings at the surface and the bottom of the Harbor repeatedly. During the two-week experiment she collected daily water samples and performed two 24-hour sampling periods where she collected samples around the clock (with some help from her labmates). She says that she will email me the sampling plan for that two-week experiment. When I look at it later, I see that it outlines the research goals, sampling plan, and other necessary field work for the experiment.

So, Maria says, she has physical samples and sensor data from both continuous sensing and from short-term data collection "campaigns," meaning that she has time series as well as point sampling data. But, Maria continues, she is also investigating the effects of the daily solar cycle and the bi-daily tidal cycle on microorganism growth, and thus has light and tidal measurements from two other kinds of sensors. She says she is still figuring out how to deal with the complexity of analyzing all of these kinds of data, saying "I'm not a statistician."

She opens a few folders on her computer and comes to a folder labeled "time series." This is where she stores the sensor data from the Harbor. She has separate folders for 2009 and 2010. The files of sensor readings are named by the location and date that they were downloaded, as in "H_031510," where "H" stands for "Harbor." She also has a

sub-folder called “light,” which contains the readings from the small light sensor that Evelyn attached to the outside of the WQM sensor package (described in Vignette 5).

Maria says that now that Evelyn is doing the bi-weekly field maintenance and data collection trips her data are more organized than they were before. Maria says that the aquatic team had a learning curve for running the sensors, including having to develop calibration and documentation methods. For example, she describes how she has a sub-set of her sensor data from 2008 that has a different depth reading than the data before and after it in time. As she says, all of the sudden the depth reading goes from 0.5 meters to 2 meters without obvious reason. Maria said that she has the notebooks from the technician who was doing the data collection at that time, but is still trying to figure out why that depth change happened, “*I haven't been able to figure out from the [former technician’s] notes if that's a real change in where that sensor was deployed or if it's, you know, if the sensor got sent back and was put back in water without calibrating the depth sensor or pressure sensor*.”

Maria then opens up an Excel file to show me how they keep track of their sensor calibrations. The file, named “Sensor_field_notes,” contains information related to sensor calibrations they have performed recently. It has date-by-date listings, dating back a little less than a year, of the sensor calibrations that are associated with individual sensor data files. For example, it lists the pre and post-chlorophyll calibrations, which I had discussed with Evelyn during previous field trips. I asked Evelyn if she transcribes the calibration from her field notebook into this file. She says yes. I ask if the file is continually updated when she performs a new calibration, or if there are multiple versions of the file. Evelyn

says that there is just one file, and they keep adding successive calibrations to it. Maria shows me how there are a few blank cells where calibrations did not get done, like for one week in June. I ask Evelyn if that was the time when I was out with her and she had problems connecting to the sensors. She says yes.

I ask Maria how she will use the chlorophyll calibrations to adjust her sensor data. If she has two weeks of data, and the calibration information indicates that the chlorophyll value shifted by some value over those two weeks, how does she adjust the full data set? Does she generate an equation? She says yes, though she has not done it yet. She says that if she has 600 data points over the two weeks, and the chlorophyll value shifts by 0.5, then you would adjust each point by 0.5/600. She mentions that she sometimes has to back fix data if the sensor calibration was bad. I ask how she knows how far back to go when back-fixing data. She says that you just look at the data to find where calibrations have gone wrong. You will be able to see where data values are wrong.

At this point, another student knocks on the door and comes into Maria's office. He wants to ask Maria about a "recipe" for making a certain lab solution. They talk about that for a few minutes.

After the student leaves, I ask Maria about physical samples. Maria shows me the "raw" data from a water sample analysis: an Excel table of numbers which represent organism counts. She stores the physical sample counts in separate files from the sensor data. Maria says that to produce the organism counts from a particular water sample, she looks at a portion of the sample under a microscope, identifies what kinds of

microorganisms are in the sample, and then make counts of each species that is present. I ask how she identifies organisms. She says that you can identify them visually based on the physical characteristics of the organisms, such as shape, size, and appendages. She pulls a book off the shelf and shows me pictures of microorganisms. She says that she can also use this book as a reference if she is unsure how to identify a particular organism. To produce the organism counts, she does not count each cell in the full sample, or even a full microscope slide's worth of organisms. Maria describes how she counts about 200 cells at random places on the slide, and from there calculate the relative biomass. She says that she will also take microscope pictures of organisms within samples from time to time. She can then use those pictures for her own reference, as well as using them as illustrations in papers and presentations.

I ask how long a frozen sample will last. Maria says that samples held at -80 C will last basically forever, and that at -20 C, the temperature at which her samples are stored, they will last a long time. I ask how they get back to an individual sample, and how they correlate a physical sample with a sensor data set. Maria says that they are all labeled with the same name, for example "H_040809," which they can then use to track across all kinds of data.

I say, "just to pull back a bit to an earlier discussion, You said that you are more organized now that Evelyn is doing the sampling. What's the difference between what you are doing now and before?" Maria says that the biggest difference is that Evelyn has keeps a strict schedule for doing the sampling and sensor data collection. Because of this regular sensor maintenance in the field, they are catching problems faster. Evelyn says

that they also have redundant sensors, the WQM and the Hydrolab. Maria agrees and says that this helps with data consistency. I ask if the WQM data are stored in the same "time series" folder as the Hydrolab data, and they say yes. Maria opens up the folder again, and shows me. She says that the data from the two kinds of sensors are clearly marked, the one has a "WQM" in the file name, and the other does not.

I ask if anyone else will use this data. Maria says yes, people occasionally will use her sensor readings or samples. Maria goes out into the main common area outside her office and pulls out a few folders that are sitting on a shelf. One is a sample preparation method binder for the various methods that they train people to do. I ask whether the methods in the sample preparation binder are standard methods that I might find any lab at any university doing. Maria says yes, and pulls out a book, *A manual of chemical and biological methods for seawater analysis*, by Timothy R. Parsons, Yoshiaki Maita and Carol M. Lalli (1984), which she says is the standard manual that everybody uses for sample preparation.

The other binder Maria shows me is a binder of sample data showing which organisms can be found in any particular sample. Each sample is labeled with codes like D, R, C, which mean:

- D = Dominant (>50%)
- A = Abundant (25-49%)
- C= Common (10-24%)
- P = Present (1-9%)
- R = Rare (<1%)

Maria says that these demarcations are based on live observations of samples and are "largely qualitative," to be used as a quick indicator of which samples might be useful to someone else in the lab.

5.3.4 Commentary on Vignette 6

This vignette vividly illustrates a data deluge. In this case, however, the deluge is not measured in megabytes or gigabytes, it is measured in numbers of files, data types, sensors, and calibrations. For example, one data file downloaded from a Hydrolab sensor containing two weeks of sensor readings is 63 kilobytes in size. Thus, memory and storage are not a problem. Integrating hundreds of such data files, however, is a considerable task. In addition, as Maria describes in this vignette, these sensor readings must be integrated with organism counts from water samples, pressure sensors that measure tidal cycles, and light sensors that measure solar cycles.

Compounding this complexity is the unreliability of the sensors themselves. On numerous occasions Maria remarked on the unreliability of the Hydrolab sensors, which I documented in Vignette 5. Because the Hydrolab sensors are so unreliable, the team installed the WQM sensors at each Harbor location. The WQM sensors largely measure the same parameters as the Hydrolab sensors, but according to Maria are much more reliable. Thus, she will have duplicate data sets for the periods in which both Hydrolab and WQM sensors are installed. Maria described how this will allow her to "*do comparisons between the dissolved oxygen and chlorophyll sensors that are shared between the two types of sensors. The WQM should be a little bit more robust, a little bit less sensor drift, things like that*."

Unlike Vignette 4, in which Claire did not use the calibration information from her calibrations of the Graphite Furnace machine, Maria uses the information from Evelyn's calibrations of the Hydrolab chlorophyll sensor to adjust the sensor values to account for bio-fouling. She gets this information from Evelyn through a shared Excel spreadsheet. Evelyn records the calibration information in her notebook while in the field, transcribes it to the shared Excel file once back in the lab, and Maria is able to access it from there.

The shared Excel file, however, only contains calibration information for the chlorophyll sensor. As Maria describes, the other sensors are less of a concern from a calibration point of view, "*Temperature and conductivity, as long as you're keeping the sensors clean shouldn't jump very much. …[W]e basically just kind of check, make sure that it looks decent. … And every time the sensors get recalibrated, we do the full suite*." The Excel file does contain notes about other sensors, including the WQM and light sensors, such as "Light Sensor deployed on WQM," which relate to a particular data file or data collection trip. The full calibrations, however, are recorded in notebooks and shared as needed amongst the team.

As this narrative illustrates, the data deluge takes on different forms depending on the characteristics of the project. Table 5.6 illustrates the kinds of data identified in this vignette. For the CENS aquatic biology team, the challenge is in integrating many small data sets from many different sensors, data collection campaigns, experiments, and analysis methods. This integration also involves adjusting data based on sensor calibrations. As Table 5.6 illustrates, records of these calibrations are one source of

metadata for this integration process. Sensors produce headers that describe sensor sampling rates and intervals, and researchers record many details of their lab and field work in their notebooks. Pulling these sources together is a technical, social, and organizational task.

Table 5.6 – Data and Metadata for Vignette 6 and Commentary

Data

- Continuously recording sensor data
- Regular collection of physical water samples
- Short term collection of sensor data and water samples (two-week long experiment)
- Light and tidal measurements
- Excel table of numbers which represent organism counts from water samples
- Microscope photos of organisms in a particular physical sample

Metadata

- Folder labels: “time series,” "2009," "2010"
- Sensor data filenames: named by the location and date that they were downloaded
- Notes in technician's field notebooks
- Excel file containing information related to sensor calibrations and field sensor maintenance
- Reference books for physical sample preparation and identifying microorganisms
- Maria describes how they do a count of about 200 cells at random places on the slide, and from there calculate the relative biomass
- Physical sample labels which use the same syntax as sensor data filenames
- Research planning document that details field methods
- Sample preparation method binder with laboratory methods
- Binder that lists which organisms can be found in physical samples held in the lab

5.4 Soil ecology

Ecology applications for CENS sensing technologies range widely. CENS technologists have worked with ecologists to develop sensors for both above ground and

below ground sensing. In my study, I focused on the CENS projects dedicated to below-ground soil sensing. The CENS soil ecology project uses a combination of sensing systems. First, they are developing a suite of technologies that are able to collect high resolution images of soil, roots, and other underground biological phenomena. Researchers bury one-meter long clear plastic tubes in the ground. The tubes are buried at an angle such that one end sticks out of the ground. CENS researchers have developed both manual and automatic imaging systems that fit into these tubes, as illustrated in Figure 5.9. To manually collect images, a researcher inserts a camera system into the tube on the end of a rod. The camera is connected via a cable to a computer, which allows the researcher to see the field of view and hit a button to capture an image. The automatic imaging system, by contrast, is installed in the tube with an electro-mechanical gear system that controls the camera motion and image capture process. CENS researchers can access the automatic imaging system remotely, via the internet, and start the imaging process, scheduling the machine to start collecting images in a pre-defined pattern at any specified time.

Figure 5.9 – Soil ecology field installation.

Second, beside each imaging tube, CENS researchers buried an associated set of sensors that capture soil moisture, soil temperature, and carbon dioxide (CO_2) readings at three depths: 2, 8, and 16 centimeters below ground. These sensors provide environmental parameters with which CENS ecologists can correlate the growth of root structures as identified via the images.

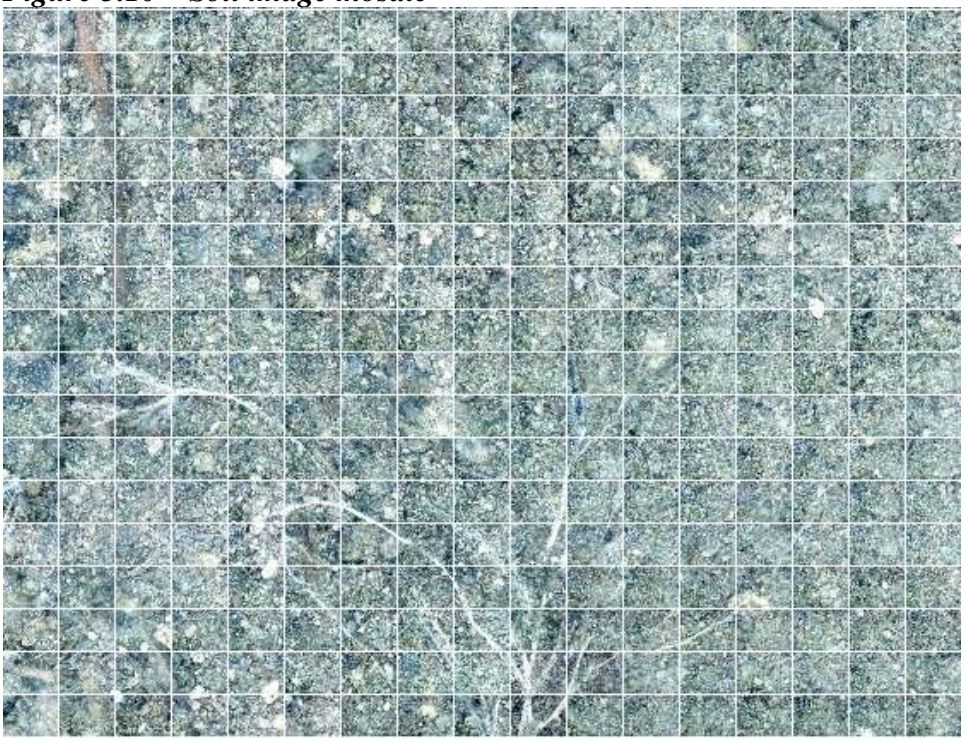
Figure 5.9 shows a field installation for the CENS soil ecology project. In Figure 5.9, the cables emerging from the three pipes on the top right of the image are attached to the underground temperature and moisture sensors. The two capped tubes at bottom right are used for taking soil images manually. The large triangle at left is insulation covering the automated soil imaging machinery.

5.4.1 Vignette 7 - Soil ecology image data

Caroline's desk is located in the main Ecology lab. Caroline, a staff researcher on the CENS soil ecology project, was hired less than a year ago to work out a method for analyzing the root images that have been collected by the automatic imaging system. I am familiar with the broad outlines of her work, having spoken with her before on a previous visit to her lab, and having visited on multiple occasions the field site where the imaging systems are installed. The first time I met Caroline, she said that when she was hired, her task could be summed by her supervisor's statement, "we have 500,000 images, can you do something with them?" Having seen presentations of the CENS soil imaging system at CENS research reviews and poster sessions, I am aware of the fact that analyzing these soil images has proven very difficult to automate. Computer vision specialists within CENS have tried to use image analysis techniques to measure root structures in the images, but have not had much success. Thus, I am interested to know what kinds of processes ecologists have developed in order to analyze the images. I have asked Caroline if she would walk me through her processes for working with the image collection, as she is now leading that effort.

She starts off by showing me an online image database that was built by the lead imaging system engineer as a tool to store and display the soil images. Caroline shows me how you can look at scheduled soil imaging runs, completed imaging runs, and imaging runs that are currently in progress. Within the database, selecting an individual imaging run returns a mosaic of all the images collected during that run. Figure 5.10 shows a partial soil image mosaic from this database.

Figure 5.10 – Soil image mosaic



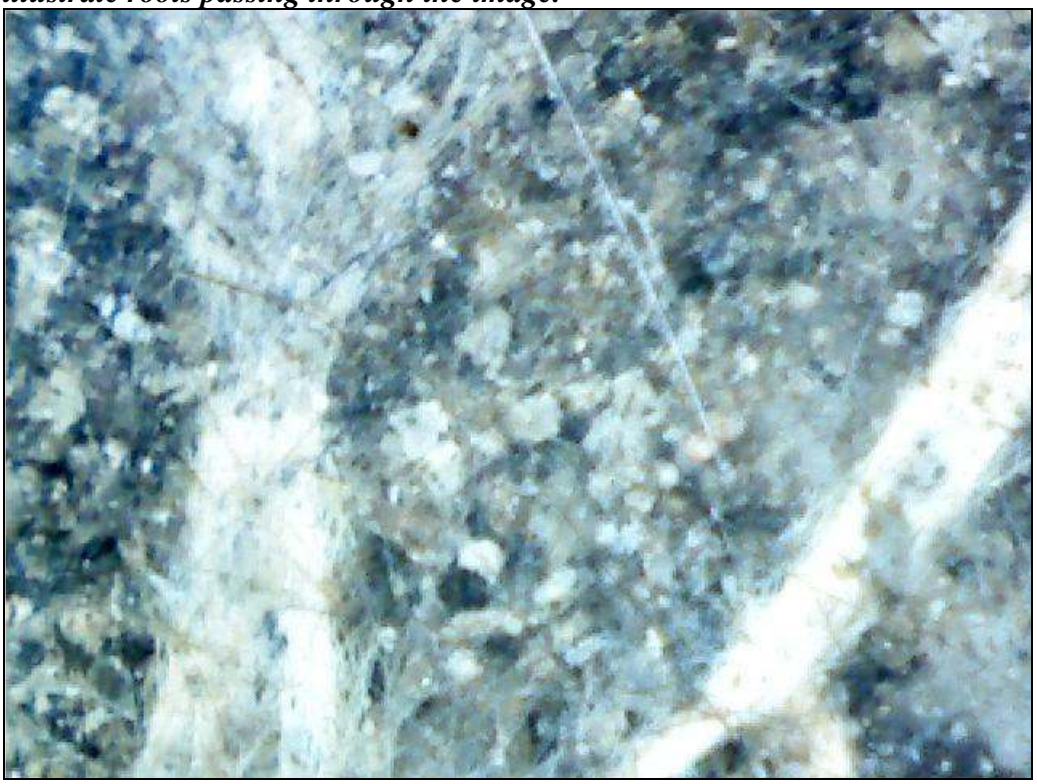
On the screen next to each mosaic, the database displays a number of characteristics of the imaging scan (shown in Table 5.7). These characteristics include when the scan was taken (Scan Time), where the scan started and stopped in the tube (Starting and Ending X and Y), what the size of each image is (DX and DY), along with a number of others. (Later, in an interview with the ecology Principal Investigator, these characteristics were called out as the image metadata, "*The images [metadata] is really much more descriptive, and again it's what's the resolution, when was it taken, where was it taken? And that's pretty much the main, the main metadata for that*.")

Table 5.7 – Image scan characteristics displayed with soil mosaics

- Scan ID: 3033
- Name: Soils 3 Recurring Scan (18:00) 1X per day
- Scan Time: 2011-03-25 18:00:00
- Starting X: 100 mm
- Starting Y: 100 mm
- Ending X: 150 mm
- Ending Y: 600 mm
- DX: 3.01 mm
- DY: 2.26 mm
- Dwell Time: 150 ms
- Scan Lines: Horizontal
- Scan Mode: Raster
- Step Units: mm
- Start Time: 2011-03-25 18:00:00
- End Time: 2011-03-26 00:35:06
- Scan Status: Completed
- Root grows down: No
- Notes:
- User: [name]
- Total number of images: 3774 (17x222)
- Total Disk Space: 417.557 Mb
- Total Travel distance: 22699.65 mm
- Estimated Scan Time (HH:MM:SS): 06:50:16
- Scan Time (HH:MM:SS): 06:35:06

Caroline says that she does not know what all of these characteristics mean. She knows some of them, like start time and X/Y coordinates, but not all. She then selects an individual image within a mosaic, which pops opens a larger version of that image. Figure 5.11 shows an example of a single image.

Figure 5.11 – Image taken by automated soil imaging system. The thick white lines illustrate roots passing through the image.



Caroline then shows me how the database allows her to select images by location, where "location" means orientation on the tube. Each "location" is a specific spot on the tube that has been imaged as part of an imaging run. Off the top of her head she says that they have 4,158 images per "run," with a run being an individual scan of the tube. She does not use all of the images from a run. Instead, she selects sub-sets of the images via targeted sampling. She looks through the images on the online mosaic visualization, and decides which locations to use for analysis based on the presence of root structures that she would like to investigate further. She selects multiple locations from a mosaic, so she can have "replications" to analyze. For example, Caroline says, she wants to look at the daily patterns of root growth and die back. So she takes some number of image locations,

say 30, then gets all the images for those particular locations that have been taken over time, and once she has all of the images from her chosen location in hand, she can analyze the images at each location as a time series. Thus, she would have 30 individual time series of images to analyze.

In order to actually acquiring the images that she wants to analyze, she works through the main engineer for the imaging system. She provides a list of tube locations to the technician, who then compiles the images for each location into files, which he then puts on a group server. Caroline can then download the image files to her computer using the FTP protocol. She says that this system works pretty well, she has to go through the technician to get the images, but he is very good about getting the image series' back to her. In a later discussion with the main engineer, he indicated that he was working on a way to automate this process to allow Caroline, or other members of their team, to download the images that they want without having to go through him.

Caroline says that before they developed this online database, she would compile the time-series of images for each location "by hand" using the filenames as identifiers for particular locations. She describes how the image filenames have a particular syntax, which included the date-time stamp, the tube number, and the X and Y coordinates (which indicate the location in the tube where the image was taken). Since the filenames have the X and Y coordinates in them, she could do a global search in the database for all images with the appropriate X and Y coordinates in the filename. She says that this process worked to compile image runs, but it would tie up her computer for a long time because they have so many images to search.

I ask about the filename syntax. Caroline cannot remember the exact syntax of the filenames off the top of her head, so she looks through a few documents to try to find it. She looks at a document in her computer, and also through a binder of papers that she has by her desk. When she does not find what she thought was the right explanation for the image filename syntax, she asks an undergraduate assistant at the next desk to pull up a raw image file to show us the name. The undergraduate looks around on her computer, but only finds images that have been renamed, so Caroline tells her to look on the external hard drive that was attached to the computer. The undergraduate assistant pulls a file up from the hard drive. Caroline says that the filename syntax has changed for the 2010 images; it is similar to the 2009 syntax, but slightly different. We look at images from both 2009 and 2010, and can see that the date-time stamp was re-ordered, but the differences are hard to determine because the filenames are about 20 digits long with dashes. Caroline sits back down at her computer, looks at the document she has open, and realizes that it is, in fact, the document that describes the filename syntax after all.

I ask Caroline about the re-named images she found during while they were looking for the raw images. She says that once the time series' of images have been compiled into files, the image files need to be renamed because the root image analysis software that she uses requires the images to be named in a particular syntax. The root image analysis software was developed at a non-CENS university, so its specified image filename syntax is different from the CENS image filename syntax.

Caroline suddenly starts laughing, and says that she just remembered that she had written up a methods document a few months ago that describes this whole process that

she is showing me today. She says she forgot that it existed. She looks around on her computer for a minute or two, finds it, and prints out a copy. It is a five page long single spaced document that details 12 steps of her root image processes, from how to download and run the necessary software to how to record data. She says, “I should just follow this in explaining the rest to you.” So she started going through the rest of the document, continuing where we had stopped: image renaming. They use a program called “File Renamer” to rename the image files. The root image analysis software needs the images to be named using a “project_tube#_imagelength_etc” syntax. Caroline says that she uses a program called “File Renamer” to rename the whole run of images with this syntax. During the re-naming, the image files also get assigned a number from “001” to “300” in order of their time series sequence. But, Caroline says, she names the images in reverse sequence, that is, the most recent images are given the lowest numbers, and the image numbers are incremented going back in the time series. She uses this numbering scheme because she does her image analysis starting with the most recent images and then going backwards.

She stores the time series images in a set of folders on her computer. Each time series of images has its own folder. Once she re-names the files, however, the X and Y locations are dropped from the filenames. Because the X and Y locations are critical to identifying the images, she includes the X and Y coordinates in the name of the folders to ensure that she can easily find them if necessary.

5.4.2 Commentary on Vignette 7

Researchers rarely use all of their data. Instead, researchers select and filter their data, pulling out particular chunks or coherent sub-sets. In this narrative, Caroline showed me her processes for identifying, selecting, and then creating sub-sets of images. For Caroline, the relevant unit of analysis is an image time-series. In order to study the growth and die-back of roots over time, she must have images of particular roots as they grow and die. This involves knowing two things, 1) that she is looking at the same roots in each image, and 2) when and where her images were taken. As Caroline described, to look at the same roots over time, she finds a particular location within the tube where roots are growing. From there, she needs every picture that was taken of that particular location over the time period in which she is interested.

Her primary source of metadata in this process is the image filename. Each image is automatically assigned a filename by the imaging system. The imaging system engineers developed the filename syntax to facilitate retrieval of particular images. As the chief engineer told me: "*The file name itself ... tells you everything you need to know about the file. It tells you basically what machine, what drive it's on. …I mean it's got the directory, …a subdirectory. And then there's a unique ID set up for every scan so it has a scan ID. Then it has the date and time and I mean down to hour, minute and second. Then it has the XY position of the motors, so you know exactly where that image was taken on what machine and then where it should be looking, so... And that's the same data we use to retrieve the images*." Note also in this quote that he refers to the filename as "data" in the last sentence. At an earlier point in our interview, however, in response to

my question, “what kind of metadata are you collecting?,” he said, “[w]*ell, of course the date and time, filename…*,” indicating the filenames’ ambiguous status as data or metadata depending on the topic of discussion.

The filename syntaxes themselves, however, have changed a few times during the project. Caroline described how the images from 2009 have a different filename syntax than the images from 2010. She showed me a document that she had created that outlined what all of the components of the filename mean.

In the second half of the vignette, Caroline showed how she has to manipulate the image filenames once she has pulled time-series of images out of the full image collection. Her analysis software, developed outside of CENS, required its own particular filename syntax. Because the outside software was not created with the automated imaging system in mind, it has different filename requirements. The X and Y coordinates of an image, for example, which are so critical to identifying images that are collected using the CENS automated imaging system, fall completely outside of the analysis software’s assumptions. All conventional soil imaging systems have no X and Y coordinates. Thus, when Caroline re-names her time-series of images, she has to remove the X and Y coordinates from the filename of each image. As a work-around, she records the X and Y coordinates in the name of the time-series folder, because each image within that folder should have the same X and Y position.

Thus, metadata facilitate data selection and filtering. In this case, the main metadata used for selection and filtering of images are the image filenames. Table 5.8 illustrates other forms of metadata identified in this vignette. The metadata displayed by

the image database for a given imaging scan, listed in Table 5.7, are not used in the selection and filtering process. Additionally, as sub-sets of image data are created, researchers create new metadata, such as new filenames or folder structures. Image selection and filtering are one of the first steps in the analysis process. The next vignette picks up where this one left off, focusing on how images are turned into numbers in the course of building an argument.

Table 5.8 – Data and Metadata for Vignette 7 and Commentary

Data • Images of soil, roots, and other underground biological phenomena • Sensors that capture soil moisture, soil temperature, and carbon dioxide **Metadata** • Image scan characteristics (Table 5.7) • Image filenames as identifiers for particular images and imaging locations • Document describes the image filename syntax • Document that describes her image analysis methods • Folders of time-series of images, with folder names that identify what is inside

5.4.3 Vignette 8 - Soil ecology numerical data

Continuing from Vignette 7, I ask Caroline how she organizes all 500,000 images. She is a bit confused by the question, so I ask again, how does she look at all of the images? If, for example, she wanted to find an image from a particular day, how would she go about finding it? She says she can show me the "actual data." She gets up and goes to the lab bench, and takes some papers out of file boxes. She says that these are the actual data collection sheet, and shows me a bunch of gridded sheets full of numbers.

Her goal is to develop quantitative models of root growth and die-back. The numbers on these sheets are measurements of root lengths in individual soil images. To

get the numbers recorded on these sheets, Caroline says, she has to go through a multi-step process. The general work flow is the following: with the folders of time-series images for the sampled locations in hand, the images are analyzed one-by-one with the root image analysis software. The software has an interface that allows her to trace lines over the root images, indicating where roots appear in the image. The software then uses these traces to calculate the amount of roots that appear in the picture, giving a number that quantifies the roots in the image. Caroline, or her undergraduate assistants, then records the number by hand on the paper data collection sheets. To get these numbers back into the computer, they then transcribe these paper sheets into Excel spreadsheets.

Caroline shows me an Excel file for a particular analysis. This particular Excel file has 50 individual worksheets. About half of these worksheets are transcribed values from the paper data collection sheets, and the other half are various calculation steps. Caroline says that it requires this many steps in Excel to get from the numbers they record on the paper data collection sheets to the final value of interest, which is root growth and dieback per time unit.

Caroline says that she created template Excel sheets for herself and the undergraduate assistants who work on the project. The Excel templates explain the analysis steps in step-by-step fashion. Caroline shows me what the Excel template looks like. It is blank except for column headings and a series of numbered instructions. These instructions include details about where to copy and paste columns from one sheet to another, where to sort columns, and where to perform mathematical operations.

Going back to the actual data Excel file, Caroline shows me how they make red flags in the Excel sheets to leave notes. She shows me a few notes. Each note is attached to a particular cell in the Excel file, and, because multiple people take part in the analysis, each note indicates who wrote it. The notes range from simple, "analysis 3 completed on 2/23/10," to detailed, "YOU NEED THESE TO COMPLETE THE FORMULA FOR CELLS THAT CHANGE (limited to one day change): =Adjacent cell + (23:59:59-Below cell) + 00:00:01." Caroline says that the notes are mainly to help her remember important things about the process.

One of the notes, attached to a cell heading named "Calibration (mm)" says "Same as [previous column] but divided by 10 to account for [software] calibration." I ask Caroline what that means. She says that they need to adjust the values in that column to correct for the image analysis software. The image analysis software is not designed to analyze images at the high resolution at which the CENS images are taken. I do not follow her the first time she explains it, so she shows me the paper data collection sheets again, and shows me the numbers, pointing to one cell that says "4.73mm." Caroline says that the true value is actually 0.473mm, but the software is only configured to give two digits after the decimal point. So, she tells the software that her images are 30.00mm wide instead of the actual 3.00mm wide, so that she can get an additional digit of precision in her values. Otherwise, the value of 0.473 would get rounded to 0.47. Because of this work-around, all of their numbers on the paper data collection sheets are inflated by a factor of ten, and thus, they have a step in their Excel analysis process where they re-

calibrate the numbers by dividing by 10. Caroline says that the note in the Excel sheet is there to remind them of why they are doing the division by ten.

As our discussion is wrapping up, I ask Caroline if she had any thoughts about what she would like for data management if she could change anything. She indicated that it would be very nice to have some way of automating the analysis steps that they have to do repeatedly in Excel. She has to do the same Excel machinations over and over again, copying columns, dragging cells, performing calculations. If they have 30 time-series of images, she has to perform the same dozen Excel actions 30 times. Caroline's idea is to put together a database that automates some of these steps, so that you just enter the values that they currently write on the paper data collection sheets into an interface and the system would give you back your root growth value without all the Excel copy and paste work.

5.4.4 Commentary on Vignette 8

Quantitative analysis is the standard in most scientific arguments. Mathematical models, statistical analysis, numeric graphs – rarely does a scientist make a claim without some form of quantification. Hence, the prototypical conception of "data" is a table of numbers. Vignette 8 illustrates how images are converted into numbers in the CENS soil ecology project. Caroline uses the image analysis software to quantify the volume of roots in individual images. She then records the values generated by the software on paper data collection sheets, transcribes these paper sheets into an Excel file, and performs a number of Excel manipulations to come up with final values for the parameter that she wants: root growth and die-back per time unit.

A big challenge in this process is that the CENS root imaging system is still a developmental technology. When asked about the main research question for the project, the Principal Investigator, a senior ecologist, said, "*Can we build it*?" Developing the imaging technology, the sensor packages, and the user interfaces have been the primary goal. Along the way, however, the hundreds of thousands of high quality soil images that have piled up have tantalized the ecologists. The lead engineer for the imaging system describes this dynamic nicely:

"*Well, it was a real struggle because I wanted to get first light on the instrument. And soon as [the senior ecologist] saw the images, he was like 'Oh we got to take more of those'. And it's like, well, …we pull the thing out of the ground most every day to tweak something. It's like we're still in development and he wants to do science, so it's always a constant battle of those things…. [But] it's really fun. I mean, it was challenging from everybody's standpoint. And it's very nice that Caroline and everybody is very understanding of, they've all been very good about that. I'll say I'm sorry we missed a month of images because X happened. And [they say] 'Oh okay'*."

The analysis practices that Caroline described in this vignette are thus newly developed as well. The labor-intensive hand transcription and Excel manipulation serve as a test-bed for the development of new analysis methods and processes. As the ecologist PI states, a lot of her work is "*still methods development and then picking pieces and asking specific little questions that we can address based upon the visual imaging, coupled with the sensor type of data*."

The novelty of their process was not lost on Caroline. A few times during our discussion she remarked at how explaining her analysis process to me was harder than she expected, even though she had developed a number of metadata processes: creating Excel templates, writing Word documents, and making notes in the Excel data files. The metadata flags in her Excel files, as noted in the vignette, were associated with individual cells. Excel indicates that a cell has an associated note by adding a red marker in the corner of the cell. The note is displayed by clicking on the red marker with the mouse.

The Excel files reflect Caroline's final comments about the repetition involved in her analysis processes. Whole tables have clearly been copied from one sheet to another, with metadata notes intact. Differences among notes indicate where some change was made from one analysis to the next. For example, in one Excel data file that contains twenty separate root image time-series analyses, the worksheet for the fifth analysis contains a column titled "Change in Length (mm)" that has a note from Caroline saying, "DID I DO THE CHANGE OF LENGTH INCORRECTLY ON ALL THE OTHER ANALYSIS (IN THE WRONG DIRECTION)?!!!!" The "Change in Length (mm)" column for the next analysis, the sixth, has a note from Caroline saying, "This is the correct way to determine change of length." All subsequent analyses, the seventh through the twentieth, retain the "DID I DO THE CHANGE OF LENGTH INCORRECTLY…" note, indicating that they were copied and pasted from the previous worksheet without the note being changed after the pasting.

Converting images to numbers is a multi-step process. In developing a novel method for quantification, CENS ecologists use custom software, open-source software

developed outside of CENS, paper worksheets, and general purpose data analysis tools like Excel. Because methods development has been an explicit goal for the project, the researchers on this team have been conscientious about developing metadata processes for documenting how they manipulate and transform data from one form to another. Table 5.9 illustrates what those data and metadata are in this process. These forms of metadata, such as the Excel templates that outline analysis steps, and the annotations within the actual Excel data files, serve as tools in making arguments about why this is a legitimate data analysis method, and why their findings are scientifically valid.

Table 5.9 – Data and Metadata for Vignette 8 and Commentary

Data • Images of soil, roots, and other underground biological phenomena • Gridded paper data collection sheets • Excel spreadsheets transcribed from paper data collection sheets **Metadata** • IDs, units, name of analyst on paper data collection sheets • Template Excel sheets that outline analysis steps, including column headings and numbered instructions • Notes in Excel data sheets, identified by red flags

6. RESULTS II – EXPLORING METADATA PRACTICES

The vignettes given in Chapter 5 illustrate typical settings, situations, and practices of CENS science research, and identify how data and metadata manifest in CENS field-based research. In this chapter, I compare and contrast the data and metadata practices across my four cases studies in two ways. First, I draw out four analytical themes that cross-cut my CENS case studies. These themes illustrate particular types of data and metadata practices that were observed in each case study. Second, I draw on the practices discussed in analyzing these themes to build out the lists of data and metadata provided in the Chapter 6 vignettes. I illustrate the commonalities and differences in the kinds of data and metadata products and processes that CENS researchers use. With these comparisons in hand, I then look at how metadata practices are manifested in CENS researchers' published papers. In looking at published papers, I see how researchers present their practices to academic peers, and what is not presented.

6.1 Analytical themes for examining data and metadata practices

In looking across my four case studies, particular themes emerged. By "themes," I refer to issues that I observed to be of importance in examining the data and metadata practices of my study participants. The themes are: 1) processing data from raw to forms suitable for analysis, 2) assessing data quality, 3) distributing metadata tasks, and 4) developing metadata products and processes over time. I use these four themes to isolate

particular aspects of metadata creation, use, sharing, and preservation. I also identify kinds of data and metadata that researchers create and use that were not included in the vignette narratives.

6.1.1 Processing data from raw to forms suitable for analysis

As the vignettes illustrate, the data that CENS researchers actually use in their analyses are many steps removed from the original sensor readings or machine output. Researchers perform transformations, calibrations, and selections to manipulate data into a form that can be input into analysis tools. I refer to this set of practices as "data processing." Metadata are an essential part of data processing and analysis. Existing metadata might be used to facilitate processing steps, and new metadata may be created during the course of processing and analysis activities.

6.1.1.1 Data processing - seismology

Seismic sensors record continuous streams of measurements. Seismic research, however, is based on analysis of discrete earthquake events. In the CENS seismic project, processing focuses on pulling these discrete earthquakes out of a large database of continuous recordings. Members of the CENS seismology team use the Seismic Transfer Program (STP) to download data from the main project database in the SAC format. STP was developed by the Southern California Earthquake Data Center to enable users to access seismic data in an automated way, and is available as a free download. CENS seismologists use STP as the main tool for retrieving data from their database. As a computer scientist on the CENS team described:

"*What the STP server does is it um it provides a nice interface for seismologists and everyone to get the data in whatever chunks, whatever length, whatever chunks of data you need. It's just a little command line or web interface where you say, 'I want the data for these stations or this one station for these channels for this hour and a half long period, or for this two minute period, or for this 10 day period.' And it will just deliver all the data in the SAC format to you*."

The SAC data are delivered with a header. The header is an important metadata component, and lists a number of properties of the downloaded data, including the number of points, the sampling rate, the minimum and maximum amplitudes, and the location of the sensor. Researchers can also use the SAC headers to find information about earthquake events that are contained within the data, but, crucially, only after they have added earthquake event information to the headers themselves. As one seismologist stated, "*the event depends on the user*." Different users of the data look for earthquake events of different magnitude. For example, one analysis of the Mexico data looked at earthquake events larger than size 5.8 on the Richter scale, while another looked at events larger than size 6.0. Another student analyzed events larger than size 6.5 from the Peru data.

Researchers must create this event metadata themselves in the SAC headers by using an "event size" field in the headers. Researchers can either directly input an event magnitude into a SAC header, or they can download catalogs of earthquake events from the United States Geological Service (USGS) and create the event annotations using the USGS catalog. As of this writing, the CENS team was using the USGS catalog for the

Mexico data, but did not have a USGS catalog for the Peru data because the project was still ongoing.

6.1.1.2 Data processing - environmental science

Researchers use Excel as their main analysis tool in the CENS environmental science project. Thus, much of the data processing in this project involves researchers getting data into Excel in a form that they desire. Vignette 4 describes how Claire writes output values from the Graphite Furnace in her notebook and transcribes them into Excel by hand instead of using the machine's built-in output format. She then organizes her Excel files by the dates in which she uses the machine and by the sequence in which she extracts contaminants from her samples. Both the analysis date and the contaminant extraction sequence are found as metadata in her sample labels. By recording the values in her notebooks and manually adding them to her own Excel file, she can format her data in any way that she needs.

Claire was not the only member of the CENS Environmental lab who manually transcribed values from a lab to a notebook and into Excel. Another student, Samantha, uses a quantitative Polymerase Chain Reaction (qPCR) machine to detect and quantify the amount of DNA in physical samples. In her case, the qPCR machine outputs a file that contains multiple forms of metadata, including calibration curves and pictures of the machine set-up, along with charts of analysis values. Samantha described how the output file can be exported into Excel or Word, but similar to Claire, Samantha said that she does not use the machine's output format because it is not in the form she needs for her analysis.

Thus, all of the other metadata that is contained in the machine output files, such as calibration information, instrument identification, sample setup and machine initialization, are dissociated from the data that are ultimately used for the analysis.

6.1.1.3 Data processing - aquatic biology

Vignette 6 described one of the main forms of processing in the aquatic biology project, namely applying calibrations. Another processing activity is removing artifacts that the sensor introduces into the data. In particular, this means removing outliers from the data. Some outliers are obvious, like a value of 20 million surrounded by values of less than one, but outliers can also manifest as much smaller changes that still fall outside of real phenomenon.

Researchers use their personal knowledge and experiences to determining when sensor values fall into this latter type of outlier. Maria, the lead student on the project, gave this example, "*where it goes from 0.05 volts up to 1 volt, like that's a really huge change. It's not really feasible in a 30-minute time period*." For example, researchers know that a 50 degree change in temperature in 30 minutes is "not feasible" because the ocean temperature will never change that much that fast, and so sensor values that display such a change are related to equipment malfunction. When outliers are identified, researchers remove them manually. Maria described how she determines outliers by calculating standard deviations and then removing any points that fall outside of five standard deviations from the norm. When I asked why she uses five standard deviations, Maria said that the literature she found suggested using two standard deviations, but using two standard deviations resulted in her removing data features that in her

experience did not look like outliers. "*So we do a little just playing around with different parameters, plotting it, seeing you know, 'Okay, here's the original time series, Here's what was taken out...' I'm [comparing two standard deviations] versus three versus five*."

Individual data points that have been removed must then be replaced by another value in order to retain continuity. Maria described how she fills in such gaps with interpolated values. "*I'm really just interpolating those missing points. Like, once I remove the outliers [I'm] basically putting those data points back in based on an interpolated value between the two samples on either side*." The Principal Investigator on the project describes how in his terms this "massaging" is a critical component of using this sensor data:

"*When you get the raw data in and it has to be considered just that, raw. It's not processed data. So …[we have a] very specific set of algorithms the [we] put the data through to make sure that it's quote-unquote massaged, so that any bad information is taken out. And we always try to err on the side of being conservative, meaning I'd rather throw data away than be concerned that I may have data that I can't trust. So we go through that process which, for us, converts it from quote-unquote metadata into real data. But that's kind of the massaging of the data processing that we do*."

This quote also shows an interesting use of the term "metadata." Here he uses it to refer to data that has not been processed in any way. In the same interview, he also used "metadata" to refer to compiled data sets that are used for meta-analyses.

6.1.1.4 Data processing - soil ecology

The process of going from images of roots to numeric values, as illustrated in vignettes 7 and 8, involves numerous processing steps: visually scanning image mosaics, selecting and extracting series of images, marking up root images via specialized software, recording values onto paper data collection forms, and transcribing those paper forms into Excel. The processing in this case requires considerable manual work. This manual effort includes metadata processes, such as developing Excel templates, annotation practices, and writing up internal methods documents, and allows the images to be analyzed with great precision. As Caroline, the main student on the project notes, this is because "*it's a brain and a computer together that is quantifying them*." Machine vision techniques are not yet able to detect roots in images as well as humans. Both Caroline and her ecology PI noted that they believe that automatic detection algorithms are not far off, which will be essential as image volumes increase even more. As Caroline notes:

"*So, I see the work that I've done so far as being very accurate albeit very time consuming. So, what will be really advantageous is to develop some sort of, what I see as eventually a data mining algorithm, …that will be able to extract a feature from an image, ...whatever it is we are interested in extracting, automatically. And the work that myself and my undergrads have done, I think will serve to determine the precision and accuracy of whatever that algorithm is*."

As this quote reflects, the manually analyzed images can serve as a test-bed for the development of automatic systems. But re-purposing these images will require

knowing how the images were processed, making the Excel templates, annotations, and internal methods documents essential metadata.

6.1.1.5 Data processing – synthesis

Data processing practices are thus highly situated to labs, technologies, disciplinary standards, and individual skills. Metadata are interwoven within these practices. Metadata are created, as in the case of the seismic headers and ecology Excel annotations. Metadata are used, as when aquatic biologists apply calibration information to adjust sensor readings. And in some situations metadata are problematic, as when environmental scientists choose not to use machine output files because of their cluttered and difficult formats.

6.1.2 Assessing data quality

Researchers in any field are concerned with data quality, and CENS researchers are certainly no exception. Knowing when data are of low quality is critically important in ensuring the validity and accuracy of scientific results. What counts as "bad" data, however, and the reasons for that characterization, are determined by a number of factors. In this section, I consider "bad" data to be any data that researchers find unusable for any reason. I discuss how researchers determine the question of "what makes data unusable?" and discuss the role of metadata in researchers' practices around "bad" data.

6.1.2.1 Assessing data quality - seismology

"*[W]e have to keep absolutely track of time. We have to know down to the millisecond when these records started and so, um, it's very important that we keep good track of time and that we use the same standard of time*." This quote from a PI of the

CENS seismic project illustrates how critical timing is to seismologic research. Seismic waves travel through different materials at different speeds. For example, some kinds of seismic waves can only move through solid rock, not through any liquid. Researchers thus use the timing of arrival of seismic waves at individual locations as a means to characterize the earth structures through which the waves have traveled.

Because timing is critical to seismic research, timestamps are critical metadata. In the CENS seismic sensor deployment that took place in Mexico, bad timestamps proved to be a significant problem. The seismic stations were configured to update their internal computer clocks based on timing corrections that came from an attached GPS device. The GPS device provided the stations with the correct time, so when the GPS device malfunctioned, the station computer clocks deviated from the correct time. According to a member of the seismic team, about seven percent of the sensor readings from the Mexico project were incorrectly timestamped. Typically, if the timestamp is wrong, the data are considered "bad" and are not used in analysis by the seismologists. A seismology student, for example, who used the data from the CENS Mexico project disregarded the incorrectly timestamped data in his dissertation analysis.

The incorrectly timestamped data are not thrown out, however. In fact, methods development for fixing incorrect timestamps is an active research topic for both seismologists and computer scientists within CENS. One of the PIs on the project described one effort he had made to fix timestamps on data from one seismic station in Peru: "*We had one station which the clock is sort of adrift, okay? And what I do is I take that station and I correlate it to another station where I'm really confident it's working*

right. ... I put the corresponding correction into the [STP] server, so the server now knows that this station is drifting by this amount. ... So, the server knows the metadata now for that, that's sort value-added data we put on."

Once that timing metadata has been added to the STP server, researchers who are downloading that data can ask for the timing corrections to be applied. Note that the timestamps are not changed within the original data. The incorrect timestamps remain in the database. Instead, metadata are created that indicate where a timing error has occurred, as well as any timing corrections that apply to particular data. This timing-correction work might fix data for future uses, but in the meantime conference presentations, journal articles, and dissertations have been written that ignore the data with the bad timestamps.

6.1.2.2 Assessing data quality – environmental science

Multiple individuals in the Environmental Lab use the qPCR machine for DNA detection and quantification. To use the qPCR machine, researchers insert a gridded tray of up to 100 cells, with each cell containing a sample or a duplicate of a sample. With the qPCR machine, however, the amount of DNA in a sample might be on the low edge of detection. The issue is detection vs. quantification. The machine might detect DNA, but it might be in such small quantities that it is near the lower boundary of detection. In these cases, as one student described, "you cannot trust the numbers coming out." They need to set cut-off points for what counts as a positive detection and when they can trust the quantification.

For example, depending on the cut-off point, 30 out of 80 cells in a tray might be good, or 60 of 80 might be good. The machine measures copies of a genome per microliter. According to a member of the lab, the EPA practice is to cut-off at a value of "25," meaning any sample that has a reading of higher than 25 counts as a positive detection. But, the lab member continued, in research settings higher reliability is necessary, so they use higher cut-offs. The qPCR machine that the Environmental Lab owns is highly sensitive, but also a new kind of machine. They have only been used for about three years or so, so they are not clear how to deal with this problem.

Depending on the amount of DNA in a particular sample, researchers might be able to use a sample for detection, but not the quantification. And depending on the cut-off thresholds that a researcher is using, this ability to use a sample for detection and quantification might shift.

6.1.2.3 Assessing data quality – aquatic biology

As noted throughout the vignettes, calibration information is very important in establishing and ensuring trust in data. Documenting calibrations is a critical task. Poor documentation of calibrations can render data unusable. In the CENS aquatic biology project, this issue came up. As Maria, the lead student on the project, described how she has six months of data that she is not using because they are inadequately documented.

According to Maria, the technician who was responsible for maintenance and calibration of the installed sensors did not document calibrations in a regular and rigorous way. Maria said that she has access to the technician's field notebooks, but has trouble looking through them and cannot really understand what the technician was doing:

"*Unfortunately, from when Kimberly left until when Evelyn came that record keeping kind of fall by the wayside and so we don't have a lot of those records. There's like a six month gap in the good calibration data*."

She said that she could spend the time to try to figure it out, but the amount of work that it would take was not worth it to her. So she has a gap in usable data because of technician-to-technician changeover. Because of this, Maria says, she is focusing on this year's data because she knows that she can trust it. Thus, in this case, trust in data is achieved through trust in people, and in their metadata practices.

6.1.2.4 Assessing data quality – soil ecology

In using the CENS automated root imaging system, ecologists identify images through their location in the tube, using X and Y coordinates to identify a particular location. These coordinates are analogous to coordinates that astronomers use to identify a star in the sky. In the CENS imaging system, images are automatically assigned X and Y coordinates by the attached computer based on the internal mechanics of the robotic camera actuator. Caroline, the student leading the image analysis project, mentioned that in the past the imaging system has had problems with consistency. During a few two-week periods, the camera system has been out of alignment, and images have been assigned incorrect X and Y coordinates. Because her analyses are based on extracting time-series of images from particular tube locations, images that have the wrong coordinates are unusable. It is as if an astronomer was measuring the light from a star over time, but for two weeks in the middle of the experiment their telescope was looking at the wrong star.

In Caroline's case, the camera got a bit out of orientation a couple of times, meaning that it was tilted slightly in the wrong direction. So there are certain sets of images from that month that are not consistent with the images collected before and after. Caroline showed me a document that she had created to keep track of these imaging "intervals." It listed time periods in which the images collecting by the system were internally consisted, but inconsistent between the intervals. In this case, she says, "consistency" means consistency of camera orientation.

I asked her if she would be able to tell by looking through their online image database when the images are not consistent in that way. She said no, "*there's no annotation or flag that appears [in the image database]. So someone who is a guest maybe that is using that data would have, may see the change in the field of view and say, you know, 'This is a technical... There's some errors in here.' But someone who maybe is less experienced may not. That's definitely a problem*."

The lead engineer on the project described the same problem with camera orientation, and indicated that it had been fixed, along with a number of other bugs that cropped up during the development process. He also described how they have a database which they use to document equipment fixes. The database includes business as well as technical details, "*we have how much the estimated cost is going to be, what the actual cost is be, what PO it's billed against and then there are very detailed notes about our discussions and why we're doing this.*" Because it includes business information, this equipment fix database is not open to the public or connected to the image database itself.

6.1.2.5 Assessing data quality – synthesis

In every project in my study, researchers make decisions about when to use data and when to not use data. In some cases, as with the seismology timestamps and the missing calibration information for aquatic sensors, bad metadata might be what makes the data themselves be considered unusable. In other cases, such as the ecology root image database, "good" and "bad" data might be indistinguishable if a user only looks at the immediately attached metadata; documentation of what makes data "bad" might be in other forms. In the soil ecology case, researchers have created individual Word documents that describe the time periods in which the soil images are out of alignment with the rest of the database. Similarly, in the aquatic biology case documentation of calibrations, which are key to establishing and ensuring data quality, takes place in field notebooks and individual Excel files. Calibration records are not attached to the data files themselves.

6.1.3 Distributing metadata tasks

In group research, roles and tasks are distributed according to the capabilities and expertise of individual team members. CENS projects are interdisciplinary collaborations by design, and research roles are often split along disciplinary lines. CENS collaborations allow individuals without technical backgrounds to benefit from the use of advanced technology. As one seismology graduate student stated: "*I think it's been great and the people have been great. They've been very helpful. So I suppose I would have been upset or more frustrated if I didn't have such a good group of colleagues because I really don't know anything [about computer science].*" But collaboration brings its own set of

challenges. Researchers must develop shared languages, practices, and tools that meet both individual and team priorities.

Disciplinarity, however, is not the only axis along which research practices are striated. Faculty, students, and staff have different research responsibilities, with different individuals having knowledge of different parts of a project. Additionally, individuals move in and out of a project over time, with the corresponding losses in institutional memory. Metadata practices are distributed along all of these axes.

6.1.3.1 Distributing metadata tasks - seismology

Within the CENS seismic project, metadata practices are distributed disciplinarily, spatially, and temporally. Field notes are shared amongst the team via an email listserv, as noted in Vignette 1. At the beginning of the Peru project, multiple researchers (between two and five) were at the field site at a given time to help with equipment installations. During this time, the field notes were sent out by the individual who was nominally in charge, typically a senior graduate student. After the installations were finished, most on-the-ground work was done by a member of the research staff. Consequently, almost all of the notes relating to field work that have been sent to the email list in the later stages of the project have come from one individual.

From a technical perspective, the metadata being produced by the equipment itself, including headers and network health logs, are the responsibility of the technical staff. A computer science student was responsible for creating the database of network health metadata, and for developing the interface that allowed scientists to see near-real-time updates regarding the status of the seismic stations, as described in Vignette 2.

Metadata knowledge is also distributed along institutional lines. A faculty researcher based at a partner institution described how he has little need to understand how the data move from the field to UCLA: *"Basically, from my perspective, what it does is it all ends up back on a server at UCLA and that's where I enter the picture. So, it magically for me goes from the sensor in the field to [UCLA]. My role is I pick up the data at that point…."*

This comment is slightly facetious given that this same researcher also described how he had visited the research sites earlier in the project, and had been a leader in the project planning process. But his point is well taken, namely, that any questions about data prior to it arriving at UCLA are best directed to other members of the project team. This PI went on to describe how his seismology students are primarily users of the data, and as such do not have knowledge of the data and metadata processes that result in the database, "*Well, you can talk to any of the students but you're going to get the seismology part and they are pretty blissfully ignorant on how the data flows around, in fact I tell them there's more data in there*." I received a similar sentiment from a seismology student who knew enough about the data to be able to use it, but had no interest in knowing all of the details of the equipment and data collection process.

6.1.3.2 Distributing metadata tasks – environmental science

The CENS environmental science project is less of a collaboration between scientists and technical researchers than the other three cases in my study. Collaborations take place within the lab between students. Students help each other out in their

respective field data and sample collection trips, they share samples, and in some cases share notebooks.

Within the Environmental Lab setting, students develop their own data management and documentation practices, as described in Vignettes 3 and 4. The Principal Investigator of the lab told me that she does not actually touch or see the data from most of her projects, except in graph form. The students are the ones who deal with the data from day-to-day. From the PI's perspective, students should use whatever processes, software, and data management methods with which they are comfortable.

Students develop data management practices as part of their development as scholars. Individual metadata practices, such as transcribing notes from notebooks into Excel, develop through being part of a social laboratory setting. As one student described, "W*ith data management, …we talk amongst ourselves just to... Just because we do, so sometimes you'll find what works better for somebody, 'Oh, maybe I'll try that and that will help.'*" Additionally, metadata practices are passed down from faculty mentors. The PI's lab notebooks from her own Ph.D. years are stored in the Environmental Lab as a communal resource. Students described to me how the PI's notebooks provide an example of lab documentation from which they can develop their own practices.

6.1.3.3 Distributing metadata tasks – aquatic biology

The CENS aquatic biology project also has metadata tasks distributed amongst a number of members. As noted in Vignette 5, the documentation of field site and sensor maintenance, as well as calibration activities, is the responsibility of a technician. Technicians have been students and research staff at different times in the project.

Another interesting example of the distribution of data and metadata knowledge relates to an effort to build a centralized database for the project. A computer science student worked with the biologists to develop an automated method for sending data from sensors in the field to a project database using a wireless data transmission system. Similar to the CENS seismic project, this system would allow researchers to visualize data shortly after they were collected, potentially allowing the biologists to see when interesting events were occurring, and allowing technologists to identify any equipment problems faster. An initial version of the system was developed, but shortly thereafter, the lead computer scientist student on the project left for about a year to visit another oceanographic research center. During his absence, the wireless data transmission system and associated database fell into disrepair and were not used by the biologists. Because the new technology was never fully adopted before he left, established practices, such as manually downloading the data, were maintained. As the computer science student stated, his absence interrupted the process of building practices around the new system. "*I think the main problem is I am [gone], so I have been physically been unable to maintain this thing as I would want to. …[W]e weren't able to hit the level of reliability we would want, when [the biologists] can trust the data. So they always went and downloaded data anyway*."

If this new data transmission and storage system is more fully implemented after the computer scientist returns, as was his plan, the documentation of sensors and data might become more dependent on his technical knowledge.

6.1.3.4 Distributing metadata tasks – soil ecology

CENS ecologists rely on technologists to document the automated root image capture system. And, correspondingly, much of the metadata about the images stored in the image database (see list given in Vignette 7) take on a technical bent, with the image database listing parameters such as the image size and camera dwell time. The technicians are also responsible for the image filename syntaxes, which are heavily used by the ecologists to identify and retrieve particular images.

Knowledge of the associated soil sensors, however, is the responsibility of the ecologists. Additionally, ecologists are responsible for documenting of the processing and analysis practices described in Vignettes 7 and 8. Piecing all of these disparate data sets together, and documenting that process, is out of the purview of any one individual. The PI ecologist described his own role in this integration process:

"*Yeah, you know one of the difficulties of every project that is not long term designed for that …has been trying to actually hire a metadata when you really don't have funding for a metadata person. Ideally we would have somebody who would be specifically in charge of the data management, but I'm finding that more and more there is less of the need at the scale of the kinds of systems that I run for data management than there is for all of us to understand a little bit about data management and then it's my job to make sure that all of these things actually do come together and I think that's the way that smaller projects, like this are going to have to work and that's why things like workshops and understanding metadata and data analysis are so critical broadly rather than a group of specific experts."*

In this case, then, the PI has let students develop methods for analyzing and documenting images and sensor data, but sees it as his task to understand broadly how data sets and documentation should integrate.

6.1.3.5 Distributing metadata tasks – synthesis

Knowledge of how data are, and should be, documented is striated in a number of ways within these projects: along disciplinary line, along institutional lines, based on seniority, and based on work roles. Lab and field technicians often take on documentation tasks by default, as they are primarily responsible for maintenance of equipment, and ensuring that data collection proceeds without problems. When a project does not have lab or field technicians, those tasks default to the students.

The distribution of metadata tasks is largely an informal process, with students developing documentation practices and products over time as a matter of necessity. Students do not formally learn data management methods through courses. Students across all four projects said that no such courses exist. Instead they learn their practices through working with more senior students, and by trial and error. The next section focuses on the temporal considerations more directly, outlining how metadata practices evolve over time within a project.

6.1.4 Developing metadata over time

Research projects evolve over time. Students come and go, goals and priorities shift due to monetary or other considerations, and individual researchers build knowledge through lab and field experiences. As part of the same process, metadata processes and products change throughout the course of a project. This section lays out how researchers

in my four CENS case studies develop new metadata practices as part of the development for both practical and professional reasons. Researchers develop new metadata practices as part of new data collection and analysis methods, in response to data quality issues, and for new collaborations.

6.1.4.1 Developing metadata over time – seismology

Section 7.2.1 described how the CENS seismic team experienced significant data loss in their Mexico project due to incorrect timestamps. Data loss also resulted from weather-related equipment damage, malfunctioning sensors, and vandalism. Detecting and diagnosing such problems proved to be such a significant challenge that in the last year of the project, a seismology graduate student lived in Mexico full-time in order to monitor and maintain the sensor stations. As described in Vignette 2, when the seismic team moved their deployment from Mexico to Peru, CENS researchers developed new methods of recording technical metadata to make detecting and diagnosing equipment problems much easier. By collecting and building a display for information about wireless network health, data transmission routes, memory disc space, and errors encountered, members of the team were able to see when data files were missing, when portions of the wireless network were down, and when a memory disc was starting to falter.

As Kyle, the lead field technician noted, having access to this information allowed them to build a more robust and reliable network of stations. *"[In Mexico] I then had to hire someone to spend the [last] year down there keeping it [all] working, …And that was continuous pretty much until the end of the deployment we had somebody on site…*

In Peru, …I now spend an average, for the last 6 months, I spend an average of less than 2 days a month fixing things, troubleshooting things. …You know …in April we visited a site that we haven't opened the box on in over a year. That was great, that was just great. I was frustrated I was having to visit that site, but the point was that it had been a long long time, and that was exactly what we had hoped where we would get to."

Difficult experiences with the Mexico deployment led the CENS seismic team to develop and use a new system of metadata collection and display. Though the primary users of this technical metadata are the students and staff responsible for the maintenance of field sites, the full team benefits from more reliable data collection and transmission systems.

6.1.4.2 Developing metadata over time – environmental science

In the Environmental Lab, students will often share samples if, for example, two people were sampling on the same day in the same location. If one of them is missing a sample or has a bad sample for whatever reason, they will use a sample from the other person. In these situations, multiple people are analyzing the same samples. Samantha, a student, said that they were used Google Docs as a tool for sharing data, both for shared sample data and for data that they shared with undergraduate assistants. But, Samantha described, earlier in the project they ran into situations where they did not have documentation of sample names and numbers, and lost track of what sample names meant. Because of this, Samantha said that she has gone back through the Google Doc data and created a new Excel spreadsheet in which she created a sheet explicitly for

metadata. With that metadata sheet, they document sample names and numbers, and other important information.

Samantha described how they did not collect much documentation in some of the early stages of their projects. As the projects developed, they worked out new methods for documentation, largely motivated by the need to keep track of physical samples. Samantha also described how more senior students in her lab started writing up lab protocols, both for their own use and for training new students and researchers. They now have a student initiated lab web site where many of the protocols are posted and shared.

6.1.4.3 Developing metadata over time – aquatic biology

Vignettes 1, 3, and 5 illustrate the importance of technicians to field-based sensor research. In the CENS aquatic biology project, students and faculty rely on technicians to document field activities. When technicians change, documentation practices also change.

As noted in Section 6.1.3, poor calibration documentation practices led to a six-month gap in usable sensor data. Maria, the main student who uses this data, has been a part of the project since the beginning stages, but as she has transitioned into the dissertation stage, she described how she had an imperative to have more consistent data. To ensure that this data consistency, Maria developed an Excel file in which to document when data were downloaded from sensors and when field calibrations take occurred.

When Evelyn was hired as the lab technician following this period in which data were not documented, she was trained with the new field documentation Excel file in place. She assumed her role as the primary field technician with an understanding of what

was to be documented, and how that information would be shared with the primary users of the data. Thus, since Evelyn started Maria's data is documented better and, as a result, more trustworthy.

6.1.4.4 Developing metadata over time – soil ecology

A main goal of the CENS soil ecology project is to develop effective methods for analyzing the images being collected by the automatic root imaging system. Developing metadata practices are one component of this larger methods development goal. As the processing pipeline solidified, Caroline, the lead researcher on the project, created Excel templates and individual methods documents.

These documents ensure that her methods can be replicated, particularly when she is not there. Near the end of my work with the CENS ecology team, Caroline left the university to work at another university. She continued to work on the CENS project after she left, working with collaborators remotely. When I asked her whether she brought her notebooks and other forms of documentation with her to her new position, she responded. *"I left all of the documentation that I have in my notebooks on my computer. So, any of the metadata and all of the analysis are still on the computer because our undergraduate researchers are still employed… Now that information is in two places."* As the projects continue in different universities and with different researchers, metadata practices might diverge in response to new obstacles and opportunities.

6.1.4.5 Developing metadata over time – synthesis

Researchers develop new forms of metadata for specific reasons, often in response to particular situations. In my case studies, the main motivator for new metadata

products and processes was data loss. The seismic team wanted to make their field deployment more robust by documenting when and where failures occurred in their network. The aquatic biology team needed better documented of field activities and calibrations to ensure that a student would have high quality data for her dissertation. And the environmental science team developed new forms of documentation to prevent mixing up physical samples. The metadata practices of the soil ecology team, on the other hand, developed in a more evolutionary way, rather than in response to particular data losses. This difference makes sense in light of their general goal of methods development. Thus, changing data collection goals can influence the development of new forms of metadata products and processes.

6.2 Data and metadata comparisons

Looking across the four themes discussed in the previous section, researchers face similar issues across projects: needing to manipulate data into usable forms, identify and manage problematic data, work within team settings, and adjust their metadata practices as projects develop. In this section, I provide more direct comparisons of the types of data and metadata researchers collect across my case studies. These comparisons build on the lists of data and metadata provided in the tables at the end of each vignette in Chapter 5, as well as the examples described in the four themes discussed in the previous section.

6.2.1 Data types across case studies

Across my four case studies, CENS researchers collect and use many different kinds of data. In this section, I identify the main types of data I observed researchers collecting and using, and in what format those data existed. First, I look at the range of

data types. Table 6.1 shows the main data types that CENS researchers in my four case studies collect. This table is a synthesis of the data types identified in the vignette data tables in Chapter 5. The data that CENS researchers collect and use can be grouped into five main types: time-series sensor readings, point-sampled sensor readings, physical samples, and images. Time-series sensor readings are continuous streams of readings that the sensor collects at pre-defined and programmed intervals. Point-sampled sensor readings, on the other hand, are sensor readings that are taken opportunistically, not regularly or in a pre-programmed fashion. Physical samples are objects or substances that researchers collect from a field site, and analyze using specialized equipment, either in a remote laboratory setting or in a web-lab set up on site. Images, as the name suggests, are photographs that focus on particular phenomena of interest.

Table 6.1 – Data types collected and used by CENS researchers

Types of data	**Seismology**	**Env. Science**	**Aq. Biology**	**Soil Ecology**
Time-series sensor readings	X		X	X
Point-sampled sensor readings		X	X	
Physical samples		X	X	
Images			X	X

As Table 6.1 shows, the seismology team only collects time-series sensor readings. The environmental science and soil ecology teams each collect two types of data, but without any overlap in data types. The environmental science team collecting point-sampled sensor readings and physical samples, while the soil ecology team collects time-series sensor readings and images. The aquatic biology team, on the other hand, collects all four of these kinds of data.

In addition to this range of data types, CENS researchers' data exists in multiple formats, as illustrated by Table 6.2. By format, I mean the physical or digital structure of data, such as file formats or paper formats. The first data format category is "discipline-specific format." This refers to standardized formats that have been created and adopted by researchers in particular disciplines or data communities. The only examples of discipline-specific formats that I observed researchers using were the SEED and SAC formats for seismic data. No other teams within my study were using discipline-specific data or metadata standards. Researchers in the other three projects used a combination of general purpose simple text file formats, such as .txt, .csv, and .DAT, Excel spreadsheets, and other formats. Members of the environmental science project initially write down measurement values in paper notebooks, to transcribe into Excel later. Members of the aquatic biology and soil ecology team perform analysis using MATLAB, a mathematical software package, and thus have data existing in MATLAB's custom file format. Lastly, the soil ecology team records root lengths and sizes using paper data collection sheets, and then transcribes them into Excel. These paper data collection sheets are a type of data format as well, and one that is important to the work practices of that team.

Table 6.2 – Formats in which CENS researchers collect and use data

Data formats	**Seismology**	**Env. Science**	**Aq. Biology**	**Soil Ecology**
Discipline-specific format	X			
Simple text files (.txt, .csv, .dat)		X	X	X
MATLAB files			X	X
Excel spreadsheets		X	X	X
Paper notebooks		X		
Paper data collection sheets		X		X

Taken together, these two tables illustrate how the seismology team collects one type of data, time-series sensor readings, in discipline-specific formats. The other three projects collect multiple types of data, which then exist in multiple formats. No teams completely overlap in the types of data or the data formats, although all three of the non-seismic teams use simple text formats and Excel spreadsheets in their work.

6.2.2 Metadata types across case studies

Identifying metadata types across my four case studies is more problematic because metadata practices are much more variable from project to project than are the data types. Table 6.3 shows a cross-comparison of each project's use of different documentary forms. To create Table 6.3, I compiled all of the metadata lists generated in the vignette commentaries (Tables 5.1 - 5.9) and added any new entries mentioned in my discussion of the four themes in Section 6.1. I then began grouping like items to abstract out from the specific metadata processes and products within each project to larger categories of documentary forms. As this table shows, most of these documentary forms cross all four projects. For example, researchers on all projects use textual documents to write up research plans and methods. Similarly, researchers on all projects also use Excel spreadsheets to keep track of particular objects or activities, from sensor calibrations, to equipment serial numbers, to physical sample identifiers. The question marks indicate that I have no direct evidence that aquatic biology project creates and uses "Annotations within data files" and "Field site diagrams," but, based on my experiences with the other projects, I expect that I would observe the aquatic team using these documentary forms also. The empty boxes in this table are 1) the seismology team not annotating data files,

which is related to the use of discipline-specific data format that uses a header as the main documentary form, 2) the environmental science team not using data headers, which is due to them recording readings from handheld sensors in field notebooks, not autonomous sensors that generate data files, and 3) the seismology and soil ecology team having computer generated files about the performance of CENS-built equipment, namely the seismic network health logs and the soil image file parameters, while the environmental science and aquatic biology teams have output files regarding the analysis of physical samples that were generated by purchased laboratory machine.

Table 6.3 – Documentary forms across the four case studies

Documentary Form	**Seismology**	**Env. Science**	**Aq. Biology**	**Soil Ecology**
Word Documents	X	X	X	X
Excel spreadsheets	X	X	X	X
Paper notebooks	X	X	X	X
Annotations within data files		X	?	X
Data file headers	X		X	X
Field site diagrams	X	X	?	X
Descriptive filenames	X	X	X	X
Project web pages	X	X	X	X
Computer generated files about equipment performance	X			X
Laboratory machine output files		X	X	

Table 6.3 illustrates two things: 1) metadata practices within each team incorporate a wide variety of documentary forms, and 2) the overlap between projects in documentary forms is considerable. With regard to the first point, researchers use many general use tools that are easily available, such as Microsoft Word and Excel, web pages, and paper notebooks, along with documentary forms that are more specific to their

particular data, such as data file headers and machine generated output files about equipment performance. With regard to the second point, Table 6.3 suggests that abstracting metadata practices to general documentary forms is not a useful way of making distinctions among my case studies.

To try to draw out distinctions among projects in more useful ways, I developed another form of comparison. In particular, I focused on the function of metadata practices to create a categorization of metadata types across projects. To create a metadata typology, I created another compiled list of individual kinds of metadata from the tables given in Chapter 5, and asked the question, "What are particular metadata practices intended to achieve?" I then grouped particular practices into functional categories to generate a distinct typology.

In looking across my case studies, compiling and synthesizing the metadata types identified in the vignettes and in my examination of the four themes analyzed in Section 6.1 (data processing, Assessing data quality, Distributing metadata tasks, and development of metadata over time), I identified six types of metadata:

- *Metadata for data identity:* metadata that establishes or ensures that one set of data can be distinguished from another. Examples: digital filenames and sample labels.
- *Metadata for data collection equipment:* documentation of equipment installation, use, problems, and maintenance. Examples: calibration records, wireless network health logs, annotation of equipment maintenance in field notebooks.

- *Metadata for data characteristics:* metadata that describe characteristics of the data themselves, including where and when they were collected, person who performed the data collection, and what the data are representing. Examples: column headings, units, timestamps, and annotation of seismic events in data headers.
- *Metadata for data quality:* metadata that documents known or potential data quality problems, where "data quality" refers to being able to say that data are what they purport to be, that is, that they accurately represent the desired characteristic of the phenomena of interest. Examples: calibration records, Excel spreadsheets of problems encountered, annotation of deployment issues in field notebooks.
- *Metadata for data collection methods:* documentation of the processes used to collect data. Examples: Word documents of research plans, methods, and lab protocols.
- *Metadata for data analysis methods:* documentation of how data are pre-processed, transformed, combined, and adjusted during the analysis process. Examples: Notes in Excel data files of analysis steps, lab protocols that outline sample preparation and analysis processes.

In this typology, the first two categories include metadata that document particular things or objects. The data identity category includes documentation of particular data files or sets, and the data collection equipment category includes documentation of machines. The second two categories, metadata for data characteristics

and data quality, include documentation of properties of those things or objects, including what a data set is purported to measure, and whether the data collection machine was working correctly. The last two categories, metadata for data collection and data analysis methods, include documentation of processes in which those things (data sets or collection equipment) are created and used.

Note that particular metadata processes or products may serve as multiple metadata types. For example, a data file name serves as identity metadata, but the particular file name of interest (such as an aquatic biology sensor data filename or an soil ecology image filename) might also indicate particular data characteristics, such as location and time of data collection. Similarly, researchers' lab and field notebooks may contain metadata of multiple types, including metadata related to data quality, data collection equipment and methods.

These metadata types cross my four case studies, in the sense that each team has developed practices that involve the creation and use of metadata for these purposes, but this functional view of metadata is more illuminating in seeing differences across projects. In the next set of tables, I illustrate how the forms of metadata created and used in each project fit into this typology. The first column in each table is a list of the particular forms of metadata identified in my discussion thus far for each project. The other columns indicate the metadata type categories in which each particular form of metadata in the list falls.

Tables 6.4-6.7 show the range of metadata processes and products within each case study. Comparing these tables illustrates a number of interesting features of

metadata practices across projects. I focus here on four features of this cross-project comparison:

- Data identity – The metadata practices of the seismology around establishing and ensuring data identity are much more streamlined than the other three projects. In the seismology project, with only one main kind of data - ground motion acceleration - establishing data identity only needs to take place when the data are initially collected. Because the other projects have more kinds of data - image data, physical samples, and sensor data - they have more data resources to identify and thus need to keep track of those identities across more formats. The "Data identity" columns illustrate this by showing fewer checked boxes in the seismology table (Table 6.4).
- Timestamps – As noted in the vignettes, in seismological research timestamps are central to data quality. As such, the seismic data are time stamped with high precision. As one faculty seismologist stated, "We have to know down to the millisecond when these records started." In fact, data filenames include timestamps down to the second, such as "20110430002151," where the last six digits of that number are hour, minute, and second. In the soil ecology and aquatic biology projects, sensor readings are also timestamps, but the degree of precision is less critical. The soil ecology timestamps are in five minute intervals, for example 9:00, 9:05, 9:10, etc., while the aquatic biology timestamps increment by 30 minutes each, such as 9:30:00, 10:00:00, etc. On both of these projects, the data filenames include the date of data collection, but not the time. In the

environmental science project, data are recorded and identified by dates, but the times when sensor readings and physical samples are collected are not written down in notebooks (Note that the notebook page shown in Figure 5.6 does not indicate time). Researchers in the environmental science project, however, are often aware of time at a finer resolution than just the day, and will often try to collect data at roughly the same time every day, such as 10:00 AM, in order to mitigate any effects of time on a project's outcome.

- Location – Similar to the timestamp issue, precise location metadata is important for seismic data, as the discussion of GPS coordinates in Vignette 2 described. In the other projects, precise GPS coordinates are not as important. In fact, the term "gps" does not appear in my interview transcripts and field notes from the non-seismology projects. Location is important in determine data identity, that is, knowing that data have been collected at the same location is important in order to ensure data comparability over time. But with the aquatic biology, soil ecology, or environmental science projects, if data were collected in a different place, the research would not change. For example, during one trip to an aquatic biology field site, the researchers were moving a set of sensors across the harbor to another, because their previous sensor location was undergoing construction. Their criteria the new sensor location was that it was roughly at the same distance inland from the harbor outlet. Most critical for the non-seismology projects with regards to location is knowing what the surrounding area is like, for example that the data were collected in a conifer forest, near storm drain, or in an urban harbor.

- Across all four tables, only a few items have checks for all six metadata types. In the environmental science project, lab notebooks and lab protocols both include all six metadata types. Similarly, within the soil ecology project, researchers have created Word documents that describe the image analysis process, sensor installations, and sensor data files in ways that use all six metadata types. With the exception of the environmental science lab notebooks, the other three metadata forms that included all six metadata types are documents explicitly created to document research processes for other team members. I did not observe any individual forms of metadata within the seismology and aquatic biology projects that were quite as wide in scope.

Table 6.4 – Metadata form and type matrix for the CENS seismology project

Form of metadata	**Data identity**	**Data char.**	**Data quality**	**Data collection equip.**	**Data collection methods**	**Data analysis methods**
Seismic station site numbers	X					
Seismic data filenames	X	X				
Timestamps	X	X	X			
Equipment serial numbers	X			X		
GPS readings		X	X			X
SEED header components		X		X		
SAC header component		X				X
Sparklines graphs			X	X		
Timestamp corrections			X	X		X
Names of file directories and scripts		X	X	X		X
Notes of field work			X	X	X	
Logs of equipment installation locations				X	X	
Pictures of the installations				X	X	
Diagrams of field sites				X	X	
Network health logs	X			X		
Sensor response specifications				X		X
Seismic project wiki				X		

Table 6.5 – Metadata form and type matrix for the CENS environmental science lab

Forms of metadata	Data identity	Data char.	Data quality	Data collection equip.	Data collection methods	Data analysis methods
Bottle labels	X	X				
Sampling site names	X	X				
Names of samples written on test tube label	X	X				X
Sample names input into the computer	X	X				X
Notes in lab notebook	X	X	X	X	X	X
Notes from lab notebook transcribed into Excel		X	X			
Lab "protocol"	X	X	X	X	X	X
Machine output files	X		X	X		X
Packing list				X		
Measurement units written in field notebook		X				
Field sampling plans				X	X	
Graph of the calibration curve			X	X		
Lab wiki				X	X	X

Table 6.6 – Metadata form and type matrix for the CENS aquatic biology project

Forms of metadata	Data identity	Data char.	Data quality	Data collection equip.	Data collection methods	Data analysis methods
Sensor data filenames	X	X				
Sensor data file header	X	X		X		
Data file folder labels	X	X		X		
Physical sample labels	X	X				
Label written on the sensor	X			X		
Notes in technician's field notebooks	X	X	X	X	X	
Timestamps		X				
Calibration values written in field notebook			X	X		
Excel file of calibrations and field activities			X	X	X	
Sensor manual from manufacturer		X	X	X	X	
Laboratory reference books						X
Field work planning document			X	X	X	
Laboratory sample preparation protocols			X	X	X	X
List of physical samples held by laboratory		X				

Table 6.7 – Metadata form and type matrix for the CENS soil ecology project

Forms of metadata	Data identity	Data char.	Data quality	Data collection equip.	Data collection methods	Data analysis methods
Image scan parameters	X	X		X	X	
Image filenames	X	X				
Word document describes the image filename syntax	X	X		X		
Word document that describes image analysis process	X	X	X	X	X	X
Folder filenames for image time-series'	X	X				
Paper data collection sheets	X	X				
Sensor data file header	X	X				
Notes in field notebooks	X		X	X	X	
Word document describing sensor field installations and data files	X	X	X	X	X	X
Timestamps		X				
Database of equipment fixes			X	X		
Template Excel sheets for analysis process		X				X
Notes in Excel data sheets			X			X
Excel spreadsheet with sensor calibrations			X	X	X	

6.3 Researchers understandings of metadata

In my initial research questions, I hoped to uncover researchers' own understanding of "metadata." What does the term "metadata" itself mean to scientists in distributed projects, and what other language do researchers use in reference to data description activities? In this section, I focus on how CENS researchers understand "metadata" as a concept. In my interviews, I asked researchers what the term "metadata" meant to them in their work. I did not receive answers to this question in all interviews, but researchers' familiarity with the term "metadata" varied from person-to-person. Only one interviewee, a seismology student, responded with the literal "data about data": "*The metadata is basically it's data about data... It just has information about the [seismic] wave form that you are looking at. The document that comes with data itself."*

Most responses to my question about the meaning of "metadata" focused on the function and role of metadata in the research process. An environmental science student described metadata as "information about how data were collected and information needed to know about how to use the data." Researchers would often provide very specific examples of what metadata meant in their work, as illustrated by this ecology faculty member, "*Metadata, to me, is the description of the data in that particular column or that particular unit. So it's what that number that in the column represents and where does it come from, how is it developed so that, basically, what you have is a pile of data and the metadata tells us how it's... Or what it's from and where and what are the assumptions that are inherent in that*." Similarly, a staff researcher on the ecology project described metadata as "a header," showing me a file that listed variable names and units,

and followed up by enumerating other specific kinds of metadata, including lists of sensors, calibrations, and sensor replacements, as well as "where things are located."

I received this kind of response, a long list of examples, from a number of researchers. A computer science student working on the aquatic biology project responded that his metadata consisted of "*information about the deployment, the sensor calibration and who deployed the equipment. Basically, everything about the deployment and the sensor, that's all*." He then proceeded to list what "everything" was to him: "*So that if I have to use the data, I know that all the sensors were calibrated and I have those calibration values and know when the data start. So I mean, so let's see, for me, the metadata would be revolving around the sensor. Because that's where the data is being gathered and I want to know when it was deployed, who deployed so that there's a point of contact, what are the calibration values for the different sensors and, yeah, and that's how I'll define metadata. The calibration, deployment information. Oh, yeah, by the way I forgot, and service information but I guess that it's more at the level of the manufacturer because if it went for servicing and came back. That usually changes the calibration parameters. So you might want to log that … ."* I received similar a list from the field engineer for the seismology team.

Other researchers interpreted "metadata" in unexpected ways. A faculty biologist interpreted "metadata" in two ways, both of which were different from any other interviewee. His first interpretation, which I also noted in Section 6.1.3 on data processing in the aquatic biology project, was of "metadata" as a pre-cursor to "real data," essentially equivalent in meaning to "raw data":

"*When you get the raw data in and it has to be considered just that, raw. It's not processed data. So …[we have a] very specific set of algorithms the [we] put the data through to make sure that it's quote-unquote massaged, so that any bad information is taken out. And we always try to err on the side of being conservative, meaning I'd rather throw data away than be concerned that I may have data that I can't trust. So we go through that process which, for us, converts it from quote-unquote metadata into real data. But that's kind of the massaging of the data processing that we do*."

His second interpretation of "metadata" was analogous to "meta-analysis," that is, an analysis of data from multiple independent projects. In this interpretation, "metadata" are the data that are combined to be used for meta-analysis: "*There are other scientists who love to collect and look over datasets. So, this is why I asked you to define metadata because to them they do a meta-analysis. They will take raw data or raw processed data, good data from various places or maybe from different studies in the same place and they will look at just the information. They haven't collected any of it, but they may know how to manipulate it and get new insights out of it that somebody has not yet done*."

The other unexpected interpretation of "metadata" was from an environmental science student. When asked about what "metadata" meant in her project, she responded in a manner that indicated she had only heard the term in relation to computerized geographical information systems (GIS). "*We never use [the term 'metadata'] in the lab. I think there are couple of us have taken GIS classes or whatever that have talked to them about metadata but we never do it our own*." She then went on to describe how her lab notebook is her primary means of note-taking and annotation, indicating that while she

had some familiarity with the term "metadata," it was not part of her day-to-day environment.

Thus, while researchers discuss "metadata" in different ways, their conceptions do not significantly deviate from Greenberg's (2005) statement that the term "metadata" "addresses data attributes that describe, provide context, indicate the quality, or document other object (or data) characteristics" (pg. 20). Researchers also used other terms to refer to related types of activities, including "recordkeeping" and "documenting," but the use of other terms was sporadic outside of the seismic team. In interview passages, a faculty member of the seismic team used "auxiliary information" and "value-added data" along with "metadata" to refer to annotations of station locations and timing corrections in their database. Additionally, researchers on the seismic team use a number of permutations on the terms "system health parameters," "logging data," and "network health metadata" to refer to their automatically generated metadata about the wireless network and sensor functionalities.

6.4 Reporting out – Data and metadata in published articles

For CENS researchers, reporting on research is a common activity. Research reporting primarily taking the form of published articles in journals and conference proceedings. In this section, I outline how the data and metadata practices discussed thus far were presented in CENS research publications. Note that I do not provide citations to the papers that I discuss here or use lengthy quotes, because these would quickly identify the individuals who have participated in my study. Instead I discuss them in general terms to identify patterns and trends in discursive forms.

In my four case studies, published papers reported on scientific results and on technology development. The first feature of note in these papers is that discussions of "metadata" themselves were essentially non-existent in the papers I examined, with one exception, namely a paper from the seismology project that was focused on the network health logs and visualizations. In searching for the term "metadata" (or "meta-data") across papers from all projects, I got three results. The seismology paper on the network health logs used "metadata" eight times, one other seismology paper referred to "metadata" once (in reference to earlier work on these network health logs), and one soil ecology paper used "metadata" twice, once in reference to "location metadata" for sensors and another time in reference to "metadata tags" of EML (Environmental Metadata Language), though they were not using EML themselves.

In looking these published papers in relation to the typology of metadata that I developed earlier in Section 6.2, papers from all projects were very detailed in giving information that would fit into the *data characteristics* and *data analysis methods* metadata categories. For example, soil ecology papers are very detailed about the variables collected and the ways that those variables are combined, filtered, and modeled to produce results. Similarly, the seismology papers provide detail about what data were used, such as the size of earthquake events analyzed or the time period in which earthquake events were analyzed, and focus extensively on the processes used to interpret and model those data.

In contrast, descriptions of work that would fit into the *data quality*, *data collection equipment*, and *data collection methods* categories in my metadata typology

are much less consistent. There is a significant difference in the amount of description of field activities between the papers from the seismology project and the papers from the other three projects. The seismology project has produced multiple papers that present scientific results, and multiple papers about the deployment itself, specifically about the wireless network and the network health logging and display tools. The deployment papers describe the field setups, data collection, storage, and transmission processes in great detail. On the other hand, the papers that present seismological results do not describe the field data collection locations or processes in detail. Of the three such papers that I examined, the first mentions that the project utilized a "100-station broadband array," and provided a map of the locations where the sensors were installed. The second paper gives more detail, giving a five-sentence description of the project background and process, including listing the manufacturer of the sensor and giving the sampling rate. A third seismology article that used data from the CENS project along with data from other projects mentioned that a portion of their data came from the CENS project and cited the first article I mentioned above. Two of these articles include supplementary materials (called auxiliary materials in one case) that can be found on the publishers' web sites, but neither of these gives a characterization of field work or data collection methods, instead focusing on data analysis and mathematical operations.

In comparison to the seismology science articles, articles from the environmental science, aquatic biology, and soil ecology projects, whether reporting on scientific results or technology development, give extensive characterizations of field settings, equipment, sampling methods, and sensor details. In the case of the soil ecology project, for example,

papers that focus on the scientific results have sections specifically devoted to "sensor technologies," "imaging systems," "study site," and "data collection." Similar sections exist in papers from the other two projects. These descriptions are written up explicitly for the published paper. I did not observe any similar narrative descriptions of field sites or data collection processes in the day-to-day metadata practices of these researchers.

I argue that this difference between the seismology project and the other three projects with regards to their description of *data quality*, *data collection equipment*, and *data collection methods* work reflects back to the differences in the types of data and metadata being collected and the degree of standardization of those data and metadata types. In response to my interview question, "What data are being collected on this seismic deployment?," multiple people on the seismology team responded by saying "seismic data." The understanding of what "seismic data" is very uniform within the seismic research community. In contrast, the environmental science, aquatic biology, and soil ecology research communities are much less uniform in the kinds of data that are collected, and the methods used to collect data. Thus, describing the particular kinds of data being collected, field sites, and data collection methods is very important. Zimmerman (2007) illustrates how data re-use is strongly tied to whether data users can project their own experience in doing field data collection onto someone else's data presentation.

Another notable feature in the published papers was the use of acknowledgement sections to call out team members who performed data collection and other research work. These acknowledgements in some cases referred to particular people, and in other

cases referred to "many volunteers" who helped with lab or field work. Of the 33 articles I examined, eleven acknowledged team members' efforts in helping with lab or field work. Seven papers from the soil ecology project included such acknowledgements, as well as two papers each from the seismology and environmental science projects. No papers that I examined from the aquatic biology project included such acknowledgements. It is possible that these differences in acknowledgement behavior is related to the number of authors listed per paper for each project. Table 8.1 shows a breakdown of the number of authors per paper for each project and the number of papers with acknowledgements for field work. The papers from the soil ecology project have the smallest number of authors per paper for these four projects, while the aquatic biology and environmental science project have the highest number of authors listed per paper. If I exclude the environmental science project on account of the smaller number of papers examined, these numbers suggest that higher numbers of authors per paper correlates with lower likelihood that the paper will give an acknowledgement for field work. Because of the limited sample, this finding is tentative, and points to future work.

Table 6.8 - Comparison of the number of authors per examined published paper with the number of papers that acknowledge lab or field work performed by a non-author.

CENS Project	Number of papers examined	Number of authors per paper	Number of papers acknowledging field work
Aquatic biology	10	9.3	0
Seismology	8	5.6	2
Soil Ecology	10	4.1	7
Environmental Science	5	9.6	2

Summing up this section, published papers vary across projects in the degree to which they include information that maps to my metadata typology. The *data characteristics* and *data analysis methods* categories from my metadata typology are generally well covered, while the *data quality*, *data collection equipment*, and *data collection methods* categories are covered with more variation. In addition, CENS papers often thank people who conducted lab and field work in the acknowledgements sections of papers. The importance of these acknowledgements is that they indicate that additional people were involved in the project beyond the paper authors. As my ethnographic study shows, lab and field workers are involved in day-to-day metadata tasks, even if they are not involved in writing up publications.

6.5 Summary of case study comparisons

In this chapter, I compared data and metadata across my four case studies. Researchers in each project derive analyzable data sets through processing methods, they decide how and why to work around "bad" data and metadata, they create metadata within team projects, and they change metadata practices over time as a project develops and obstacles arise. The types of data collected in each project have an impact on the kinds of metadata that researchers create. The seismology project collects one kind of data, time-series sensor data, and is the only group within my study to use discipline-specific data and metadata standards. The environmental science, aquatic biology, and soil ecology projects each collect multiple types of data, and largely use generic data formats such as simple text formats and Excel.

Looking across the metadata practices of CENS researchers, I identify six types of metadata, specifically, metadata that document: data identity, data characteristics, data quality, data collection equipment, data collection methods, and data analysis methods. The ways that researchers create metadata types varies from project to project. In particular, the seismic project has a need for much higher precision in the time and location metadata than the other three projects, as reflected by their use of high precision timestamps in their sensor data filenames and their use of GPS coordinates.

I then discussed researchers' own conceptions of the term "metadata," showing how researchers viewed "metadata" in functional terms as lists of particular documents or types of annotation. The kinds of metadata that are then included in publications largely consisted of *data characteristics* and *data analysis methods* metadata, with the non-seismology teams also providing descriptions that are similar to metadata relating to *data quality*, *data collection equipment*, and *data collection methods*.

In the next section, I turn from my ethnographic study to the user test of the CENS Metadata Registry. In doing so, I shift my focus from the kinds of metadata that CENS researchers create and use in their everyday work to metadata that are intended to be used by someone outside the immediate CENS projects.

7. RESULTS III – CENS METADATA REGISTRY

Chapters 5 and 6 provide a baseline that illustrates typical metadata practices and how they fit into broader research activities. I now shift to a description of CENS researchers' use of the CENS metadata registry prototype. I illustrate how CENS researchers approached the task of describing their data using a standardized schema. This discussion includes an illustration of the kinds of descriptions that people gave for their data, what kinds of problems people encountered, and what resources they draw on to describe their data. In the sections below, I provide examples of responses that testers submitted in our form, and quotes from the tests and post-test interviews. When I give an example, I identify the testers with a number and the project in which they are involved, with form response in bold and quotes in italics, for example:

- Form Response: **"Seismic data: Mexico."** (Tester 2 – Seismology)
- Quote: "*So, I used the Mexico dataset. There's not a specific subset. I actually used the entire thing*." (Tester 2 – Seismology)

I assigned the test numbers according to the chronological order in which we conducted the tests. For the testers that are part of projects other than my four case studies, I give their project as "other CENS." These five individuals included three computer scientists, one electrical engineer, and one environmental engineer. In passages in which I quote dialogue between a tester and the test administrator(s), I do not distinguish which test

administrator is speaking.

In the next section, I describe characteristics of testers approach to our test and the responses that they give. In the following section I then go over how the results of this test relate to the four themes of metadata creation that I developed in Chapter 6: data processing, metadata relating to data quality, the distribution of metadata knowledge, and the metadata issues relating to project timelines.

7.1 Use of a community metadata registry

In this section, I give a brief overview of some test statistics. Following that, I present some important issues that arose during the tests and post-test interviews, specifically: 1) Sense Making: what challenges testers faced in making sense of the metadata creation task, 2) Talking vs. writing: how testers' verbal descriptions differed from what they ended up writing in the form, 3) Projected/Reverse sense making: how testers projected that their metadata descriptions might be used, and 4) Use of existing documents: how testers utilized existing forms of documentation in the tests.

7.1.1 Overview of test responses

Counting the post-test interviews as part of the test, the tests took an average of 34 minutes, with the longest taking 57 minutes and the shortest taking 21 minutes. Testers varied widely in the length of their responses. Table 7.1 shows some statistics about the number of words that testers used in their responses to each field. Note that I counted dates in the form of 2011-01-01 and URLs as one word.

Table 7.1 - Response word counts per metadata field (n = 11 testers)

Data description fields	Average	Max	Min	Median
Data set title	5.7	16	1	3
Dates of data collection	2.5	5	1	2
Data collection site	4.1	9	0	2
Other contributors	5.0	17	0	6
Data type (chosen from list)	1.5	3	1	1
Research question/why the data were collected	24.9	59	2	15
Variables collected	17.9	60	1	10
Process and equipment used for collection	17.3	34	0	17
Data format	2.8	12	1	2
Data sharing permission level (chosen from list)	1.2	3	0	1
Data sharing permission level (free text)	4.3	19	0	0
Funding source	9.6	58	1	2
Keywords	8.6	16	4	8
Location of the data (URL)	0.7	2	0	1

The results shown in the "Average" column of Table 7.1 are not surprising. On average, testers provided the longest responses to the three data description fields: "Research question," "Variables collected," and "Process and equipment used for collection," followed by the "Funding Source" and "Keywords fields." We anticipated that these fields would elicit longer responses, and in fact provided larger text boxes in our form for these five fields than for the other fields. It is possible, and probably likely, that the larger text boxes contributed to the longer responses on average for those fields, but we did not test this.

Looking at the average column in comparison to the maximum, minimum, and median columns, however, is more interesting. In particular, the fields with large relative differentials between the maximum and median columns are notable. Four fields have a maximum that is at least six times greater than the median value: Variables Collected, Data Format, the free text option for Data Sharing Permission Level, and Funding

Source. This cutoff of six times greater is an arbitrary one, but differentials this large indicate wildly varying behavior on the part of the testers. In the Variables collected field, two responses were much longer than others at 60 and 56 words respectively. The 56 word response was from a tester who copied and pasted a list of column heading and units from a sensor data header file. The 60 word response was by a tester who did not have a straight-forward set of columnar data, thus making his task of describing his "variables" more nuanced. I discuss this case more in Section 7.1.2 on sense making.

Similarly, in the Data Format field, most testers gave used between one and three words, usually just a format name, such as SAC, JPG, netCDF, or as one tester responded, "**text file**" (Tester 3 - other CENS). The longest response came from an electrical engineer student on the aquatic biology project who was used to providing data to science and engineering partners in a multitude of formats. His response was, "**Stored internally in a Database. Can be provided as text files/netCDF etc.**" (Tester 5 - Aquatic).

For the Data Sharing Permission Level field, we asked testers to choose permission levels from the Creative Commons open source licenses, or they could state their own data sharing permission criteria using a free-text box. Five of the eleven testers used the free-text option to state their own sharing criteria. These five responses were:

- "**see our web site**" (Tester 11 - Ecology)
- "**Database is not directly accessible. We can provide the data in almost any desired format on a case-by-case basis**." (Tester 5 - Aquatic)
- "**Attribution, with discussion how data will be used**." (Tester 8 - Aquatic)

- "**Need to get permission from the PI of the Project**" (Tester 1 - other CENS)
- "**check with the data collector website**" (Tester 9 - other CENS)

Further investigation is needed to determine whether these testers have a greater awareness of data sharing permission issues, or were simply confused by the Creative Commons license options.

The Funding Source field also saw widely varying response sizes. The median number of words used in the Funding Source field was two words, as most testers responded with one or two organization names, such as "NSF" or "CENS." The maximum, however, was 58 words used. Three testers used more than ten words with 58 being large outlier. The second and third most words used were 17 and 13. In all three of these cases, testers copied and pasted funding information from the acknowledgements section of a personal publication, such as "**Supported by the NSF through the Center for Embedded Networked Sensing at UCLA**" (Tester 1 - other CENS).

Another interesting feature of Table 7.1 is the column that shows the minimum number of words used for each field. What immediately jumps out from the "Min" column is the number of zeros. In filling out the metadata form, testers would often skip over fields that they either did not understand or did not want to fill out. Testers would occasionally go back to a field that they skipped initially, after looking more closely at the rest of the form. Testers would also occasionally give token answers, for example only responding to the Variables Collected field with "**Soil temperature**" (Tester 11 - Ecology), instead of responding with a full variable list.

Notably, the largest minimum word count value was four for the "Keywords" field. Testers were very consistent about responding with keywords, with the average and median word counts for the Keyword field being both being right around eight. Researchers might skip or give token responses to other fields, but they will stop and type out eight keywords. Note that I counted the words individually, meaning that a keyword of "**stationary camera**" (Tester 4 - other CENS) counts as two words, but the total word count is still a good indicator of the efficacy of that field, because "stationary camera" tells a potential data user as much, if not more, about data than "stationary" and "camera" would individually. These consistent responses to the Keyword field indicate that our CENS testers were familiar with the task of writing out keywords in a way that they are not as familiar with the other fields. Researchers regularly have to create keywords for publications, conference presentations, among other reasons, so the task of making sense of the field is easier than for other fields of which the testers had less familiarity. In the next section, I discuss more broadly the sense making challenges our testers had in creating metadata using our form.

7.1.2 Sense making

"Sense making" is a process of comprehending a given situation by using knowledge, intuitions, opinions, personal experiences, effective responses, evaluations, and questions (Dervin, 1992; Savolainen, 1993). Sense making involves determining the "plausibility, coherence, and reasonableness" of a particular situation or task (Weick, 1995, pg. 61). The first challenge the testers faced was simply understanding what it was that they were being asked to do. None of the testers were familiar with the Dublin Core

Metadata Schema. In a few cases we had to clarify that we were not collecting the data themselves, we were only collecting descriptions of data. We also had testers respond to our test introductions by making statements like, “I may not be a good person for your test because my data is not like typical CENS data.” In these cases we assured the tester that nobody has “typical CENS data,” and that we wanted testers with as many different kinds of data as possible.

Beyond those basic mis-understandings, other confusions resulted from the ambiguity of “data.” A number of testers struggled to determine what data we were wanting them to describe. As one tester noted, his methods were highly computational, without a clear data set to which he could point to as his “final dataset.”

“*I have trouble thinking of what the actual final dataset actually is, like numbers in the file. Because there’s lots of… the final result is like a little time series graph and it's just so much data reduced to just that. So, I wouldn't call these time series graphs a data set, because it's something that just changes with any parameter, sort of kind of on the fly based on however you want to interpret the data*.” (Tester 2 - Seismology)

After some discussion, he decided instead to describe the data that he started with prior to any of his computations.

Another example of the same sense making challenge is that more than half of our testers reported being part of multiple projects. Testers described how their data consisted of many constitutive pieces, such as multiple files and database tables, and they may be spread around multiple locations, such as lab servers and personal computers, even

located in multiple institutions. In creating metadata descriptions for our test, researchers had to decide what data they would categorize as being under a single project.

"*We have two separate things but if you come to us, we could supply both of them, in a sense that my database currently only stores the [communication] data and the vehicle data. But, I mean, I also have access to the other dataset which is what the vehicle has been collecting but it's not being stored in the same database. It is stored [by the] people who collaborate at the biology lab*." (Tester 5 - Aquatic)

This multiplicity of data sets emerged for Tester 5 as he proceeded through the form. He initially wrote a very descriptive data set title, using about 10 words. Later in the test, however, as he realized that he was describing multiple data sets, he shortened the title considerably. As he discussed with us, and the above quote illustrates, he had access to multiple data sets that might be of interest to secondary users. But, he noted, he was only the primary user of one particular kind of data. In the end, he decided that he would prefer to only be contacted by potential users of his primary data. His final submitted title was two words long: "**Glider data**."

The challenge of sense making extended to the metadata fields themselves. Metadata fields may not make sense to a researcher who has not seen them before. In our tests, the researchers routinely asked for clarification of what they were expected to include in particular fields, requesting examples or further explanation. Although, as noted in my method section (Section 4.2), we tried to customize the field terminology to be more understandable to CENS researchers, terminology confusions were still common. Researchers approach the task of describing their data with the terminology of

the communities in which they are embedded. This was most visibly illustrated by testers' confusions about the list of options they were presented with for the "Data type" field. A number of testers were confused by the options, with their confusions stemming from their own ways of referring to their data. Testers preferred to describe their data types in terms that were not on our list, such "geo-located time series," "sensor streams," or, as in the following quote, as "acceleration time series": "*I don't know what it means, the "type of data." Like, I saw you put image, sound, text, but what would you call my data? Acceleration? …I'm not used to using this word ['numerical']*."

When he was then asked if he knew of a better word to describe his data, he responded: "*No, maybe numerical. If someone asks what is the type of data, time series, I think people would say, like time series, acceleration time series. But this is numbers. So it is numerical. Maybe I'm not used to this word. … I think time series. Because I think the type also should reflect, I want to say the field, but it also reflects the type of methods that people use. So if you're working with time series, you know the references; the books, for example, to looking for, you know. At least for me. And I'm someone who is doing algorithms for example*." (Tester 3 - other CENS)

Researchers were also confused by other fields on a case-by-case basis. The three data description fields - Research Questions, Variables Collected, and Process and Equipment used for Collection - generated varying responses. One tester indicated that the "Variable" field did not make much sense to him, as illustrated in the following exchange, with "Q:" being the investigator who is administering the test.

Tester 2 - Seismology: *'Variables collected.' [pause] I don't know what this means. Maybe it means being more specific about the data, like list out accelerometer time series data, wind velocity and direction data and ice and bathymetry data. Or… That's what it must be, since nothing else here asks for that.*

Q – Test Administrator: *Are you saying there's something else that it could have been or that you were thinking it might be?*

Tester 2: *Something that affects how I process the data. Like, I mean, for me, variables are things in a model. So the data isn't really a variable to me. It's the little parameters that I adjust to come that interprets the data differently.*

Tester 2 then proceeded to include a processing-related description in his response to the "Variable" field, by indicating that he used the data at a different sampling rate than the readings were originally taken: "**Seismic data: time series of 3 channel accelerometer data at 100Hz. I used it at 1Hz and 10Hz."** His response to the "Process and Equipment used for collection" field, in turn, gave a list of URLs for the web sites from which he downloaded data.

7.1.3 Talking vs. writing

The last example of Tester 2 and the "Variable" field illustrates a typical conversation between the testers and the test administrators. Testers would often ask questions to get clarifications about a particular field. In these situations, we tried to "talk through" a field until the tester decided what to write. These verbal discussions about what should or should not go in a given field were not always reflected in what testers

ended up writing in the actual metadata fields. Often a rich verbal discussion resulted in a brief written statement.

The “Data sharing permission level” field proved often to elicit lengthy discussions, for a couple of reasons. In this field, testers were asked to choose permission levels from a list of Creative Commons licenses, or provide their own data sharing policy in a free-text field. First, testers often needed clarification on the difference between the Creative Commons licenses, and second, most testers did not have an established data sharing policy within their research teams. The following exchange illustrates this confusion:

> ***Tester 8 - Aquatic:*** *I mean, I'm not entirely sure for the sharing permission level. I don't feel like that’s something that we could decide or plan as like a group. So I wouldn't really know what to put.*
>
> ***Q-Test Administrator:*** *So if somebody was just to ask you, "Would you be willing to share your data?" like just how would want to say it? Just kind of person-person, how would you describe it?*
>
> ***Tester 8:*** *Person-to person. I mean, yes. Obviously, I think there would need to be attribution and I think... This is always I think the issue that people run into, is like how is that going to be used, right?*
>
> ***Q-Test Administrator:*** *You could put something in the other data permissions, you know, that you discussed with the PI about use.*

After this exchange, the tester wrote “**Attribution, with discussion how data will be used**” in the “Data sharing permission level” field. Later, however, in the post-test

interview, Tester 8 brought up this field again. The following comment illustrates how her view is more nuanced than her brief written response might indicate.

"*I mean, I think maybe it's hard for the whole attribution thing. You might work something... Basically saying, we're actively working up this data, or you know publications are out or on the way out, or... You know what I mean, so you get a better idea of, 'Yes, we have this data. Yes, we would be happy to talk to you about it, but if you can wait six months…'.*"

With another tester that had trouble with the Data Sharing Permission Level field, we had a four minute conversation about the various license options and his own thoughts about data sharing permissions. This conversation resulted in him choosing to apply two Creative Commons license, with the additional data sharing permission condition of "**see our web site**" (Tester 11 – ecology), even though his web site did not have a data sharing permission statement.

7.1.4 Projected/reverse sense making

A stated goal of the CENS metadata registry is to make CENS data more discoverable to outside users. We thus asked testers to create descriptions with the thought that they might be used by someone outside of their immediate projects. The potential users and uses of research data, however, are often not obvious, even to the researchers who collected them. Researchers must try to project what metadata descriptions potential future users will need to make sense of their data.

As a post-test question, we asked testers to identify which metadata fields included in our test were most useful for describing their data. Answers to this question

spanned our field set, with the "Variables" field being the most common response. Table 7.2 shows the full list of responses.

Table 7.2 – Responses to the question, "Which fields were most useful for describing your data?" (n = 11 testers)

Field	No.
Variables	6
Data collection location	3
Process and equipment used for collection:	3
Research question (why collected)	2
Data type	2
Keywords	1
Dates of data collection	1

Similarly, we received a wide range of answers to the question, "Is there anything that you think is missing from the form that would help you describe your data better?" Responses to this question tended to be very specific to the data being described.

"*The other part that would have been useful for image people are like, they want to know or what are the nuisances of the images, like over the course of... It's collected over the course of the day. So, I have lighting issues and there's also the wind. There's a lot of moving things in the background, those kind of things could help people decide whether or not they want to use the dataset*." (Tester 4 - other CENS)

"*So one thing which, I'm not sure how to put it. So you have the ['Data format field'] where you type it in free form. The thing is, though, someone can tell you this is such and such type of file but there might be other information that you need, … maybe just put more guidance on that box, say include version number of the format if there is one. Or*

maybe even a link to [the data format], a lot of these data formats have like links. And then someone can click on the data 'Oh, we can download that one,' even though they've never heard of this type of data. I've had this problem a lot with open data sets, I download it and they tell me some vague name of what the format is and I have no idea." (Tester 6 - Aquatic)

"*Of course the frequency, once the location, the spatial and temporal resolution, useful for a lot of people*." (Tester 1 – other CENS)

"*People want to say 'what is the frequency of your collection?*' (Tester 9 - other CENS)

"*You know, when I look for data, most likely what I look for is…for example, leaf area index, of course we have to know where it was collected, that's for sure. And then we want to know the time-frame, was it collected the last 100 years, the last ten years. And how often this was collected*." (Tester 11 - Ecology)

These quotes illustrate how testers use their own personal experiences with finding and using other peoples' data in thinking about what our metadata fields do not cover. Their examples largely focus on methodological and process related issues, like the data format, sampling rates, and nuanced descriptions of the data collection periodicity. One tester gave a very simple answer to the question: what is missing from this form? "*Something that lets you explain what my biggest issue was*." (Tester 2 –

seismology). This answer touches on both the importance of personal experience, and of unexpected methodological problems.

7.1.5 Use of existing documents

Much of the data collected by CENS researchers are not textual, and thus non-self-describing. To describe image, audio, or numeric data researchers needed to either create textual descriptions from scratch or adapt existing text from research publications or technical reports to the data description task. The majority of the responses that testers wrote up to describe their data fell into the first category, they were created from scratch. Testers wrote up most fields without looking at any document, web site, or other reference material, generating titles, descriptions, keywords, and variable lists off the top of their heads. When testers did refer to sources, they did so for specific fields. I recorded any time a tester referred to an document, taking an expansive view of "document" to include any files that the tester opened (or browsed through folders to find), as well as servers and web sites visited via a web browser. If a tester referenced the same document for two different fields, I considered those to be two distinct document references. Additionally, I only counted the last document that a tester accessed in looking for a particular piece of information. For example, if a tester accessed a personal web page to find a publication that had information for a particular field, I counted that as one reference to a publication.

Researchers on average referenced documents twice per test, for 22 total references. Two testers referenced documents four times, while one researcher did not reference any documents during the test. 19 of the references we considered to be

"successful" in that the tester found the information for which they were looking. Three references were "unsuccessful," meaning that the testers looked at a document but did not find the desired information. See Appendix IV for a full list of the document references that I observed. Table 7.3 lists the fields for which testers used a reference. Testers most often referenced web sites to find URLs at which data could be found online. The other references were widely distributed among seven other fields.

Table 7.3 – Metadata fields for which testers used a reference (n = 11 testers)

Metadata Field	**No. of references**	**% of total**
Location of the data (URL)	6	27%
Process and equipment used for collection	4	18%
Funding source	4	18%
Variables	3	14%
Dates of data collection	2	9%
Data set title	1	5%
Data collection location	1	5%
Data sharing permission level	1	5%
TOTAL	22	100%

The actual documents that testers referenced varied even more widely. Table 7.4 shows the distribution for the types of documents referenced. Testers most commonly referenced their own authored publications, public web sites, and their laboratory web sites.

Table 7.4 – Document types that testers referenced (n = 11 testers)

Document type referenced	**No.**	**% of total**
Authored publication	5	23%
Lab or Personal web site	5	23%
Public web site	4	18%
Data file folder (for the folder name)	1	5%
Software code	1	5%
Name of data file	1	5%
Date stamps inside data file	1	5%
Column definition file (for variable names)	1	5%
Paper print-out of a database record	1	5%
Personal list of sensors	1	5%
Personal file directories	1	5%
TOTAL	22	100%

This list illustrates how testers largely generated responses to the metadata fields in our test without referencing documents. When testers did reference documents, they largely used their published writings to find equipment manufacturers and funding sources, and accessed public, lab, and personal web sites to find URLs from which a user can access their data. Most of the other referenced documents can be categorized as forms of metadata that were identified in earlier chapters - names of data folders and files, date stamps, lists of data column names, lists of sensors – however, the relative infrequency with which those kinds of metadata documents were referenced indicates that the day-to-day metadata products were not considered important to testers in completing this task.

In summary, in performing our metadata creation test, researchers referenced specific documents in order to fill out specific metadata fields that they could not fill out off-hand. With only three testers conducting “unsuccessful” references, that is, looking in a document for particular information but not finding it, the referencing process is highly

directed, and typically did not involve the forms of metadata identified in the ethnographic portion of my study. Instead researchers looked for funding and equipment details in publications and web sites.

7.2 – Applying metadata themes to the user test

In this section, I describe how the four themes of everyday metadata creation that I developed in Chapter 6 were manifested in our CENS metadata registry user tests. I try to illustrate how these themes were visible in how researchers responded to our metadata fields, and how they manifest in different ways from the everyday situations.

7.2.1 Data processing

Researchers did not include much description of their data processing steps in their metadata descriptions. This is likely due to the fact that they were asked to describe the process and equipment used for data collection, not analysis. One tester took our form to task for this emphasis on describing the data collection and not the analysis, "*Well, you see, you have a data. And then now, you have to process the data in zillions of ways. You do some different filtering, you know, integration, differentiation, whatever the hell you do. And then you come up with a completely different data set that allows you to do some kind of conclusions. That data set is very different from the original data set. Because at the very end, you deal only with the subset of data that is really relevant to the problem. It is not at all similar to what you started with*" (Tester 10 - Seismology). He then proceeded to write the following response for the "Process and equipment used for collection" field, "**Typical seismic instruments like a digitizer, sensor, solar power source, data collection unit**." In the "Data format" field, he responded, "**miniSEED**,"

which is the archival format for seismic data. He did not mention that data are converted from the Mini-SEED format to the SAC format, or that they have generated catalogs of seismic events that took place during this data collection period.

In an opposite example, one tester responded to the "Data format" field with "**matlab**" (Tester 11 - Ecology). MATLAB is a mathematical computing software product that is widely used across science and engineering disciplines as an analysis tool. In this tester's case, soil ecology, his sensor data were collected initially in a plain text format, and he had imported them into MATLAB for analysis. Thus, again data processing is not described, but is implicitly present in the tester's response.

Both of these examples show researchers giving responses that correspond with their expectation of a data use very similar to themselves. In the seismology case, when I asked if potential users of his data would want the original data or his data after the processing, he responded, "*Oh, you have to share the entire thing. Because then you have the data initially. Then you do different processes or steps, let's say 5, 6, 7 different steps. If you just extract one of the steps it is not useful unless you share everything else... If you want to repeat everything you have to start from zero, from scratch*." (Tester 10 - Seismology). Thus, describing his data as "miniSEED," the archival form of the original data, provides an account of his data that speaks to a typical seismologist. Similarly, in the ecology case, the tester describing data as in the "matlab" format and not the simpler plain text format accounts for his data in a way that speaks to more computational and mathematically inclined ecologists.

7.2.2 Assessing data quality

Data quality issues were commonly discussed in the tests and post-test interviews. Researchers were aware of potential data quality problems in their own data, as well as being aware of how potential users of their data might assess the quality of their data. The extent to which these discussions were captured in the testers' responses is another story. The fields in our form did not directly ask for data quality related responses. No testers noted any data quality related issues in their responses, though one tester mentioned that the main thing missing from our form was "*Something that lets you explain what my biggest issue was*." (Tester 2 - Seismology)

Tester 3 described to us his thoughts about the quality of his data: "*Would you be willing to share your data? Yes. But no one asks for it. [chuckle] Because it's, [pause] It's very, [pause] It's very specific, I guess. But no, many people would be able [to use them]? I mean these sensors are not very reliable, that's why mainly. So maybe if someone is using a different accelerometers maybe it's gonna be different, but I'm not sure, maybe not*." (Tester 3 – other CENS) In his response to our metadata form, he did not indicate anywhere that he had data quality problems related to unreliable sensors. In another example, Tester 10 from the seismology team talked about the metadata logs of technical information that the seismic stations automatically collect, and how they are important to understanding the corresponding data: "*Then the question about this metadata, these things, GPS location, voltage, wireless and everything else, this is very specific to the data about this particular site. So, if you see for example, the data is very noisy, we'd like to find out why. Sometimes it helps if you look at the weather conditions;*

you know, the temperature, the voltage, whatever you got. And those extra data can explain what happens to this data. Which means you can use some of this extra data to help you to make sense of the actual data that you would be given." (Tester 10 - Seismology) As noted in Section 6.4.1, the seismic team developed these metadata logs for their Peru project based on the problems they experienced on their Mexico project. Because Tester 10 was describing the data from the Mexico seismic sensing deployment in his test, he did not include any description of these Peru metadata logs, but he did talk about how if he was to fill out the form for the Peru data, he was not sure where the logs would fit: "*if you include the temperature and GPS and voltage then it's not really variable for this data that you collect, it's something different, okay? So you can find some specific stuff about the data and you can find some other stuff like it's not exactly the data. Some of it's metadata which isn't exactly same. So I don't know if you're want to mix it all and put in one pot or those are different things. I guess it depends*." (Tester 10 - Seismology)

Some data users might want to know about data quality issues for other reasons. One tester was a computer scientist who was working on building algorithms for detecting data anomalies. He indicated that he had tried to use data from the CENS aquatic biology project, but the data sets he received from them were only a few days long, which was not long enough in duration for his project because they did not contain enough anomalies (he did not specify what "anomalies" he was trying to find). So he described how he switched to using a larger data set from a non-CENS project after he received a recommendation from a researcher who was also working on anomaly

detection.

7.2.3 Distributing metadata tasks

CENS research takes place in group settings. Individual researchers may not know what to include in certain fields, but do know who in the group to ask. For example, a couple of the testers said that they would need to ask their principal investigator about how to fill out the "funding" and "permissions" fields. In the post-test interviews, we asked testers whether they felt it would be beneficial to allow multiple team members to contribute to their metadata description. Most users responded positively to this idea. For example, one tester said that colleagues could help to ensure that they he had not missed anything important. Another, and engineer on the aquatic biology project, indicated how he would like his science partners to describe their portion of the data because it would take him a long time to put together the description himself:

> ***Q-Test Administrator****: Is there any particular fields on here that you think like you would like to ask somebody else to help you with, just kind of in your immediate project? I mean, you mentioned [your PI] with the permissions or something?*
>
> ***Tester 5 - Aquatic:*** *Right. [pause] No, I wouldn't really require anyone else's help for anything here.*
>
> ***Q – Test Administrator:*** *Okay.*
>
> ***Tester 5:*** *At least related to my own data.*
>
> ***Q – Test Administrator:*** *Right.*
>
> ***Tester 5:*** *If I have to give a real science data, I might require some help. But it*

wouldn't be a straight process where I could give you the data within 20 minutes. It will take maybe a few hours to put that data in place, or maybe one whole day. [chuckle]

7.2.4 State of a project

Different CENS projects are in different states of completion. Many CENS researchers are involved in multiple projects, and have taken part in collection of multiple data sets. The stage of the project, beginning, ongoing, or completed, has an impact on what metadata our testers created, and whether they choose to create metadata for a particular data set at all.

For example, one tester described how he chose to describe a data set that he is already finished using, instead of the data set that he was currently using. His gives his reasoning as follows:

Q-Test Administrator: *Okay. So the first question is just, why did you choose this particular dataset? Are there, for example, are there others which you could've chosen or are there other interrelations?*

Tester 7 - other CENS: *I think this is the most, this dataset was the one that, basically, I had publish on there and so … first of all, if someone sees my publication and wants to replicate my results, they have access to the data set. Second, if they want any sort of derivative works, I think it's my interest to put it out, so they don't have to record data again.*

Q-Test Administrator: *Okay. Were there others that you'd use this year that you could have chosen or...*

Tester 7: *There is one more but I'm sort of still going through it and make sure it's okay once it's final, I'll put it up. So…*

Q-Test Administrator: *Makes sense. So it's about, is it more about your wanting to use it first or more about making sure that it's clean and all these kind of things?*

Tester 7: *Mainly it's me wanting to use it first.*

Another tester noted that he chose to describe a data set that he was sure could be released. He indicated that he had worked on another project, but it was currently ongoing, and its funding status made him unsure whether he would be permitted to share the data now or in the future.

Q-Test Administrator: *Why did you choose this particular dataset, so are there others that you could've chosen which are related ones that you could have chosen?*

Tester 6 - Aquatic: *So this one is particularly useful, I think. It's kind of, it's got two robots deployed off the coast at the same time working different paths at the same time, which is pretty hard to get, so it's probably pretty useful to people. Also I've been doing some other things since then that are, one is a DARPA project, and that's still ongoing. So probably be hard to release stuff from that. I don't know what the actual official difficulties of releasing it but, yeah. So this one is … easier, and there are less issues.*

Another issue related to the state of a project is that with ongoing projects, data might be continuously growing. In these situations, creating metadata descriptions at one

point in time is necessarily an open-ended task. As one researcher noted, **"***I mean the project is evolving and we're interested in different things. So we're collecting always data for the new, uh objectives*." (Tester 3 - other CENS) To indicate this open-ended aspect of his data, Tester 3 responded to the "Dates of Data Collection" field with "**June 2010 – present."** Another tester who was also describing ongoing data sets took another approach, and just described a sub-set of his data. Although he had been collecting data for a number of years, he just responded to the date field with "**2009**" (Tester 11 - Ecology).

It is well known how knowledge of the details of data set collections decreases dramatically after projects are completed (see for example Michener, et al., 1997). With completed projects, forms of external documentation become more important. One researcher looked at a publication from a finished project to find the details about his sensors and other equipment. In another case, mentioned in the "References to Existing Documents" Section, Section 7.1.5, a tester who had completed his projects about six months prior to our test referenced his software code because he could not remember the details of his analysis steps. He repeatedly re-remembered details as he was writing his responses and describing them to us, saying "*Can you tell it's been forever since I looked at this data*?" (Tester 2 - Seismology). The process of metadata creation proved to be a mechanism for him recalling his analysis process.

8. DISCUSSION

Metadata processes, and the metadata products created as part of those processes, are inextricably linked to people, organizations, equipment, software, and to physical and digital objects. Chapters 5, 6, and 7 illustrate how data and metadata exist within these extended actor-networks (Latour, 1987). It is impossible to talk with CENS researchers about metadata without also talking about data, methods, sensors, computers, and colleagues.

In this chapter, I investigate the way that metadata processes and products are interwoven into research practices more broadly. First, I outline how metadata serve day-to-day research needs and longer-term goals of research projects. I then compare the types of metadata I identified in Chapter 6 with the metadata typologies found in the information studies literature that I introduced in my Chapter 2 review. With this picture in hand, I contrast everyday metadata practices with how researchers approach the task of creating metadata for the CENS Metadata Registry user test. I then examine every day practices side-by-side with researchers' use of a structured metadata submission form in order to discuss how metadata serve to ensure accountability of the research process by. I conclude this chapter with reflections on my own metadata practices in relation to this study.

8.1 Metadata in everyday practice

How and where are metadata created, by whom, and for what purpose in the research process? This question, which I gave in Section 2.2.3 as Research Question 1, was the starting point for my study. The vignettes in Chapter 5 provide narratives of lab and field settings that illustrate how metadata practices manifest across projects. In this section I discuss my findings from the vignettes in relation to Research Question 1. The most direct answer to my initial question is that metadata in everyday practice is directed toward achieving specific goals. As Coyle (2010) states, "the metadata that we find ourselves using every day is the metadata that we can use to accomplish some task" (pg. 6). Goals vary by projects and work roles within my case studies. Scientific faculty and students are focused on characterizing physical or biological systems, such as the earth's structure, aquatic micro-organism behavior, and soil carbon dioxide flux. Technology and methods development are another research goal, with researchers from multiple ranks and disciplines collaborating around new soil imaging technologies and soil image analysis methods, new methods for measuring environmental contaminant in situ, and new ways of visualizing the status of seismic deployments. In addition to these research goals, team members such as field technicians and software developers have the goal of ensuring that other members of the team can do their jobs.

Researchers' metadata practices reflect these immediate goals. In regards to characterizing physical and biological systems, as one seismology student noted, "*the data as a goal is only 5% of the entire research. …I mean, it's a tool in order to do some research, but the data itself is not really the goal of the entire interpretation*." Creating

metadata in and of itself is not a research goal, but creating metadata is a necessary activity. As I noted in Section 2.1.5 in my discussion of metadata and ontology, agreed-upon instances of "data," such as sensor readings or physical samples, have "deontic powers" that entail social obligations, responsibilities, duties, and, consequently, actions, including metadata activities (Searle, 2006). Without metadata, data cannot exist. In other words, metadata and data intertwine within the ontologies of scientific research settings. As an example, the Principal Investigator of the seismic project described the importance of metadata as follows: "*Well, the metadata is critical because a waveform disassociated from knowing where the station is makes it absolutely useless. You can't, I mean if I didn't know where it was … it's like, okay, here's a picture of somewhere in here, [but] it doesn't help me. So, um, it's critical in that sense*."

This quote focuses on location metadata. What work is required to create and manage location metadata? For the seismic project, this is a multi-stage process, one that bears looking into in detail as an example of how actor-networks extend outward from even seemingly simple situations. The primary location metadata are latitude and longitude coordinates. During the field installation of a seismic station, the students and research staff who perform the installation install a GPS unit. If installing the GPS unit does not go according to plan, as in Vignette 1 when the GPS cable was too short to reach the roof, a field worker will make a note of the particular problem either in a notebook or in the field laptop computer. This note is then shared orally with any other field workers on site, and shared with the full team via a daily update email to the project mailing list. Once the GPS unit is properly installed, the site software attaches latitude and longitude

coordinates to packets of sensor readings, along with other network health parameters (as noted in Vignette 2), and the coordinates are then transferred through the wireless network and internet back to UCLA. Once at UCLA, the metadata logs are automatically stripped from the sensor readings and are stored in a local UCLA database. The metadata log database was designed by a computer science student on the project. The GPS readings are stored in their own designated table. The computer science student created another table within the database for calculated GPS coordinate averages. These averages are necessary to remove fluctuations in the day-to-day GPS readings caused by the GPS unit's accuracy limitations. To share the GPS coordinates with the rest of the team, the computer science student then created a program that outputs the coordinates as a file. The Principal Investigator for the project was then able to download the GPS coordinates in that file, and upload them into the database of sensor readings, effectively re-attaching the coordinates to the data. At this point, when a seismologist downloads sensor readings in the SAC format from the project database, described in the commentary for Vignette 2, the SAC files are returned with latitude and longitude coordinates in the header.

As this example illustrates, collecting essential metadata, such as GPS coordinates, involves field work and multiple intermediaries, both human and machine, interacting in response to each other (Latour, 1987; Pickering, 1995). I can provide a similar example for the creation of metadata for the soil ecology image collection process. The PI of the ecology project stated "*you basically can't do any, [laughter] you can't address the research questions without the metadata*," referring to descriptions of units and columns for sensor data, soil image resolutions, and time and locations of data

collection. Ecology students work with staff researchers and engineers to perform field work, compile image sets for analysis using metadata embedded in image filenames, and keep track of their image sets via creating new folders and Excel spreadsheets in their lab computers.

These practices are "situated actions" (Suchman, 1987; 2007), in that researchers are always dealing with emergent issues. When the field technician for the aquatic biology project found that she could not download data files from the sensors because of a faulty cable connection, she had to deviate from her planned activities by taking the sensors back to the lab for maintenance after talking with the lead student on the project. Talking is critical to the day-to-day practices of working scientists (Orr, 1996), with talk being supplemented by forms of metadata. In this case, the aquatic biology technician documented her deviation from her normal schedule by adding a comment in an Excel file, named "Sensor_field_notes," specifically created for this task. She then used the same Excel file to note when the sensors were re-installed the following trip.

Around these situated actions, however, routines develop (Agre, 1997). When researchers develop metadata processes and products that are sufficient for their current tasks, their practices solidify. The computer science student on the seismic project described how their data processing pipeline had solidified, "*I, you know, I understand the SAC format now, so I had, between Mexico and Peru, I was in the [data converting] code, … and I realized, oh I could just write [the data] to SAC instead of to Mini-SEED. And I suggested this to [the PI], and [he] said, 'no, the pipeline we have is working, we don't want to mess with it, just leave it alone*.'" In an aquatic biology example, given in

Vignette 5, data files are created by the sensor with a metadata header that describes the sensor parameters and each column in the file. After the field technician downloads the data files from the sensor, she converts them to a format that is easier to import into analysis tools. As part of the conversion, the metadata headers automatically get removed. The field technician indicated that losing the headers was not a problem since it always stays the same. Once they have one copy of the header, they do not need to generate it for each set of sensor readings they download. Additionally, because the header does not change, they can always re-generate it from the sensor again if they so desire.

In the CENS projects I studied, researchers were most likely to break metadata routines and develop new metadata practices when data quality issues emerged. I noted in Chapter 6 how metadata practices are intimately involved in the determination of "bad" data, and how metadata practices develop over time in a project. The aquatic biology "Sensor_field_notes" Excel file is the perfect example of that. The lead student on the project felt that she needed better documentation of field activities and sensor calibrations in order to have more trustworthy data for her dissertation. The "Sensor_field_notes" Excel file was intended to serve in that role, and indeed did increase her confidence that she would have usable data.

Thus metadata enable researchers to use their data to achieve their immediate research goals. Note, however, that I did not include "sharing data with researchers outside of their team" as one of the immediate goals of the CENS projects I observed. In

the next section, I discuss data sharing in the context of the CENS projects in my study, and illustrate how metadata for data sharing are not a priority for CENS researchers

8.2 Metadata for data sharing

Researchers within the projects I observed do not have the explicit goal of sharing data with individuals outside of their projects. This is not to say that researchers are opposed to sharing their data. Within each case study, researchers are aware that their data might be useful beyond their immediate projects. Metadata practices, however, rarely facilitate such sharing. This finding reflects back on Kelty's (2008) notion of "recursive public." CENS projects are not organized around a shared public goal of enabling data sharing.

The CENS seismic project, for example, plans to submit their data to the IRIS archive at some point in the future, though the timeline for when that might occur was not yet decided as of this writing. As I noted in the commentary for Vignette 2, the full documentation required to submit data to IRIS, will only be created when the data are in the process of being submitted. Any data sharing that takes place before the data are submitted to IRIS consists of informal arrangements made through personal connections.

In contrast, the aquatic biology and soil ecology projects each put some portions of their data on their lab web sites. As of this writing, the aquatic biology project's web site contains graphs of sensor readings from the sensors installed in multiple Southern California coastal locations, including those described in Vignette 5. These graphs are accompanied by general descriptions of the projects, including the depth at which the sensors are installed and the temporal sampling interval. The sampling interval listed on

the web site, however, does not match the sampling interval as described by the student and technician working on the project. The web site does not provide links to download or access the full data sets from the aquatic biology project, but does include links to the main project personnel. This would allow a potential user to find out how to contact somebody on the team in order to ask for full access to their data, but again requires direct personal contact between the data authors and data users in order for data to be shared.

The web site for the soil ecology project does provide links through which potential data users can download actual data sets, but only the sensor data sets, not the soil images. The immense size of the image corpus precludes the team from making the images available for download via their web site. Alongside the links to the sensor data sets, however, the team has posted a number of metadata files (note that while I call them metadata files, the web site itself does not call them metadata), including a diagram of the field site, a list of sensors used, a header document for the sensor data files, and a document called "Note to data file," which describes how the data were "corrected" prior to their being posted. The "corrections" described in this document include how points were removed, and how certain measurements were adjusted via a set of calculations. These documents appear to be very comprehensive with regards to what a potential data user would need to know in order to understand and use the sensor data. According to the researcher who put these documents together, a staff ecologist, they are unsure as to whether anyone is actually downloading and using the data. They have yet to be contacted by a data user, and are not tracking web site usage or download metrics. Thus,

the staff ecologist responsible described to me how while he recognizes the importance of data and metadata management with sharing in mind, he saw the development of the data web site as a much lower priority than his every day scientific work.

Within the Environmental Lab, data sharing is less visible. The lab web site has a number of lab protocols available for team members to download, reflecting how the lab is as much focused on developing methods for characterizing environmental contaminants as they are on collecting data about existing contamination. One student indicated that the team had an ongoing project with a partner university that might involve data sharing in the future, but at the time of this writing, little data sharing was occurring.

As I have noted here, data sharing is not a specific goal for the CENS projects in my study. Researchers are interested in sharing their data with other interested individuals, but with the exception of the soil ecology web site, rarely document their projects or data specifically to facilitate such sharing. Research projects, like the CENS projects described here, do not have a "recursive public" organized around data sharing. They do not need to enable widespread data sharing in order to ensure the continued existence and success of their research communities. Data sharing with outside users, if and when it occurs, is considered by different team members to be either an added bonus or a source of additional work, sometimes both.

In Chapter 6 I developed a metadata typology based on my observations of CENS projects. My typology reflects the ways that metadata practices are situated within specific projects and are targeted towards achieving specific goals. In the next section, I

compare my metadata typology with typologies from information science and scientific literatures in order to identify overlaps and differences among them, and to illustrate how this low prioritization of metadata for data sharing emerges as a gap in CENS metadata practices.

8.3 Comparing metadata typologies

In Chapter 2, I outlined the conceptions of metadata in information institutional and scientific settings. The most widely used metadata typology within information institutional settings is from Gilliland (2008): *administrative*, *descriptive*, *preservation*, *use*, and *technical* metadata. Gilliland's typology was developed to illustrate the types of metadata necessary to manage and preserve information collections. In the CENS research setting, however, data collections are in the process of being created. And, as described above, the goal of building a data collection in CENS is to use it in the process of achieving research goals, not to manage or preserve it. As a result, the metadata typology I developed in Chapter 6 does not completely map onto Gilliland's typology.

My typology included six metadata types: metadata to establish *data identity*, to describe *data characteristics*, to document *data quality*, and to annotate *data collection equipment*, *data collection methods*, and *data analysis methods*. My *data collection equipment* category most appropriately maps to Gilliland's *technical* metadata category, but mapping my other categories is less straightforward. In some sense, they could all be mapped to either (or both) of Gilliland's *descriptive* or *use* metadata categories. Data set identifiers, column headings, Excel annotations, and Word document write-ups are all

describing particular aspects of CENS data sets, but at the same time would have value for someone trying to use those data.

My metadata categories appear to map best to Greenberg's (2001) metadata typology. Greenberg uses four categories of metadata: *discovery*, *use*, *authentication*, and *administration*. My category of *data identity* roughly maps to her *discovery* category, as one of the main uses of *data identity* metadata, such as filenames, is to navigate through folders and find particular data. Greenberg lumps Gilliland's *descriptive*, *use*, and *technical* metadata categories into one category, which she calls *use* metadata. She states that *use* metadata "permits the technical and intellectual exploitation of an information object." (Greenberg, 2001, pg. 919) My *data characteristics*, *data collection equipment*, *data collection methods*, and *data analysis methods* categories would all fall into Greenberg's *use* category. Additionally, my *data quality* category maps to Greenberg's *authentication* category, as *data quality* metadata, such as calibration records, notes of equipment failures and fixes, and seismic timestamp corrections, all "support the evaluation of an information object's integrity, legitimacy, and overall genuine quality (Bearman & Trant, 1998)" (Greenberg, 2001, pg. 919). I did not observe metadata practices in CENS that mapped to Greenberg's *administrative* metadata category.

Table 8.1 illustrates the mapping of my typology categories to Greenberg and Gilliland's metadata typologies. As I discussed, my metadata typology categories largely map to their *use* categories. This table also illustrates the absence of metadata that serves *administrative* or *preservation* needs within the CENS teams that I studied.

Table 8.1 – Mapping CENS metadata to existing metadata typologies

<table>
<tr><th>CENS Metadata Typology</th><th>Greenberg (2001)</th><th>Gilliland (2008)</th></tr>
<tr><td>Data identity</td><td>Discovery</td><td rowspan="5">Descriptive, Use</td></tr>
<tr><td>Data quality</td><td>Authentication</td></tr>
<tr><td>Data characteristics</td><td rowspan="4">Use</td></tr>
<tr><td>Data collection methods</td></tr>
<tr><td>Data analysis methods</td></tr>
<tr><td>Data collection equipment</td><td>Technical</td></tr>
<tr><td rowspan="2">N/A</td><td>Administrative</td><td>Administrative</td></tr>
<tr><td>N/A</td><td>Preservation</td></tr>
</table>

Mapping my metadata typology categories to metadata typologies that have come out of scientific disciplines is more difficult because of their high degree of variability. Comparing a few of them on a one-by-one basis shows a mix of overlap and non-overlap. Michener, et al., (1997) gave an extensive list of "standard ecological metadata descriptors," including: *data set descriptors* (ex: title, identification, principal investigators), *research origin descriptors* (ex: project, site, sampling, and research method descriptions), *data set status and accessibility* (ex: latest data set updates, storage location, permission restrictions), *data structural descriptors* (ex: file, variable, and anomaly descriptors), and *supplemental descriptors* (ex: computer programs, publications, history of data set usage). My typology categories map to various categories in this list, often multiple categories at once. For example, my *data identity* category roughly maps to components of their *data set descriptor* and *data structural descriptors* categories, as both of those categories list identifiers of different kinds. Similarly, my *data quality* category maps to both their *data structural descriptors* category, particularly its "data anomalies" component, and the "quality assurance" component of their

supplemental descriptors category. In contrast, my *data analysis methods* category has no clear mapping in Michener, et al.'s typology, with only a few components of their *supplemental descriptors* category being tangentially related.

In another example, my categories partially overlap with the metadata typology provided by Lawrence, et al. (2009) for grid-based data services. Their typology includes five kinds of metadata: *archive* (syntax and semantics, e.g. parameter descriptions), *browse* (context of data), *character* (citations of the data and post-fact quality assessments), *discovery* (for finding data), and *extra* (discipline- or instrument-specific metadata). My *data characteristics* category maps to their *archive* category, with their emphasis on descriptions of the parameters or variables that have been collected. Additionally, *data collection equipment* and *methods* categories are similar in spirit to Lawrence, et al.'s *browse* category, which they describe as characterizing "instruments and/or processes available for producing data; … the location(s) (and observers) of data production; … [and] the projects and campaigns etc., associated with data production" (pg. 1007). Their *extra* category, however, they consider to be "the core discipline-or instrument-specific metadata, which may be strongly typed (i.e. conforms to schema such as [an XML schema]) or consist of arbitrary documents" (pg. 1006). Given that many of the metadata examples I identified in my vignettes could be considered to be "arbitrary documents" in this sense, this *extra* category may be the clearest mapping to all of my typology categories.

As this analysis indicates, my metadata typology best maps to *use* categories in information science metadata typologies, and has varied mappings to the idiosyncratic

metadata typologies from scientific settings. In the next section, I discuss how my use-centric typology makes sense when CENS researchers' metadata practices are analyzed from an accountability point of view. Metadata are created and used by people in particular research roles. Students and research staff are responsible for documenting day-to-day research processes, such as documenting field activities events, labeling samples, and annotating Excel files. These roles give individuals particular relationships to data, metadata, and to each other. I discuss how researchers in various roles are accountable for their metadata practices to their team members and to larger social systems.

8.4 Accountability of metadata practices

My Research Question 2 focused on the social nature of metadata practices: How are metadata creation tasks learned and parceled out in research groups? This question came with a number of sub-questions:

- How do team members make sense of metadata descriptions created by others in their team?
- What counts as an adequate metadata description, and in what situations?
- How are the expectations and norms of different social systems and communities reflected in metadata practices?

To address these questions, I turn to the notion of accountability. Metadata practices (products and processes) are embedded in multiple systems of accountability (Wenger, 1998; Yakel, 2001). Researchers are accountable for performing metadata tasks. In day-to-day work, they are accountable to their team members and to their

supervisors. In one sense, "accountability" refers to the obligation that individuals have to accept responsibility for their tasks. In another sense, Garfinkel's (1967) ethnomethodological sense, "accountable" literally means account-able, that is, able to provide an account, a statement explaining one's actions. Both of these senses of accountability pertain to this discussion of metadata practices. I initially focus on the first sense, that of responsibility for actions, and then discuss the second sense, being able to provide an account of research actions.

In the first sense, the obligation to accept responsibility, faculty rely on students and staff to perform the detailed work activities that lead to publishable research. Similarly, students rely on staff to perform field work, which, as I have discussed, includes creating metadata that relate to field activities and to the data themselves. Research staff are accountable to the scientists who use the data and metadata, in that they are responsible for ensuring that sensor installations are operationable and that trustworthy data are being collected. As noted in Vignette 5, Evelyn, the aquatic biology field technician, explicitly brought up her social accountability in discussing her problems in downloading data from sensors. She stated that by leaving at least one sensor working, even with the problems they've been having, she can at least tell the Principal Investigator that they are still collecting data. If she had to take both of the sensors out, which ended up being the case, she could not say that anymore.

As much as students and staff are accountable to project leaderships, Principal Investigators and lead graduate students are accountable as well to students and staff. Among many other things, students and staff rely on project leaders to make decisions

about where sensors should be placed, what sampling rates to use, and how data should be integrated within and across projects. When researchers receive conflicting or confusing directions from leadership, their work becomes more difficult. In one case I observed, for example, a project PI said in an interview that a member of his research staff was responsible for migrating data and metadata from one server to another. The staff member, on the other hand, indicated that he had no knowledge of this task.

The seismic project also has distinct field workers. As outlined in Vignette 1, the installation process involved many people, including students and staff. As the equipment proved their reliability, the tasks of documenting the day-to-day field work largely shifted to the lead engineer on the project, as he noted himself: "*So I am the sort of the project manager, in the sense that it has been my role to sort of make the deployment happen …when we began the Peru experiment we had learned enough painful lessons that it was clear that we needed somebody from the computer science side deeply involved in many many aspects of the deployment. The original concept was that I was going to be a manager of people, as well as other things, um, in practice it mostly worked out to be I did much of the technical work, after the original site preparation. Um, You know, I configured the radios, made sure that the WiFi worked, have nursed it along for nearly three years now, and so on.*"

His lead role also extended to metadata practices, particularly the creation of metadata processes and products that documented field activities. In addition to highly detailed paper notebooks of site maintenance, which were often transcribed into email updates to the project mailing list, he developed an Excel spreadsheet that documented

the state of each station in the network during the installation process. He also described how he also had "*a rather large archive of photographs*" of field sites, which were useful for recalling where sensors were installed and also for reporting to police in the case of vandalism. The automatically generated logs of network and technical metadata described in Vignette 2 were the result of his collaboration with the lead computer science student on the project. All of these forms of metadata, and the processes by which other field workers contributed to them, helped to increase the reliability of the seismic sensor network, and as such, increase the accountability of the lead engineer as project manager.

Students are also subject to social accountability, in that their actions affect other team members. Students often help each other, and field technicians, in the field data collection process. I participated on a field excursion where two members of the ecology team, one student and one research staff, filled in for a field technician who was on vacation. They were tasked with collecting soil images using the manual imaging system. This task typically takes over three hours, and involves taking approximately 750 pictures using a handheld camera system attached to a laptop. Both of the team members had done this task before, but, on this occasion, they almost ran out of battery power for the field laptop before they could finish collecting all the images. By accident they left the laptop powered on while we ate lunch (possibly related to my own presence in the field interrupting their routines). Shortly after resuming their work, they started getting "low power" messages. They proceeded to work much more quickly and were able to finish the imaging task without having the laptop die, much to their relief. Earlier they had told

me how they forgot to bring a package to the field site; the leaving the imaging task incomplete would have compounded their error. As Caroline, the research staff member said, "*I was worried that we were going to have two things go wrong, first we forgot [the] package, and then we didn't get the data collection finished, but we did*." When I asked what they would do with the images once they returned to their lab, they indicated that they would leave the images on the laptop for the field technician to download to their project database upon his return from vacation.

Social accountability, in the sense of an obligation to take responsibility for actions, extends to metadata practices. When students or staff leave an ongoing project, they are accountable to the team members who pick up where they left off. Caroline, the lead on the soil image analysis project for the machine-collected images, described how "*I left all of the documentation that I have in my notebooks on my computer. So, any of the metadata and all of the analyses are still on the computer because our undergraduate researchers are still employed*." In this example, she was accountable to her team, and her undergraduate research assistants in particular, for their ability to use the documentation that she had produced for the project going forward. In another example, a member of the environmental science project described how metadata was more important for group projects than for individual work. For her own work, she described how she can typically use her data, "the numbers," without metadata. The exception is situations where large groups of people all go out to collect data for the same project. In those cases, she described how metadata are necessary to get data from everybody back

in the same place, metadata provide some assurance that the data are where they are supposed to be.

Researchers thus make sense of metadata created by other researchers in their team within this framework of accountability. If a researcher has trouble understanding a metadata description, they contact the accountable individual and ask about a particular issue. This is common on the seismology project. Researchers send emails over the team mailing list asking questions about incorrect time stamps, equipment maintenance, and missing GPS coordinates. In another example, an environmental science student described how most graduating students take their lab notebooks with them after they graduate. To use data files that are left behind on lab computers, they will contact the student that has left, "*like I know [a graduated student] did this type of analysis and I will just send her an e-mail and she has her notes and will get back to me*."

Beyond the immediate social accountability, researchers are accountable to standards of scientific practices that are part of their professional circles. As Shankar (2007) describes, "the academic scientist is afforded great latitude of choice in recording technology, the ordering of data elements, the integration of external documents, and so on" (pg. 1463). Researchers learn, however, as legitimate peripheral participants in communities of practice (Lave and Wenger, 1991; Wenger 1998). As noted in Section 6.3.5, nobody I talked to had taken, or knew of, formalized courses in data management. Beginning students, such as undergraduate assistants, start with simple but essential tasks. In the Environmental Lab, undergraduate assistants clean lab equipment and help with sample preparation, as noted by a more advanced graduate student: "*When [the*

undergraduates] start up, they do a lot of prep work for us. And then as they get more familiar with some of the lab techniques, we have them help us with some of the analytical processes. Or if they get to a pretty self-sufficient point then ideally we'd give them maybe like a small project that they can tackle." Similarly, in the CENS soil ecology project, undergraduate assistants crop soil images and assist in the image analysis process. Undergraduates participate in the soil image processing pipeline, from measuring roots using the image analysis software to transcribing paper data sheets into Excel. For example, the root image analysis Excel files have metadata notes written by undergraduates that indicate who compiled a certain set of data and on what date.

Shankar (2006), focusing on laboratory notebooks as "records," illustrated how researchers learn and develop "record-keeping" practices as part of the process of building professional identities. Researchers develop practices through trial and error, through social interactions among peers, and through examples from more experienced researchers, as exemplified by the leader of the Environmental Lab leaving her lab notebooks from her own Ph.D. in the lab for her students. In this way, researchers learn how to become "accountable" in the second sense of the term: able to provide accounts, statements explaining their actions.

Researchers need metadata to perform everyday work, such as time and GPS stamps on seismology data, calibration records to use aquatic biology data, soil image file name syntaxes, and labels on water samples. The metadata that people create in their everyday work, however, is always selective and incomplete. Researchers often described gaps in documentation and feelings of ambivalence regarding their metadata practices,

and in some cases laughed at their own practices. Garfinkel (1967) illustrated how seemingly "bad" practices, in his case practices for medical recordkeeping, can actually serve the "medico-legal enterprise" in which they are situated without problems as long as they meet the "expectations of sanctionable performances" (pg. 198). In other words, with knowledge of the settings and situations in which patient records are created, readers are be able to understand the place of the records within the clinic despite any problems with the records themselves.

Similarly, CENS researchers' metadata practices that seem inadequate, incomplete, or problematic to an outside observer may not be so to someone inside the community. CENS researchers account for their metadata practices in a number of ways. A seismology student, describing his need to keep track of equipment fixes in the following manner: "*You have to keep track of everything because first, like if you want to do the instrument correction …, you need to know which sensor came from where, because different sensors have different responses. … And also, you want to know if there's some problem that keeps repeating itself at the same site, you want to see, you know, what's actually happening. So … we don't have really substantial sort of notes, but good enough... ."* When I followed up this comment by asking him if they had standard types of documents to keep track of equipment fixes, he said, "*No, it just goes through the email*." Thus, "needing to keep track of everything" was achieved through emailed notes that were "good enough." In this case, he accounts for the "good enough" notes by making it clear that he is aware of how important it is to have metadata about equipment fixes.

In another case, when I first introduced my study to a student on the aquatic biology team, I described how I was interested in, for example, how data are organized after researchers bring them back to the lab. The student's immediate reaction was to laugh and say, "not very well." Later, in an interview with that same student, she expressed her own opinion of her data organization methods: "*I mean it's certainly not a very fancy organizational system, it's a bunch of folders. And it's reasonably doable right now. …But time-wise, I know it's not huge but it probably should be bigger and then probably make things easier*." She then noted that her field technician took on many of the data management, and hence metadata, activities: "*I mean as far as the organization and the upkeep of it, Evelyn's been really good... She goes out every other Friday, downloads the sensors, basically, just looks through the data, makes sure it's in a good format… .*" Thus, the student can account for the self-perceived limitations in her practices by pointing to the ways that her efforts combine with the field technician's metadata practices to meet the standards of good practice.

I described in Section 6.4 how researchers develop metadata practices over time as a project evolves. Practices can also fall away over time. A research staff member within the soil ecology team showed me a series of files that he considered to be metadata for his sensor data. Among them was an Excel file of sensor calibrations, sensor changes "*At one point I was keeping track of how we did the calibrations, [and] when we replaced the sensors for example. We kept track of those things. But … I'm not doing it any more [laughs], I'm doing it, but not keeping track anymore*." Later in the discussion he said that he did write calibration information in his field notebooks, he just stopped

transcribing them to the Excel file. He said that it would be hard to go back and find it in his notebooks, but he has it. In this case, he indicated that he stopped transcribing the calibration information into his metadata files because of the work effort involved. His view of the metadata creation process, as part of the data management process, is that it is important, but time consuming, stating, "*My thinking is always basic minimum, that's it. I don't want to get into data organizing management business, that's not my job*." In multiple discussions with him and other team members, however, it was clear that the "data organizing management business" was his job, by default if not explicitly. When the lab put up a project web page that allowed sensor data sets to be downloaded, he was responsible for putting together metadata files such as headers that listed column names and units, sensor lists, and field site layouts. By creating such documentation, he is able to illustrate his own accountability (in the first sense - obligation to accept responsibility) for metadata tasks. With this first kind of accountability in hand, he can account for his incomplete calibration files by pointing out that he has limited time for metadata tasks, and that he can look up calibration in his notebooks if necessary.

This soil ecology example illustrates how metadata practices can be deliberately dropped. In other cases, metadata practices may not have changed, but the products produced can be lost over time. In the environmental science lab, a student perfectly described the ways that best efforts can go awry, "*I'm sure if you think you're being good about like annotating what you need to take notes of, but then somewhere down the line, it's like three months, six months, a year from now, you go back to look at it, 'I swear I made a note of this'. And now, I can't find it or I actually didn't. I think that probably*

comes up more frequently than we'd like." In her case, as with other in the Environmental Lab, she reduces the impact of losing metadata by running multiple replications of an analysis and by retaining samples so that they can be re-analyzed later if necessary. Samples can be retained "more or less indefinitely," which makes re-analysis possible, but also makes it necessary to have effective systems for sample labeling, as described in Vignette 4. Researchers can thus account for lost metadata in individual situations by illustrating how their metadata practices work effectively in the majority of situations.

Looking across these four examples, the metadata practices of CENS researchers meet the "expectations of sanctionable performances" (Garfinkel, 1967) in the sense that apparent gaps, inconsistencies, or mistakes can be accounted for - researchers can provide statements explaining their actions. In cases where practices do not meet the "expectations of sanctionable performances," as with the aquatic biology technician whose poor sensor calibration documentation resulted in six months of unusable data, sanctions such as losing a job are real.

In these examples, accounts for metadata practices come in four varieties:

1. Our metadata may not be complete, but we know what we need to document and can do so if necessary.
2. My metadata practices may not be sufficient individually, but as a team our practices are sufficient.
3. I do not have my metadata processes and products all available in a displayable form, but I could if I had enough time.

4. My metadata practices may have been inadequate in one situation, but I can show you many other situations in which they were.

Through having practices that enable these kinds of accounts, researchers can have incomplete, limited, or occasionally problematic metadata and still meet the "expectations of sanctionable performance" of scientific research. Metadata adequacy, I argue, is determined by team members' shared social understandings of accountability. If researchers can account for any perceived problems in their metadata, then their identity as a researcher within a community of practice will not be challenged on that regard.

Knowing what types of metadata practices facilitate accountability requires becoming part of the community of practice. The researchers in all four of the examples I give above of "account-able" practices are experienced researchers. Three are advanced graduate students with considerable field experience from working on multiple projects, and the fourth is a research staff member who has more than a decade of field data collection experience. Through engaging in research activities over time, researchers come to understand the institutionalized rules, expectations, and norms within their social settings, even if they cannot directly state what those rules, expectation, and norms are. As Wenger (1998) notes, "[t]he regime of accountability…may not be something that anyone can articulate very readily, because it is not primarily by being reified that it pervades a community" (pg. 81). In the four examples of "account-able" practices that I list above, I interpreted researchers' practices and statements in terms of their social accountability for those practices; researchers themselves did not articulate their activities in those terms.

Researchers did, however, articulate their metadata practices in terms of whether they were sufficient to achieve their research goals. By using language such as "*good enough*," "*it's certainly not a very fancy organizational system," "my thinking is always basic minimum," "I think that [losing track of notes] probably comes up more frequently than we'd like,"* to describe their metadata practices, researchers acknowledged the inherent incompleteness of their efforts. Researchers develop an understanding of the situations in which metadata incompleteness can be tolerated and accounted for as part of their professional and personal development as scientists (Shankar, 2002; 2009). Senior graduate students and research staff have experienced the ups and downs of multiple projects. They know that field sites are unpredictable, that equipment fail unexpectedly, and that data collection processes may need to be adjusted as a project proceeds. These events become part of the social ontologies of their research settings (Weissman, 2000). Experienced researchers are adept at negotiating the regime of "mutual accountability" (Wenger, 1998) that requires them to document both routine and unanticipated events that occur during the research process, while at the same time fitting metadata practices into the other tasks, both social and individual, that they are expected to perform (such as manipulating machines, writing papers, helping team members, etc.).

In the next section, I turn this framework of accountability toward researchers' user tests of the CENS metadata registry.

8.5 Creating metadata for a community registry - Accountability to whom?

Standardized forms inevitably reconfigure work practices. Attempting to apply an outside standard to a particular type of work practices can force individuals to simplify,

constrain, and change their actions (Bowker & Star, 2000). In the CENS metadata registry user tests, I asked researchers to create data descriptions using a standardized set of metadata fields. To perform the task, researchers had to decide how to summarize their complex and situated activities. As in Chapter 7, the bolded passages below indicate actual passages from users' submissions to our test form.

Creating metadata for a community registry is a completely different task than creating metadata in everyday situations. The day-to-day system of accountability does not apply. In a centralized community metadata registry, accountability is to future users, not to immediate team members. But, researchers do not know who the future users of their data might be. In this section I argue that without a clear future user in mind, researchers give metadata descriptions that might be usable by someone to whom they might be accountable. For CENS researchers, any further accountability is a professional issue. So without particular users in mind, metadata descriptions are created such that they might be used by researchers who are part of the same community of practice.

The first example of this comes from testers who are part of the CENS aquatic biology project. One tester, a biologist on the project, responded to the "Process and equipment used for collection" field on our form by describing a series of sensors, including providing the sensor manufacturer: "**Unattended sensors, including Hydrolab series 5 water quality sondes, WetLabs WQM sensors, Sontek Argonaut ADCP. Turner Designs Phytoflash active fluorometer also used.**" (Tester 8 - Aquatic) An engineer on the project, on the other hand, did not mention "sensors" at all in his response, even though all of the variables that he listed in his response to the "Variables

collected" field, "**Temperature, conducivity/salinity, depth**" (Tester 6 - Aquatic), were also listed by the biologist in response to the "Variables collected" field. Instead, the engineer gave a description of the robotics technology and how it was used: "**2 webb slocum gliders deployed for several days, each traveled a different path and collected the same kinds of data.**"

This variation in responses, I argue, illustrates the different communities to which these testers are accountable. To the biologist, it is important that data users know which types of sensors were used. To the engineer, the sensors are not as important as the technology used to move the sensor around, even though he did not foresee roboticists wanting to use his data because of its specificity to an aquatic environment, "*It is a pretty specific data set. … This kind of data [roboticists] don't really look for because there's not much they could do with it. This very specific, like environmental type of data*." (Tester 6 - Aquatic) But lacking any other clear users of the data, his response still reflected his accountability to his own engineering community. The differences in the way that an aquatic biologist and engineer describe their "processes and equipment for data collection" illustrate how sensors are much less visible and central in the ontology of a roboticist than in the ontology of a biologist. To a roboticist, sensors are essentially a payload that could be anything, whereas to a biologist, a sensor is the mediator that allows them access to their phenomena of interest, and as such is of critical importance and visibility.

Two testers from the CENS seismic team exhibited similarly divergent responses. A student on the project used the data collected in Mexico as one part of a computer

science project. His response to the "Process and equipment used for collection" field stated, "**The seismic data was downloaded from the [project] stp server at [the partner institution].**" (Tester 2 - Seismology) In response to the "Data format" field, the computer scientist student listed "**SAC**." A seismology student, on the other hand, responded to the "Data format" field "**miniSEED**," and responded to the "Process and equipment used for collection" field with: "**Typical seismic instruments like a digitizer, sensor, solar power source, data collection unit.**" (Tester 10 - Seismology) This reflects the way that the seismologist is accountable for the data in a different way than the computer scientist. The computer scientist, although involved in developing software for the seismic projects, is describing his role as a user of the data from a computer scientist point of view, including the way that he accessed the data and the format in which he used the data. In contrast, the responses of the seismologist reflect a need to know about the field installations, and the format, "miniSEED," is the data format in which seismologists are most likely to share data, even though the seismologist also primarily uses the data in the SAC form.

I also saw examples of this professional accountability in the responses of testers from CENS projects outside of the four that I included in my ethnographic study. Two engineers from separate projects gave responses that indicate their expectations of data users. One tester responded to the "Research Question" field with, "**Assimilate these real-time measurements to improve model predictions,**" (Tester 1 - other CENS) but nowhere in his responses to other fields did he indicate what he was modeling or trying to predict, though his response to the "Variables collected" field, "**Soil electrical**

conductivity; Soil moisture; Total dissolved solid; and soil temperature," provide clues that this data is most likely to be used by researchers doing soil research. Another tester from a CENS electrical engineering project responded to the "Research Question" field with, "**Containing a variety of anomalies to be detected using different techniques**" (Tester 9 - other CENS), but did not indicate elsewhere in his responses what "anomalies" he was trying to detect. In his case, he was using publicly available data that he found through a recommendation from another researcher who was doing similar work. "Anomaly" is a well-known term in sensor-network engineering community that refers to an observation that appears to be inconsistent with the majority of a data set (Barnett & Lewis, 1994; Rajasegarar, et al., 2006). To the community who is familiar with this data set, "detecting anomalies" is a term that that refers to a particular type of research. In these two examples, then, the metadata accounts created are targeted toward the communities in which the testers are members.

In another interesting comparison, two testers, both computer scientists in CENS, were working on similar problems but in different projects. Both were working on using sensors to detect and characterize the motions of a human body. This type of research has many potential medical applications. One tester on such a project gave very brief responses to the "Variables collected" and "Process and equipment used for collection" fields, writing "**Acceleration**" and "**Accelerometers**" in those fields respectively (Tester 3 - other CENS). In the "Keywords" field, however, he gave a much more detailed response, "**Activity classification, Accelerometer, Naive Bayes Classifier, Support Vector Machines, Decision trees.**" The last three terms in that list, "**Naive Bayes**

Classifier, Support Vector Machines, Decision trees" are terms for particular algorithmic methods in the computer science machine learning community. In contrast, the other CENS computer science tester working on a similar project gave very detailed responses to the "Research Question," "Variables collected," and "Process and equipment used for collection fields. For example, his response to the "Variables collected" field was, "**Accelerometer and Gyroscope information from the right iliac crest (hip), Treadmill walking speed information, Metabolic cart information providing VO2 rates.**" He also provided a very detailed description of the sensors he used. Interestingly, after the test, this tester showed us his personal web page where he had posted the data from his project. On that page he gave a brief description of the data set, a four sentence description of his data sharing policy, and a requested data citation in the BibTeX format. This individual was the only tester to have developed such a page, or the associated descriptions on his own. His detailed responses to our metadata form reflect his already developed thought on sharing his data, and sharply contrast the very brief and community specific response provided by the other CENS researcher doing similar work.

Across our CENS community metadata registry testers, the metadata descriptions they created reflect their uncertain ideas of to whom they might be accountable for their metadata descriptions. Their descriptions were typically brief, written without consulting external documents (as noted in Chapter 8), and provided accounts that spoke to their membership within particular communities of practice. Their metadata descriptions used very particular terms of art, acronyms without explanation, and glossed over portions of their study that were of lesser importance to members of their own research topic. In

short, without well-defined future users of their data, testers created metadata descriptions that were intended to be accountable to their primary professional communities.

8.6 Reflection – Metadata practices for studying metadata practices

In studying the metadata practices of CENS researchers, I have also had the opportunity to examine my own metadata practices. In this section I reflect on my own practices in relation to my findings and discussion. This reflection serves two main functions. First, it serves as a post-study examination of my own research methods and experiences. Second, it provides me my first opportunity to apply my findings to a new type of research: individual-based social science research.

Looking now at my own data and metadata practices, my data consist of field notes, interview audio recordings and transcripts, photographs, CENS research publications, and assorted other documents and files that I gathered from CENS researchers during my ethnographic work. Additionally, I have data from my CENS Metadata Registry user tests: audio recordings, transcripts, videos, and submitted responses.

My own metadata practices primarily consist of creating descriptive filenames and organizing data files into folders, with my folder system organized around the data types. I have folders for interviews, field notes, publications, etc. As my project began and I started accumulating files, I began having trouble finding documents related to a particular field trip or interview because the file name syntaxes I was using were not easy to scan visually. For example, I used a syntax of "[CENS project]_notes_[date]" for my

field notes, using university acronyms for the CENS project portion of the file name. Because many universities begin with the word "University," I ended up with a large mass of files beginning with the letter "u." This made it confusing to discriminate among documents related to my different case studies. Similarly, my interview transcripts were identified by the name of the interviewee and the date that the interview took place. As my interview protocols, notes, and transcripts piled up, it became very confusing to identify which documents were related to which case study. In response to this problem, I began separating files by the CENS project with which they were associated. I created four sub-folders within each data type folder for my four case studies. For example, within the interview folder, I created separate folders where I kept my interview protocols, interview transcripts, and interview notes for each CENS project in my study. I identified each interview transcript or note file with a filename that included the interviewee's first name, last name initial, the term "int," and the interview date. Similarly, I identified my field notes by the CENS project and date of field trip. Within each field note file, I wrote a short description of where and when the activities took place, along with a list of the CENS researchers who were present. I can call these "headers" because they are at the beginning of the notes.

Looking at my files at the completion of the project, however, some inconsistencies are apparent. My identifiers for each project are not consistent from the beginning to the end of the project. I did not settle on the labels for my four case studies – seismology, environmental science, aquatic biology, and soil ecology – until I began writing up my results into this dissertation. Thus, some folder labels and filenames refer

to these projects by different terms. Additionally, the headers that I created within each field note file are not consistent in what they include. Some include precise times and people names, others do not.

Another inconsistency is that all of my documents for our CENS Metadata Registry user tests identify testers by their last name, instead of their first name as with my ethnographic interviews. Thus, transcripts for the people who I interviewed both in the ethnographic and user testing portion of my study do not use the same identifier in the filenames. This is not a problem for me, as my study sample is small enough that I do not have duplicate last names to disambiguate, but for someone outside my study, this link between data from the same individual is not explicitly noted in any of my documentation.

Following from that point, I submitted a description of my data to the CENS Metadata Registry as part of my 2011 CENS annual report. My full responses are shown in Appendix V. Being very familiar with the schema, I did not have trouble understanding the fields. I did reference a document that contains my research questions to fill out the "Research question/why the data was collected" field, but otherwise did not use external documents in creating my descriptions. Most notably, however, I realized in creating and submitting my data description that our system was not created with our own data in mind. In response to the question, "Would you be willing to share your data?," I had to answer "no" because my data are covered under an Institutional Review Board (IRB) agreement that stipulates that my data will not be shared with people from outside my project. Thus, I can create a data description that allows people to discover them, but

I cannot actually share my data. I did indicate that "Data release is prohibited by IRB restrictions" in the "Data Sharing Permission Level" field. Thus, though much of my research, and the research of my immediate team members, is devoted to understanding and facilitating ways to make research data more widely sharable, we are not able to share our own data. In relation to this issue, we are now investigating ways to create IRB agreements that both meet the institutional imperatives for ethical human subjects research and allow us to serve as examples of our own research agenda with regards to data sharing.

I now turn to the framework of accountability that I developed earlier in this chapter. As dissertations must be, my dissertation study is an individual project. I did, however, work with Jillian Wallis, a fellow graduate student, on the CENS metadata registry portion of my study. As I noted in Chapter 4 in outlining my research methods, Jillian took part in the design of the CENS metadata registry user test protocol and was present as a co-administrator for nine of the eleven tests. Jillian and I thus exchanged numerous documents related to the tests, including test protocols, notes, and transcripts, as well as audio and video files of the tests themselves. We shared these documents via email and an online file storage service. Our own metadata for these documents are often inconsistent, as we did not discuss syntaxes for naming test documents, and audio and video files were often shared while still identified by the default automatically-generated camera names. Thus, our shared file folder contains files related to the same test that are not clearly identified as such. For example, I named my post-test write-ups as "notes," while Jillian named hers as "memos." In addition, we used different first and last name

combinations to identify the corresponding testers for our notes/memos. Because Jillian and I saw each other in person on a regular basis, these discrepancies were not an impediment to continuing our work. As I was the first person to use the data that resulted from our user tests, I took on the principal responsibility for writing up the test protocols, having the test audio-recordings transcribed, and the testers' submissions compiled. In this sense, I was accountable to Jillian to ensure that the documents were documented such that she could understand and use them.

In addition to my immediate accountability to my colleague, I needed my metadata to allow me to be accountable to my dissertation advisor and committee. In this sense, I needed to be able to give accounts of my work that met the expectations of competent ethnographic and social science research. I thus created a number of documents in which I kept track of my field trips and interviews, recording when and where they took place, as well as the duration of a particular activity. I used these documents when writing up my methods sections in Chapter 4.

Another reflection about my own accountability for my metadata practices relates to the aforementioned IRB agreement surrounding my data. The IRB data sharing restriction illustrates an important difference between my project and the CENS case studies included therein. Because my research involves human subjects, I am also accountable to the people I am studying. Part of my IRB agreement is that I must anonymize my descriptions of my research subjects. My personal data and metadata are not anonymized, mainly because doing so would be a difficult task and I have no incentive to perform this task because I cannot share my data anyway due to IRB

restrictions. In this dissertation, however, I use pseudonyms instead of real names, and removed identifying characteristics as much as possible. These descriptions can never be perfectly anonymized; any attempt to do so would completely obfuscate the points that I am trying to make in my study. Walking the line between anonymization and utility has required me to decide upon descriptions that are "good enough" to serve both purposes, in the same way that CENS researchers negotiate what counts as a "good enough" metadata practice in their own work (Luttrell, 2000). Determining when a description is "good enough" is an ethical stance, in that it entails balancing obligations and consequences (Collins, et al., 1994). In my case, I have had to balance my obligation to report on my research in a professionally competent manner with the consequences of potentially identifying my research subjects.

As my final reflection, it is important to state that I am accountable to my research subjects for creating descriptions that depict them as active and embodied human beings, not as stereotypes or automata (Clarke, 2005, pg. 73-78). To achieve this, I have illustrated the complexities of their work, my own interactions with them, and the multiplicities of the role of metadata in their daily practices.

9. CONCLUSION

In this study, I investigated how, when, and whether researchers within the Center for Embedded Networked Sensing (CENS) create metadata for their data. Through an ethnographic investigation of day-to-day metadata creation and a user test of metadata creation for a community metadata registry, I illustrate how metadata are one component of scientists' and engineers' work practices. By identifying types of data and metadata being created and used in each of my four case studies, I develop a metadata typology that reflects researchers' everyday metadata practices. This typology describes the six types of metadata I identified within the everyday practices of CENS researchers, namely metadata related to *data identity*, *data characteristics*, *data quality*, *data collection equipment*, *data collection methods*, and *data analysis methods*. These categories only partially map to established information science and scientific metadata typologies, with the best mapping being to categories of *use* metadata.

I then discuss how metadata practices fit into social systems of accountability, using the framework of accountability in two senses. In the first sense, researchers are accountable for performing metadata tasks that contribute to the continuation and success of a research project. These tasks include documenting details of field work and equipment, annotating data processing and analysis steps, keeping track of physical samples, and noting any data quality issues. Researchers are also accountable for

metadata tasks in a second sense: they must be able to provide accounts of the metadata practices that meet with the "expectations of sanctionable performances" of academic research. Researchers cannot create metadata that document *every* detail of their work and data. As Garfinkel (1967) has shown, any attempt to do so would be inevitably futile. Thus, researchers must be able to provide accounts of their metadata work that explain their actions, particularly in situations in which metadata appear inconsistent, partial, or lost. In accounting for situations in which metadata appear problematic, researchers describe how any metadata problems are anomalous, correctable, or in fact not problematic at all. These accounts allow researchers to perform the articulation work (Schmidt & Simone, 1996) necessary to align team research projects, new technologies, and complex field environments in order to achieve their research goals.

My study also underlines a fundamental tension between metadata created for immediate use by researchers themselves and metadata created to facilitate the sharing of data with researchers outside of immediate research teams. Data sharing is not an immediate research goal for the CENS projects I studied. Researchers were interested and willing to share their data, but their day-to-day metadata practices were not aligned toward facilitating sharing. In creating metadata for a centralized metadata registry using a standardized schema, researchers also create descriptions for which they are accountable, but the focus of accountability changes. Instead of being accountable to immediate research partners, researchers creating metadata descriptions for a centralized registry are accountable to potential future users. Because most CENS researchers are unaccustomed to sharing data with people outside of their immediate projects, the

"potential future users" of their data are not clear. Researchers thus create metadata descriptions that are accountable to the most likely future users of their data: other researchers in the same community of practice. Researchers selectively describe their data by using terms of art and emphasizing the particular attributes that are of most interest to other researchers like themselves. This reflects their primary accountability to researchers within their broader social communities.

The metadata practices I observed, both in day-to-day research activities and in creating descriptions for our CENS metadata registry, illustrate how data sharing is not an activity that leads to professional awards within the field-based sciences in my study. Professional advancement within these fields requires students to be accountable to their immediate projects, but not for the usability of their data beyond the lifetime of their projects. Researchers' personal and professional motivations do not stem from thoughts about data posterity. Documenting data so that they can be used by an outside user is a fundamentally different task then documenting data for one's own use. As such, from the point of view of the cyberinfrastructure visions laid out at the beginning of this dissertation, researchers' metadata practices could be seen as insufficient and short-sighted.

In contrast, I argue that the cyberinfrastructure visions are fundamentally mis-aligned with the realities of the day-to-day metadata practices of researchers in small-scale field sciences. Metadata are situated in regimes of mutual accountability in which researchers learn what is important to document, what counts as sufficient documentation, and how documentation practices are to be accounted for in social

research settings. Researchers develop social ontologies in which "metadata-for-data-sharing" have very low visibility. Cyberinfrastructure developments are predicated on having data sharing as a central goal, mission, task, and feature. In their everyday work, however, researchers are not accountable for data sharing beyond their immediate research teams. Cyberinfrastructure systems that try to institutionalize accountability for "metadata-for-data-sharing" in small-scale field-sciences will likely conflict with the community norms, expectations, and practices of researchers within those fields. As Wenger (1998) notes, "an institutional system of accountability is unlikely to be very effective unless it is integrated into the definition of competence of the communities of practice it is meant to align" (pg. 245). The question for cyberinfrastructure developments that emerges from my study should not be, "how do we get field scientists to create better metadata?" Instead, it should be, "how can we create institutional structures in field-based sciences such that researchers organize their data and metadata practices around data sharing and re-use?" In other words, the question should not be about figuring out whether "good" metadata cause data sharing or vice versa, rather, the question should be about how we can construct self-propagating systems (social and technological) in which metadata practices and data sharing efforts align in mutually beneficial and accountable ways.

9.1 Implications

My discussion suggests that there is a mismatch of goals, incentives, and roles for metadata in the development of cyberinfrastructure data systems, such as distributed community data repositories, for the small-scale field based sciences. Researchers in the

small-scale field science projects I studied have nobody with the explicit role of creating metadata. Instead, multiple team members contribute to the documentation of different parts of a project. In addition, researchers have no clear future user base for their data, although they can usually imagine how their data might be useful to someone else. How can metadata be critical to the success of distributed systems for data management, sharing, and curation, while at the same time being highly situated, personal, and often invisible work (Shankar, 2007; Trace, 2007) for the researchers who are the target contributors and users of such systems? This apparent mismatch has a number of implications for the development of shared data and metadata repositories within cyberinfrastructure projects. In the next few sections, I outline implications for three particular stakeholders within these projects, the working researchers themselves, funding agencies who support cyberinfrastructure and data repositories, and data curation professionals.

9.1.1 Implications for working scientists

Working scientists and engineers in the CENS projects described in this dissertation have no immediate reason to prioritize formal metadata products that conform to disciplinary or community standards. Researchers' main interests lie in using data, not documenting data so that it can be used by someone else. They use informal metadata processes such as direct personal communication, email lists, one-off methods documents, personal notebooks, and data file annotations to serve as metadata that "describe, provide context, indicate the quality, or document other object (or data) characteristics" (Greenberg, 2005).

Systems that require formal metadata products can be impediments to their daily work. A couple of researchers described how they had been part of standardized data collection efforts in the past, and how it proved to be difficult to achieve. The aquatic biology Principal Investigator described his experience in the following manner:

"I've worked with larger oceanographic programs where we had to contribute all of our data to a central data management group. The idea was, we were going to do several large, and I mean large, meaning multinational programs in the ocean and different parts of the world ocean. Those datasets would be generated in a somewhat consistent way, put it into a central repository and then in the end, modelers took over, I mean, they had a last call for proposals where it was just modeling and it was using those databases. Some of the information chunked in very easily so you can put salinity, temperature, dissolved oxygen numbers into a spreadsheet, lickety-split. The difficulty was when you went into anything somewhat specialized. So we did measurements of phytoplankton growth and herbivory, grazing on those phytoplankton. Nobody knew at the beginning how to put those things into a database and so it was total chaos trying to figure out how to normalize and characterize and catalog these things."

Similarly, a member of the research staff on the soil ecology team had experiences in using standardized data collection as part of an ecological field-based data collection effort for a government agency. As he described, the data collection processes and structures were rarely changed in that work, because the data being collected rarely changed. In his current work in a research university, however, "*data do change*" and are always evolving. Cyberinfrastructure data repositories require regularity and structure,

but for researchers involved in small-scale field-based projects, change and flexibility are essential characteristics of their research practices. Formalized and structured community data repositories that do not enable change and flexibility create challenges in reconciling and simplifying their practices to standardized submission forms. As I document in this study, describing data so that they may be understood by someone outside of one's project is not a typical task for many researchers. Using a standardized form to create such descriptions only makes the challenge more acute.

9.1.2 Implications for funding agencies

Funding agencies are heavily investing in data curation through a combination of technology and institution development, as evidenced by the projects being funded by the National Science Foundation under the DataNet initiative (NSF, 2010b). A significant question for projects of this scale is what institutional structures facilitate data and metadata management and curation in small-scale field science? The findings in my study indicate that it is unrealistic to expect data management and curation initiatives to spring up on their own within small-scale field-based scientific projects. Researchers focus their metadata activities on forms of documentation that are necessary for their own data uses. Any creation of metadata for other purposes, such as for a centralized repository, is additional work.

One way funding agencies can promote data management and curation institution building activities in small-scale field research might be by supporting research programs dedicated to the re-use of existing data. If data re-use were supported and valued in the same way as the collection of new data, researchers might have more professional

incentive to prioritize the creation of "metadata-for-data-sharing." Ironically, as of this writing, one of the National Science Foundation's most successful initiatives in this regard, the National Center for Ecological Analysis and Synthesis (NCEAS), is in the process of winding down its activities due to the loss of core NSF funding (Stokstad, 2011). The NCEAS was explicitly dedicated to only supporting ecological research that re-purposed already existing data, and was highly successful and influential in the ecological community. Because NSF funding for research centers is time-limited, NSF's support for NCEAS's funding will be withdrawn in 2012. From the perspective of cyberinfrastructure development, with its emphasis on data sharing and re-use, dropping support for successful data sharing and re-use institutions appears very short-sighted. How the ecological community responds to the weakening of a significant data sharing and re-use institution is an open topic for future research.

Another way that funding agencies can promote data management and curation at institutional levels is by increasing the profile of data management and curation in higher education. As I noted in my results, none of the researchers who participated in my study had taken, or were aware of, formal courses in data management. Such courses currently do not exist in most universities. Information-related schools increasingly are offering such courses, but targeted towards information professionals, not researchers in the sciences (or social sciences or humanities for that matter). Funding for the development of courses that cross the information and scientific disciplines could serve to introduce data management techniques and practices into disciplinary curricula. Data management

and curation workshops for undergraduate and graduate students might also bring more visibility to data and metadata activities within a variety of disciplines.

9.1.3 Implications for data curation practice

My study has a number of implications for the development of data curation practices of information institutions and professionals. First, in thinking about collection development policies and practices, what data and metadata should be curated? In Section 6.1, I discuss the importance of data processing. Should data curation institutions collect the original data, the pre-processed data, or both? And how should the decision about what data to curate, and the links between these multiple states of data be documented? These questions need to be answered on a case-by-case basis by talking with both the researchers who collected the data, and potential data users, as the data creators and users might have different views.

A directly related question is "to whom should information professionals talk to in order to assess and curate a particular data collection?" Knowledge of data and metadata are distributed amongst teams, and different members of a team will give different accounts for data and metadata processes and products. Information professionals will need to get multiple points of view when curating data, particularly when data are coming from the types of collaborative field science that I studied.

Similarly, when should information professionals determine metadata needs and practices? Information professionals, particularly archivists, have traditionally accessioned materials at the end of their use within organizational activities. With research data, on the other hand there is growing interest in making data available as

early as possible, making it necessary for information professionals with data curation responsibilities to work with data authors at an earlier stage in the data life-cycle (Borgman, et al., 2007; Cragin, et al., 2010). Researchers develop new forms of metadata as a project develops, often in response to data quality issues. How should information professionals assess the quality and integrity of data and metadata collections? The results of my study suggest that information professionals should study the timeline of a project, particularly at how metadata practices develop, as an indicator for potential data quality issues, and for the solutions to such data quality issues that researchers themselves develop over time.

Finally, my study has implications for information organization research. What metadata structures and organization systems facilitate curation of metadata products and processes? In my study, the metadata that researchers created and used in their day-to-day practices were very different than the metadata that researchers created for a centralized metadata registry. Researchers day-to-day metadata practices are not standardized, but they are often highly routinized. My study suggests that information institutions that are developing data curation initiatives should develop ways to: 1) curate informal metadata processes and products, and 2) enable flexible, multiple, and emergent viewpoints to be represented in metadata. First, informal metadata processes and products, such as email archives and individually created documents, were the primary types of metadata created and used by the researchers in my study. Is there a place for such documents alongside the formal standardized metadata forms in cyberinfrastructure data systems? Standardized metadata facilitate interoperability and consistency, but informal metadata

may facilitate more use of data by enabling users to visualize and understand the situatedness of data and metadata (Zimmerman, 2007). Second, in multi-disciplinary and collaborative research settings, such as CENS, multiple people create metadata for the same data resources. The everyday metadata products that are created and used about data in such research inevitably present multiple and different perspectives. Scientists and technologists document different parts of a project. Students, staff, and faculty in a given project have different perspectives on the origins, quality, and utility of data and metadata. Enabling all of these perspectives to be represented in metadata has the potential to make data more broadly discoverable and useful. How can cyberinfrastructure data systems enable multi-faceted perspectives on data resources? Srinivasan and Huang's (2005) work on "fluid ontologies" provides one potential method. A "fluid ontology" is one that lets "knowledge structures emerge from the interaction with the very communities that are using" the resources in information systems (pg. 194). Srinivasan and Huang illustrate how "fluid" ways of representing resources in information systems allow users to better interact with informational resources, and make meaning from those resources. By making the creation of metadata open and flexible, cyberinfrastructure data systems might benefit both scientific and public data users by allowing more discussion and participation around data, and making the accountabilities for data more apparent (Christie, 2006).

9.2 Study Limitations

The primary limitation in my study is that it is based on case studies of one particular kind of research environment: small-scale interdisciplinary field science. CENS

is a multi-faceted center, with a wide variety of projects that cross disciplinary and institutional boundaries. But with a focus on multi-disciplinary research that emphasizes technology development, CENS is a particular research environment. In addition, my sample for the CENS metadata registry test was limited in size and breadth. Further studies of the ways that researchers use of the official online version of the CENS metadata registry system are planned, and will help to solidify the grounding for the claims I make about researchers use of our prototype system.

CENS scientists and engineers have had to develop mechanisms that allow both groups to benefit from the collaboration, including by publishing papers in venues in each of their research areas. CENS teams are based within university disciplinary structures, including having conventional laboratory and office spaces, and students within these science teams have been developing dissertation topics within those university structures. I thus argue that CENS scientific research is not far outside the norm for seismology, ecology, environmental science, and aquatic biology, and likewise for the engineering collaborators. But further studies of other environments will help to develop my ideas.

9.3 Future Directions

The first direction that my study points is towards further case studies of research communities and collaboratories. Within small-scale field-based research domains, when and why is data re-use valued and encouraged? How can funding agencies promote data re-use? Researchers receive credit for work that meets disciplinary definitions for acceptable topics. Can ecology students receive degrees without collecting original data in field settings? Roth and Bowen (2001) note that conducting field work, often alone, is

a central experience in students development as ecologists. New types of data intensive science push against these kinds of established community expectations. The ways that scientists create data and metadata also emerge in relation to embedded norms.

Data management planning is another topic that my study points towards. With funding agencies such as the NSF promoting and now requiring data management planning, scientists, universities, and information institutions are all currently discussing best practices for creating and enforcing such plans. How will data management plan requirements impact metadata practices? Will researchers change their day-to-day data and metadata practices if they are forced to create a data management plan at the time of the research proposal? Possibly, but as Suchman has shown (1987; 2007), any changes will not be causal in the sense that plans cause particular actions. Instead, data management plans will allow the multiple stakeholders in this issue, researchers, funders, universities, information institutions, and the general public, to orient themselves to data management work in a different way than was possible before. Researchers will be able to be held accountable for data and metadata management by the other stakeholders, though in what sense of the term "accountable" is yet to be seen. If funders who initiated such data management planning requirements follow through with sanctionable enforcement, researchers will be accountable in both senses, they will be responsible for data management actions, and they will have to be able to provide accounts of their actions. If, on the other hand, funders do not enforce the data management planning requirement beyond the planning stage, researchers will only have to be accountable in

the second sense: being able to provide statements of what their data management approach is and will be.

The third main direction that my research points is toward this question: "When and why is metadata creation and management rewarded?" If metadata creation tasks are embedded in systems of accountability, researchers who work in environments where they are accountable for data and metadata that are used by others should have different types of metadata practices than those I observed in CENS. One way to investigate this issue is to find research settings where community data and metadata repositories are well established and widely used. How would the metadata practices of researchers who work in such a setting, such as astronomy (Hanisch, 2006) or climate science (Edwards, 2010), be different from the metadata practices of researchers in small-scale field-based projects?

APPENDIX I – SAMPLE SOIL ECOLOGY INTERVIEW PROTOCOL

Background

1. What are the main research questions for this work?
2. Do you consider these data sets part of a singular project, or separate projects?
3. Are these three kinds of data the only data you are collecting at [your field site]? The automated images, the manual images, and the sensor data?

Work and data flow

4. What are the connections between the three kinds of data? What are they used for in the analysis? How do they fit together?
5. Who has access to the data? Does that differ at different stages in the process?
6. How are the data organized at these different stages?
7. In one or two sentences, what does the term "metadata" mean to you?
8. What metadata are created for these data, i.e. are the data annotated or documented, and how does this occur? How are metadata different for images and sensor data?
9. What is the role of data in answering these research questions?
10. In what ways do these metadata, annotations, and documentation help you to answer the research questions we discussed earlier?
11. What is the long-term plan for these data? Will they be kept locally, submitted to a data repository, discarded…
12. What is the long-term plan for the metadata (calibration files and notebooks)?

Online and Informal documentation:

13. How do you coordinate research activities within the group? Do you have email lists or regular meetings? Who is part of the meetings?
14. How do you share data and/or metadata?
15. Is there anyone, either within your team or outside it, who you go to for expertise about data or metadata management for your projects?
16. Is there anyone, either within your team or outside it, who you recommend to students for data or metadata management? Are there particular courses available to students or other departments you might suggest they check out?
17. Caroline and I talked about her analysis of the soil images in some depth. What was your goal for her work with those images, and what do you think of her results so far?
18. How much input have you had into her analysis process?
19. How will her processes scale as you deploy more automated imaging systems?
20. Now that you have the automated soil imaging system, why do you also still collect images using the manual imaging system?
21. NEON is planning to install automated imaging systems at their sites. Have you had much involvement with NEON's planning?
22. Have you talked with people at NEON about the challenges you've had in analyzing the automated imaging systems images?

Conclude

23. Thank you for your time, is there anything else you want to mention? Who else should I talk to?

APPENDIX II – SAMPLE SEISMIC INTERVIEW PROTOCOL

Background

1. What are the main research questions for this project? Both the project as a whole, and for your research specifically.
2. What has been your role in this project?
3. In one or two sentences, what does the term "metadata" mean to you?

Work and data flow

4. What data are being collected on this seismic deployment?
5. Could you describe the path of the data, from collection to storage?
6. Who has access to the data? Does that differ at different stages in the process?
7. How are the data organized at these different stages?
8. What do you need to know about the data collection process in order to understand and use the data?
9. What formats are data collected and/or stored in? How do conversions between formats occur?
10. Could you tell me a little bit about these formats? Why use SEED vs. SAC data format? Why use Mini-SEED instead of full SEED?
11. How did you learn about the data formats – SEED and SAC?
12. In what ways are the data annotated or documented, and how does this occur?
13. Are decisions like "use magnetic north to orient the sensors instead of true north" documented somewhere?
14. What is the role of data in answering these research questions?

15. In what ways do these metadata, annotations, and documentation help you to answer the research questions we discussed earlier?

Software

16. Do you do any software development for this project? Maintenance?
17. Who has access to the data collection software, i.e. scripts? Does that differ at different stages in the process?
18. In what ways are the data collection software (scripts, etc.) annotated or documented, and how does this occur?
19. How much of the software, data file structures, etc. were carried over from the Mexico deployment? Is that ever a problem?

Online and Informal documentation:

20. What kinds of notes do you keep of your field activities? If so, how?
21. How useful is the CENS seismic email list? How do you use the email list?
22. [A team member] made a webpage that shows the status of the deployment sites and network health. Has that been useful for you? How so?

Long-term plans

23. What is the long-term plan for these data? Will they be kept locally, submitted to a data repository, discarded…
24. What is the process for submitting data to IRIS? What documentation is included?

Conclude

25. Thank you for your time, is there anything else you want to mention? Who else should I talk to?

APPENDIX III – CENS METADATA REGISTRY USER TEST PROTOCOL

- Welcome - General statement of purpose and expression of appreciation for the test user's participation
    - I'm going to read from this sheet to give you the introduction, first so I don't forget anything, and second because we want to give the same introduction to every tester, which helps to give us more reliable feedback. This is a test of our CENS Metadata Registry. NSF has asked that CENS report on data sets that have been collected as a part of CENS research. This includes data sets created or contributed to as part of the research being reported on during the yearly reporting cycle. To accomplish this, we are asking CENS researchers to fill out the form you are testing today as part of a CENS Annual Report system. Our intention is to also post these data descriptions on the CENS website to facilitate data discovery by interested users. Our goal is to facilitate the collection of well described and interconnected data that can be easily used and re-used by the researchers who collected them, and potentially by people outside CENS. Thank you for helping.
- Give permission form for tester to sign
- Introduction
    - We will ask you to test our data registration form as well as ask you a couple of questions about your experience. We will give you X tasks to complete and we will be observing and taking notes while you complete

the tasks. Some parts of the form may go very quickly, while others may take a bit longer. **Please be assured that we are testing our application and not your performance**. Our goal right now is to improve the form based on the test. Any feedback you can give us will help us. Before we start, I just have a few questions about your use of the CENS online Annual Report system.

- Pre-test questionnaire
    - Did you use the annual report system this past year?
    - If so, what did you think of it?
    - Did you fill out the data form? Why or why not?
        - Areas to probe: who has rights, who is responsible, are shared datasets more problematic?
    - If not, who created the report for your project? Did you interact with them? Why didn't you fill it out yourself?
- Training
    - We want to test how easy the form is to understand and fill out. We will have you test the web interface we have created that allows you to create metadata for your data sets. We will not help you use the application because we want to see how you would use it if you were by yourself. We want to mimic the real-world environment as much as possible. Keep in mind that this is a prototype, so things are a little rough around the edges.

At the end of the test, we will give you the output (your responses) so you can have it for the upcoming reporting cycle.

  - Now, I will give you the X tasks one at a time to complete using our form. We will be taking notes as you perform the tasks. I would like you to think aloud and tell me the thoughts that are going through your mind as you complete these tasks. I know it's not natural, but if you could tell us what you are thinking: why you clicked on a button, what decisions you are making, your reasoning, that will help a lot in understanding how you use the form. Ok? Any questions so far?

- Give test tasks
  1. Choose a data set that you used this past year as a part of [name of report from 2010 CENS Annual Report].
  2. Go to: [URL for the CENS Metadata Registry prototype].
  3. Using this interface, create a description for your chosen data set.
  4. When you are finished, click “Submit.”
- Data collection
  - Log errors and observations for each test using prepared data collection forms.
- Post-test interview
  - Why did you choose this particular data? Are there other data you could have chosen to describe instead?
  - What field(s) were most difficult for you to fill out? Why?

- Which fields were most useful for describing your data? Least useful?
- What's missing from this form that would help you better describe your data?
- How can we improve the process/form?
- What are the benefits for you in registering your data?
- Would the system be improved if we allowed multiple people to add information about your data (i.e. multiple people from the same project)?
    - If so, why? Why would this be useful?
    - If so, who else in your project would you want to contribute to your data record? What would you want them to contribute/add?

APPENDIX IV – DOCUMENTS REFERENCED DURING USER TESTS

Successful References:

Documents testers referenced successfully during their tests of the CENS Metadata Registry, and why they performed the reference.

- Looked at names of data file folders for approximate dates of data collection (Tester 6 - Aquatic)
- Looked at software code to remember the data processing steps (Tester 2 - Seismology)
- Copied and pasted a data file name as the title of the data set (Tester 9 - other CENS)
- Opened a data file to look at the dates (Tester 9 - other CENS)
- Opened up a file that had the list of columns for the variables field. He copied it and pasted it out of that file. (Tester 9 - other CENS)
- Opened up a publication he authored, did a keyword search for a particular word, and then used a passage in the text to write up the description of the equipment used in the data collection process. Did not copy and paste. (Tester 1 - other CENS)
- Used a paper print-out of a database record to fill out the variables field (Tester 1 - other CENS)
- Pulled up papers from personal webpage to find specific names of pieces of equipment (Tester 7 - other CENS)

- Went to a data web site to find details for the "Variables" field (Tester 2 - Seismology)
- Looked up a web site for the "Data collection location" field, probably to look up a correct spelling (Tester 7 - other CENS)
- Looked at published papers for the funding field (Tester 7 - other CENS)
- Opened a personal publication and looked at acknowledgements section to find the funding information (Tester 1 - other CENS)
- Opened a publication off of personal web page, then copied and pasted funding information (Tester 5 - Aquatic)
- Went to the lab web site to find the URL for their data that is displayed online (Tester 8 - Aquatic)
- Opened personal webpage to find URL for the data posted there (Tester 7 - other CENS)
- Opened a personal database log-in page to find the URL (Tester 1 - other CENS)
- Went to data web site to find the URL (Tester 9 - other CENS)
- Went to lab web site to find URL for data (Tester 11 - Ecology)
- Went to the web site of a researcher at another university to show us how the other researcher described his data sharing permission conditions (Tester 11 - Ecology)

Unsuccessful References:

Documents testers referenced during their tests of the CENS Metadata Registry, but were not able to find the information they were looking for, and why they performed the reference.

- Looked in his file directories for grant information, but did not find a file that he opened (Tester 6 - Aquatic)
- Looked in some files for sensor lists, but did not find what he wanted (Tester 11 - Ecology)
- Opened lab web site to show us how their team accesses their data internally, but did so to show us how he did not have a URL that he could provide to the public (Tester 5 - Aquatic)

APPENDIX V – MY RESPONSES TO THE CENS METADATA REGISTRY FORM

Note: The field names are in **bold** *and my responses are in* normal font.

Title: CENS Metadata Interviews

Dates of collection: Start Date: 06-01-2010 End Date: 07-01-2011

Data collection site: CENS universities

Contributors:

Name	**Email**	**Primary**
Matthew Mayernik	mattmayernik@ucla.edu	☑
Jillian Wallis	[removed]	☐
Christine Borgman	[removed]	☐

Data Type: Image, Moving Image, Sound, Text

Research question/why the data was collected: I am studying how CENS researchers document their data for their own use, and for sharing with others. My questions are: How and where are metadata created, by whom, and for what purpose? How are metadata creation tasks learned and parceled out in research groups? How do local metadata practices translate to the creation of metadata for shared community repositories?

Variables Collected: Collected semi-structured interviews about researchers own data documentation practices, and conducted user-tests of the CENS metadata reporting interface as part of the development of the online CENS annual report system.

Process and equipment used for collection and/or analysis: Interviews were audio-recorded. The audio was then transcribed. We also took pictures of lab and field settings. We also did audio and video recordings of the CENS metadata reporting interface user tests.

Data File Format: doc, mp3, wma,

Would you be willing to share your data?: No

Data Sharing Permission Level: Other data permission conditions: Data release is prohibited by IRB restrictions.

Funding Source/s: CENS

Keywords: scientific data, metadata, user test, ethnography, data practices

Location of Data (URL): N/A

REFERENCES

Note: all URLs last visited on June 8, 2011.

Agre, P.E. (1990). Plans and situated actions: the problem of human-machine communication: Lucy A. Suchman [Book review]. *Artificial Intelligence*, 43(3): 369-384. http://dx.doi.org/10.1016/0004-3702(90)90078-E

Agre, P.E. (1997). Computation and human experience. New York: Cambridge University Press.

Agre, P.E. (1998). Yesterday's tomorrow. *Times Literary Supplement*, 3 July, pp. 3-4. http://polaris.gseis.ucla.edu/pagre/tls.html

Ahern, T.K. (2002). The FDSN and IRIS data management system: providing easy access to terabytes of information. In W.H.K. Lee, et al. (Eds.), *International handbook of earthquake and engineering seismology* (pp. 1645-1656). Boston, MA: Academic Press for International Association of Seismology and Physics of the Earth's Interior.

Alasuutari, P. (1995). *Researching culture: qualitative method and cultural studies*. London, England: Sage.

Anglo-American Cataloguing Rules (AACR), 2nd ed. (2005 update). Chicago: American Library Association.

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: 191-201. http://www.jstage.jst.go.jp/article/dsj/3/0/191/_pdf

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., et al. (2004). Science and government: an international framework to promote access to data. *Science,* 303(5665): 1777-1778.

Anderson, C.R., Brzezinski, M.A., Washburn, L., & Kudela, R. (2006). Circulation and environmental conditions during a toxigenic Pseudo-nitzschia australis bloom in the Santa Barbara Channel, California. *Marine Ecology Progress Series*, 327: 119–133.

Atkins, D., et al. (2003). *National Science Foundation Blue-Ribbon Advisory Panel on Cyber-infrastructure, Revolutionizing Science and Engineering through Cyber-infrastructure*. http://www.communitytechnology.org/nsf_ci_report/

Baccus, M.D. (1986). Multiple truck wheel accidents and their regulations. In H. Garfinkel (Ed.) *Ethnomethodological studies of work*. New York: Routledge.

Barnett, V. & Lewis, T. (1994). *Outliers in statistical data, 3rd edition*. Chichester, New York: John Wiley and Sons.

Bearman, D. & Trant, J. (1998). Authenticity of digital resources: towards a statement of requirements in the research process. *D-Lib Magazine*, 4(6). http://www.dlib.org/dlib/june98/06bearman.html

Bell, G., Hey, T., & Szalay, A. (2009). Computer science: beyond the data deluge. *Science,* 323(5919): 1297-1298.

Berman, F. (2008). Got data?: a guide to data preservation in the information age. *Communications of the ACM,* 51(12): 50-56. http://doi.acm.org/10.1145/1409360.1409376

Birnholtz, J.P. & Bietz, M.J. (2003). Data at work: supporting sharing in science and engineering. In M. Tremaine (Ed.), *GROUP '03. Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work* (pp. 330-348). ACM Press.

Blanchette, J.F. (2011). A material history of bits. *Journal of the American Society for Information Science and Technology*, 62(6): 1042-1057.

Borgman, C.L. (2002). Challenges in building digital libraries for the 21(st) century. In E.P. Lim, S. Foo, C. Khoo, S. Urs, T. Costantino, E. Fox & H. Chen (Eds.), *Digital Libraries: People, Knowledge, and Technology* (pp. 1-13). Springer-Verlag Berlin.

Borgman, C.L. (2007). *Scholarship in the digital age: information, infrastructure, and the internet*. Cambridge, MA: MIT Press.

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(1): 17-30.

Borgman, C.L., Wallis, J.C., Mayernik, M.S., & Pepe, A. (2007). Drowning in data: digital library architecture to support scientists' use of embedded sensor networks. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM.

Bos, N., Zimmerman, A., Olson, J., Yew, J., Yerkie, J., Dahl, E., & Olson, G. (2007). From shared databases to communities of practice: a taxonomy of collaboratories. *Journal of Computer-Mediated Communication*, 12(2): 652–672.

Bowker, G.C. (2000). Biodiversity datadiversity. *Social Studies of Science,* 30(5): 643-683.

Bowker, G.C. & Star, S.L. (1996). How things (actor-net)work: classification, magic and the ubiquity of standards. http://epl.scu.edu:16080/~gbowker/actnet.html

Bowker, G.C. & Star, S.L. (2000). *Sorting things out: classification and its consequences*. Cambridge, MA: MIT Press.

Buetow, K.H. (2005). Cyberinfrastructure: empowering a "third way" in biomedical research. *Science,* 308(5723): 821-824.

Burnett, K., Ng, K.B., & Park, S. (1999). A comparison of the two traditions of metadata development. *Journal of the American Society for Information Science*, 50(13): 1209-1217.

Button, G. & Sharrock, W. (1998). The organizational accountability of technological work. *Social Studies of Science,* 28(1): 73-102.

Bychkovskiy, V., Megerian, S., Estrin, D., & Potkonjak, M. (2003). A collaborative approach to in-place sensor calibration. In F. Zhao and L. Guibas (Eds.), *Proceedings of the 2nd international conference on Information Processing in Sensor Networks, IPSN '03* (pp. 301-316), Berlin: Springer-Verlag.

Callon, M. (1986). Some elements of a sociology of translation: domestication of the scallops and the fishermen of St. Brieuc Bay. In J. Law (Ed.), *Power, action, and belief: A new sociology of knowledge?* (pp. 196-233). London: Routledge and Kegan Paul.

Campbell, E.G., Clarridge, B.R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N.A. & Blumenthal, D. (2002). Data withholding in academic genetics: evidence from a national survey. *Journal of the American Medical Association*, 287(4): 473-480. http://jama.ama-assn.org/cgi/content/full/287/4/473

Campbell, D.G. (2005). Metadata, metaphor, and metonymy. *Cataloging & Classification Quarterly*, 40(3/4): 57-73.

Caplan, P. (2003). *Metadata fundamentals for all librarians*. Chicago, IL: American Library Association.

Cerwonka, A. & Malkki, L.H. (2007). *Improvising theory: process and temporality in ethnographic fieldwork.* Chicago, IL: University of Chicago Press.

Chang, K., Yau, N., Hansen, M., & Estrin, D. (2006). SensorBase.org - a centralized repository to slog sensor network data. *Proceedings of the International Conf. on Distributed Networks(DCOSS)/EAWMS.*

Chin, G. & Lansing, C.S. (2004). Capturing and supporting contexts for scientific data sharing via the biological sciences collaboratory. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work* (pp. 409-418). New York: ACM. http://doi.acm.org/10.1145/1031607.1031677

Choudhury, G.S. (2008). Case study in data curation at Johns Hopkins University. *Library Trends*, 57(2): 211-220.

Christie, M. (2006). Boundaries and accountabilities in computer-assisted ethnobotany. *Research and Practice in Technology Enhanced Learning*, 1(3): 285–296. http://dx.doi.org/10.1142/S1793206806000214

Clarke, A.E. (2005). *Situational analysis: grounded theory after the postmodern turn*. Thousand Oaks, CA: Sage Publications.

Collins, W.R., Miller, K.W., Spielman, B.J., & Wherry, P. (1994). How good is good enough?: an ethical analysis of software construction and use. *Communications of the ACM*, 37(1): 81-91.

Coyle, K. (2010). Library data in a modern context. *Library Technology Reports*, 46(1): 5-13.

Crabtree, A., Nichols, D.M., O'Brien, J., Rouncefield, M., & Twidale, M.B. (2000). Ethnomethodologically informed ethnography and information system design, *Journal of the American Society for Information Science*, 51(7): 666-682.

Cragin, M.H., Palmer, C.L., Carlson, J.R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1926), 4023-4038. http://dx.doi.org/doi:10.1098/rsta.2010.0165

Cragin, M.H. & Shankar, K. (2006). Scientific data collections and distributed collective practice. *Computer Supported Cooperative Work,* 15: 185–204.

Creative Commons. (2011). *About the licenses*. http://creativecommons.org/licenses/

Cronin, B. (2008). The sociological turn in information science. *Journal of Information Science.* 34(4): 465-475. http://dx.doi.org/10.1177/0165551508088944

Crystal, A. & Greenberg, J. (2005). Usability of a metadata creation application for resource authors. *Library & Information Science Research*, 27(2): 177-189.

Currier, S., Barton, J., O'Beirne, R., & Ryan, B. (2004). Quality assurance for digital learning object repositories: issues for the metadata creation process. *Research in Learning Technology*, 12(1): 5-20. http://dx.doi.org/10.1080/0968776042000211494

Cyberinfrastructure Vision for 21st Century Discovery. (2007). Washington, D.C.: National Science Foundation. http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf

Day, R.E. (2005). Clearing up "implicit knowledge": implications for knowledge management, information science, psychology, and social epistemology. *Journal of the American Society for Information Science and Technology*, 56(6): 630-635. http://dx.doi.org/10.1002/asi.20153

Day, R.E. (2009). Information explosion. In M.J. Bates & M. Niles Maack (Eds.), *Encyclopedia of Library and Information Sciences, Third Edition* (pp. 2416-2420).

Delserone, L.M. (2008). At the watershed: preparing for research data management and stewardship at the University of Minnesota Libraries. *Library Trends*, 57(2): 202-210.

Dervin, B. (1992). From the mind's eye of the user: the sense-making qualitative/quantitative methodology. In J.D. Glazier and R.R. Powell (Eds.), *Qualitative research in information management* (pp. 61–84). Englewood, CO: Libraries Unlimited.

Doerr, M., Hunter, J., & Lagoze, C. (2003). Towards a core ontology for information integration. *Journal of Digital information*, 4(1). http://journals.tdl.org/jodi/article/view/92/91

Dodd, S.A. (1982). *Cataloging machine-readable data files: an interpretive manual*. Chicago, IL: American Library Association.

Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing*, 8(1): 19-30. http://dx.doi.org/10.1007/s00779-003-0253-8

Dourish, P. & Button, G. (1998). On "Technomethodology": foundational relationships between ethnomethodology and system design. *Human Computer Interaction,* 13(5): 395-432.

Dublin Core Metadata Initiative. (2009). *Dublin Core Metadata Element Set, Version 1.1.* http://dublincore.org/documents/dces/

Dublin Core Metadata Initiative. (2010). *Dublin Core type vocabulary.* http://www.dublincore.org/documents/dcmi-type-vocabulary/

Duerr, R., Parsons, M., Marquis, M., Dichtl, R., & Mullins, T. (2004). Challenges in long-term data stewardship. *Proceedings of the 21st IEEE Conference on Mass Storage Systems and Technologies* (pp. 47-67). NASA/CP-2004-212750.

Edwards, P.N. (2010). *A vast machine: computer models, climate data, and the politics of global warming*. Cambridge, MA: MIT Press.

Edwards, P.N., Jackson, S.J., Bowker, G.C. & Knobel, C.P. (2007). *Understanding infrastructure: dynamics, tensions, and design*. Final report of the workshop, "History and Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures." http://hdl.handle.net/2027.42/49353.

Edwards, P.N., Mayernik, M.S., Batcheller, A., Borgman, C.L., & Bowker, G.C. (in press). Science friction: data, metadata, and collaboration in the interdisciplinary sciences. *Social Studies of Science*.

Ellison, A.M., et al. (2006). Analytic webs support the synthesis of ecological data sets. *Ecology*, 87(6): 1345–1358.

Emerson, R.M., Fretz, R.I., & Shaw, L.L. (1995). *Writing ethnographic fieldnotes*. Chicago, IL: University of Chicago Press.

Estrin, D., Michener, W., Bonito, G., and the workshop participants. (2003). *Environmental Cyberinfrastructure Needs for Distributed Sensor Networks: A Report from a National Science Foundation Sponsored Workshop*. Scripps Institution of Oceanography, La Jolla, CA. 12-14 August 2003. http://intranet2.lternet.edu/sites/intranet2.lternet.edu/files/documents/Scientific_Reports/Cyberinfrastructure/cyberRforWeb.pdf

Farkas-Conn, I.S. (1990). *From documentation to information science: the beginnings and early development of the American Documentation Institute-American Society for Information Science*. New York: Greenwood Press.

Fienberg, S.E., Martin, M.E., & Straf, M.L. (Eds.). (1985). *Sharing research data*. Washington, DC: National Academy Press.

Finholt, T.A. (2002). Collaboratories. *Annual Review of Information Science and Technology*, 36: 73-107.

Friedlander, A. (2008). The triple helix: cyberinfrastructure, scholarly communication, and trust. *The Journal of Electronic Publishing*, 11(1). http://dx.doi.org/10.3998/3336451.0011.109

Frohmann, Bernd. (2004). *Deflating information: from science studies to documentation*. Buffalo, NY: University of Toronto Press.

Fujimura, J.H. (1987). Constructing 'do-able' problems in cancer research: articulating alignment. *Social Studies of Science*, 17(2): 257-293.

Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.

Garfinkel, H. (1974). The origins of the term 'ethnomethodology'. In Roy Turner (Ed.), *Ethnomethodology: selected readings* (pp. 15-18). Baltimore, MD: Penguin Education.

Garfinkel, H., Lynch, M., & Livingston, E. (1981). The work of a discovering science construed with materials from the optically discovered pulsar. *Philosophy of the Social Sciences,* 11(2): 131-58.

Gershon, N.D. & Miller, C.G. (1993). Dealing with the data deluge. *IEEE Spectrum,* 30(7): 28-32.

Gilliland, A.J. (2008). Setting the stage. In Murtha Baca (Ed.), *Introduction to Metadata: Pathways to Digital Information, Online Edition, Version 3.0*. Los Angeles, CA: Getty Information Institute. http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.html

Glaser, B.G. & Strauss, A.L. (1967). *The discovery of grounded theory: strategies for qualitative research.* Chicago, IL: Aldine.

Glibert, P.M., Anderson, D.M., Gentien, P., Graneli, E., & Sellner, K.G. (2005). The global, complex phenomena of harmful algal blooms. *Oceanography*, 18(2): 137–147.

Goble, C. & De Roure, D. (2009). The impact of workflow tools on data-intensive research. In T. Hey, S. Tansley, & K. Tolle (Eds.), *The Fourth Paradigm: Data-Intensive Scientific Discovery* (pp. 137-146). Redmond, WA: Microsoft. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part3_goble_deroure.pdf

Gorman, M. (2006). Metadata dreaming. *The Serials Librarian*, 51(2): 47-54.

Gray, J., Liu, D.T., Nieto-Santisteban, M., Szalay, A., DeWitt, D., & Heber, G. (2005). Scientific data management in the coming decade. *CTWatch Quarterly*, 1(1). http://www.ctwatch.org/quarterly/articles/2005/02/scientific-data-management/

Greenberg, J. (2001). A quantitative categorical analysis of metadata elements in image applicable metadata schemas. *Journal of the American Society for Information Science and Technology,* 52(11): 917-914.

Greenberg, J. (2004). Metadata extraction and harvesting: a comparison of two automatic metadata generation applications. *Journal of Internet Cataloging*, 6(4): 59-82. http://dx.doi.org/10.1300/J141v06n04_05

Greenberg, J. (2005). Understanding metadata and metadata schemes. *Cataloging & Classification Quarterly*, 40(3): 17-36. http://dx.doi.org/doi:10.1300/J104v40n03_02

Greenberg, J. (2009). Metadata and digital information. In M.J. Bates & M. Niles Maack (Eds.), *Encyclopedia of Library and Information Sciences, Third Edition* (pp. 3610-3623).

Greenberg, J., Crystal, A., Robertson, W.D., & Leadem, E. (2003). Iterative design of metadata creation tools for resource authors. In *2003 Dublin Core Conference: Supporting Communities of Discourse and Practice – Metadata Research and Applications. DC-2003: Proceedings of the International DCMI Conference and Workshop* (pp. 49-58). Syracuse, NY: Information Institute of Syracuse.

Greenberg, J., Pattuelli, M.C., Parsia, B., & Robertson, W.D. (2002). Author-generated Dublin Core metadata for web resources: a baseline study in an organization. *Journal of Digital Information*, 2(2). https://journals.tdl.org/jodi/article/view/42

Greenberg, J. & Robertson, W.D. (2002). Semantic web construction: an inquiry of authors' views on collaborative metadata generation. In *Proceedings of the International Conference on Dublin Core and Metadata for e-Communities 2002* (pp. 45-52). Florence, Italy.

Griffiths, A. (2008). The publication of research data: researcher attitudes and behaviour. *The International Journal of Digital Curation*, 4(1): 46-56.

Guarino, N. (1998). Formal ontology and information systems. In N. Guarino (Ed.), *Formal ontology in information systems: proceedings of the first international conference (FOIS '98)* (pp. 3-15). Amsterdam, Netherlands: IOS Press.

Hanisch, R.J. (2006). Data standards for the international virtual observatory. *Data Science Journal*, 5: 168-173. http://www.jstage.jst.go.jp/arti\cle/dsj/5/0/168/_pdf

Hanson, B., Sugden, A., & Alberts, B. (2011). Making data maximally available. *Science*, 331(6018): 649. http://dx.doi.org/10.1126/science.1203354

Helly, J., Staudigel, H., & Koppers, A. (2003). Scalable models of data sharing in earth sciences. *Geochemistry Geophysics Geosystems,* 4(1). http://www.beamreach.org/research/data_sharing_model_GC2002.pdf

Heritage, J. (1984). *Garfinkel and ethnomethodology*. Cambridge, MA: Polity Press.

Hey, T. & Trefethen, A. E. (2003). The data deluge: an eScience perspective. In F. Berman, G. Fox, T. Hey (Eds.), *Grid Computing: Making the Global Infrastructure a Reality* (pp. 809-824), New York: Wiley

Hey, T. & Trefethen, A. E. (2005). Cyberinfrastructure for e-science. *Science,* 308(5723): 817-821.

Hofweber, T. (2009). Logic and ontology. In E.N. Zalta (ed.) *The Stanford Encyclopedia* of *Philosophy (Spring 2009 Edition)*. http://plato.stanford.edu/entries/logic-ontology/

Howarth, L.C. (2005). Metadata and bibliographic control: soul-mates or two solitudes? *Cataloging & Classification Quarterly*, 40(3): 37-56. http://www.informaworld.com/10.1300/J104v40n03_03

Hutchins, E. (1996). *Cognition in the wild*. Cambridge, MA: MIT Press.

Intner, S.S., Lazinger, S.S., & Weihs, J. (2006). *Metadata and its impact on libraries*. Westport, CT: Libraries Unlimited.

Incorporated Research Institutions for Seismology (IRIS). (2010). *SEED reference manual: standard for the exchange of earthquake data*. SEED format version 2.4. IRIS. http://www.iris.edu/manuals/SEEDManual_V2.4.pdf

Incorporated Research Institutions for Seismology (IRIS). (2011). http://www.iris.edu/

Jones, M.B., Schildhauer, M.P., Reichman, O.J., & Bowers, S. (2006). The new bioinformatics: integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics*, 37: 519-544.

Jordan, B. & Henderson, A. (1995). Interaction analysis: foundations and practice. *The Journal of the Learning Sciences,* 4(1): 39-103.

Karasti, H. & Baker, K.S. (2008). Digital data practices and the long term ecological research program growing global. *The International Journal of Digital Curation*, 3(2). http://www.ijdc.net/index.php/ijdc/article/viewFile/86/57

Karasti, H., Baker, K., & Halkola, E. (2006). Enriching the notion of data curation in e-science: data managing and information infrastructuring in the long term ecological

research (LTER) network. *Computer Supported Cooperative Work*, 15(4): 321-358. http://dx.doi.org/10.1007/s10606-006-9023-2.

Kelty, C.M. (2008). *Two bits: the cultural significance of free software*. Durham, NC: Duke University Press.

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, 331(6018): 719-721.

Kohler, R.E. (2002). *Landscapes and labscapes: exploring the lab-field border in biology*. Chicago: University of Chicago Press.

Kuhn, T.S. (1962, 1970). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Lagoze, C., Lynch, C.A., & Daniel, R. (1996). The Warwick Framework: a container architecture for aggregating sets of metadata. *D-Lib Magazine*, 2(7). http://www.dlib.org/dlib/july96/lagoze/07lagoze.html

Lamb, R. & Davidson, E. (2005). Information and communication technology challenges to scientific professional identity. *The Information Society*, 21(1): 1-24.

Latour, B. (1987). *Science in action*. Cambridge, MA: Harvard University Press.

Latour, B. (1999). *Pandora's hope: essays on the reality of science studies.* Cambridge, MA: Harvard University Press.

Latour, B. (2005). *Reassembling the social: an introduction to actor-network-theory*. New York: Oxford University Press.

Latour, B. & Woolgar, S. (1979). *Laboratory life.* Princeton, NJ: Princeton University Press.

Lave, J. & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. New York: Cambridge University Press.

Law, J. (2004). *After method: mess in social science research*. New York: Routledge.

Lawrence, B.N., Lowry, R., Miller, P., Snaith, H., & Woolf, A. (2009). Information in environmental data grids. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1890): 1003-1014.

Lofland, J., Snow, D., Anderson, L., & Lofland, L.H. (2006). *Analyzing social settings: a guide to qualitative observation and analysis.* Belmont, CA: Wadsworth/Thomson Learning.

Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. (2005). Washington, D.C.: National Science Foundation, National Science Board. http://www.nsf.gov/pubs/2005/nsb0540/

Lord, P. & MacDonald, A. (2003). *Data curation for e-science in the UK: an audit to establish requirements for future curation and provision.* JISC Committee for the Support of Research. (JCSR).

Luttrell, W. (2000). "Good enough" methods for ethnographic research. *Harvard Educational Review,* 70(4): 499-523.

Lynch, C. (2008). Big data: how do your data grow? *Nature*, 455(7209): 28-29. http://dx.doi.org/10.1038/455028a

Lynch, M. (1993). *Scientific practice and ordinary action: ethnomethodology and social studies of science*. New York: Cambridge University Press.

Maddox, J. (1988). Finding wood among the trees. *Nature*, 333(6168): 11.

Marcus, G.E. (1995). Ethnography in/of the world system: the emergence of multi-sited ethnography. *Annual Review of Anthropology,* 24(1): 95-117.

Mayernik, M.S. (2010). Metadata realities for cyberinfrastructure: data authors as metadata creators. *iConference 2010 Proceedings* (pp. 148-153).

Mayernik, M.S., Batcheller, A.L., & Borgman, C.L. (2011). How institutional factors influence the creation of scientific metadata. In *Proceedings of the 2011 iConference* (pp. 417-425). New York, NY: ACM.

Mayhew, J. (1999). *The usability engineering lifecycle*. San Francisco, CA: Morgan Kaufmann.

McDonough, J. (2006). METS: standardized encoding for digital library objects. *International Journal on Digital Libraries*, 6(2): 148-158. http://dx.doi.org/10.1007/s00799-005-0132-1

Merton, R.K. (1973). *The sociology of science: theoretical and empirical investigations* (N.W. Storer, Ed.). Chicago, IL: University of Chicago Press.

Metadata Encoding and Transmission Standard (METS). (2009). Official website. http://www.loc.gov/standards/mets/

Michener, W.K. (2006). Meta-information concepts for ecological data management. *Ecological Informatics*, 1(1): 3-7.

Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., & Stafford, S.G. (1997). Nongeospatial metadata for the ecological sciences. *Ecological Applications*, 7(1): 330-342.

Miles, S., Groth, P., Branco, M., & Moreau, L. (2007). The requirements of using provenance in e-science experiments. *Journal of Grid Computing*, 5(1): 1-25.

Millerand, F. & Bowker, G.C. (2009). Metadata standards: trajectories and enactment in the life of an ontology. In M. Lampland & S.L. Star (Eds.), *Standards and Their Stories* (pp. 149-165). Ithaca, NY: Cornell University Press.

National Science Foundation (NSF). (2010a). *Scientists seeking NSF funding will soon be required to submit data management plans*. http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928

National Science Foundation (NSF). (2010b). *Sustainable digital data preservation and access network partners (DataNet)*. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141

National Science Foundation (NSF). (2011a). *Chapter II - proposal preparation instructions: special information and supplementary documentation*. http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#IIC2j

National Science Foundation (NSF). (2011b). *Dissemination and sharing of research results*. http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Newman, H.B., Ellisman, M.H., & Orcutt, J.A. (2003). Data-intensive e-science frontier research. *Communications of the ACM,* 46(11): 68-77. http://doi.acm.org/10.1145/948383.948411

Ni, K., Ramanathan, N., Chehade, M.N., Balzano, L., Nair, S., Zahedi, S., Kohler, E., Pottie, G., Hansen, M., & Srivastava, M. (2009). Sensor network data fault types. *ACM Transactions on Sensor Networks,* 5(3): 1-29. http://doi.acm.org/10.1145/1525856.1525863

National Information Standards Organization (NISO). (2004). *Understanding metadata*. Bethesda, MD: NISO Press. http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

Orr, J.E. (1996). *Talking about machines: and ethnography of a modern job.* Ithaca, NY: Cornell University Press.

Palmer, C.L., Heidorn, P.B., Wright, D., & Cragin, M.H. (2007). Graduate curriculum for biological information specialists: a key to integration of scale in biology. *The International Journal of Digital Curation*, 2(2). http://www.ijdc.net/index.php/ijdc/article/viewFile/42/27

Parsons, T.R., Maita, Y., & Lalli, C.M. (1984). *A manual of chemical and biological methods for seawater analysis.* New York: Pergamon Press.

Pepe, A., Borgman, C.L., Wallis, J.C., & Mayernik, M.S. (2007). Knitting a fabric of sensor data and literature. In *Information Processing in Sensor Networks*, *2007*. Cambridge, MA: ACM/IEEE.

Pickering, A. (1995). *The mangle of practice: time, agency, & science*. Chicago, IL: University of Chicago Press.

Pollner, M. (1974). Sociological and common-sense models of the labeling process. In Roy Turner (ed.) *Ethnomethodology: selected readings* (pp. 27-40). Baltimore, MD: Penguin Education.

Price, D.J.d.S. (1965). *Little science, big science*. New York: Columbia University Press.

Pryor, G. & Donnelly, M. (2009). Skilling up to do data: whose role, whose responsibility, whose career? *The International Journal of Digital Curation*, 4(2). http://www.ijdc.net/index.php/ijdc/article/viewFile/126/133

Rajasegarar, S., Leckie, C., Palaniswami, M., & Bezdek, J.C. (2006). Distributed anomaly detection in wireless sensor networks. In *10th IEEE Singapore International Conference on Communication Systems, 2006* (pp. 1-5). IEEE.

Randall, D., Harper, R., & Rouncefield, M. (2007). *Fieldwork for design: theory and practice*. London: Springer.

Ribes, D. (2006). *Universal informatics: building cyberinfrastructure, interoperating the geosciences.* Ph.D. dissertation, University of California, San Diego.

Richards, P.S. (1994). *Scientific information in wartime: the Allied-German rivalry, 1939-1945*. Westport, CT: Greenwood Press.

Riding the Wave: How Europe Can Gain From the Rising Tide of Scientific Data. (2010). Final report of the High Level Expert Group on Scientific Data. A submission to the European Commission. European Union. http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707

Roth, W.M. (2009). Radical uncertainty in scientific discovery work. *Science, Technology, & Human Values,* 34(3): 313-336.

Roth, W.M. & Bowen, G.M. (2001). Of disciplined minds and disciplined bodies: on becoming an ecologist. *Qualitative Sociology*, 24(4): 459-481. http://www.springerlink.com/content/g51113n07j8277j7/fulltext.pdf

Savolainen, R. (1993). The sense-making theory: reviewing the interests of a user-centered approach to information seeking and use. *Information Processing & Management*, 29(1): 13-28.

Schmidt, K. & Simone, C. (1996). Coordination mechanisms: towards a conceptual foundation of CSCW systems design. *Computer Supported Cooperative Work*, 5(2): 155-200.

Searle, J.R. (2006). Social ontology. *Anthropological Theory*, 6(1): 12-29. http://dx.doi.org/doi:10.1177/1463499606061731

Shankar, K. (2002). *Scientists, records, and the practical politics of infrastructure.* Ph.D. dissertation, University of California, Los Angeles.

Shankar, K. (2007). Order from chaos: the poetics and pragmatics of scientific recordkeeping. *Journal of the American Society for Information Science and Technology*, 58(10): 1457-1466. http://dx.doi.org/doi:10.1002/asi.20625

Shankar, K. (2009). Ambiguity and legitimate peripheral participation in the creation of scientific documents. *Journal of Documentation*, 65(1): 151-165.

Sharrock, W. & Randall, D. (2004). Ethnography, ethnomethodology and the problem of generalisation in design. *European Journal of Information Systems*, 13(3): 186-194.

Singh, G., Bharathi, S., Chervenak, A., Deelman, E., Kesselman, C., Manohar, M., Patil, S., & Pearlman, L. (2003). A metadata catalog service for data intensive applications. In *Proceedings of the 2003 ACM/IEEE conference on Supercomputing (SC '03)* (pp. 33-). New York: ACM. http://doi.acm.org/10.1145/1048935.1050184

Soergel, D. (1999). The rise of ontologies or the reinvention of classification. *Journal of the American Society for Information Science*, 50(12): 1119-1120.

Srinivasan, R. (2007). Ethnomethodological architectures: information systems driven by cultural and community visions. *Journal of the American Society for Information Science and Technology,* 58(5): 723-733.

Srinivasan, R. & Huang, J. (2005). Fluid ontologies for digital museums. *International Journal on Digital Libraries*, 5(3): 193-204. http://dx.doi.org/10.1007/s00799-004-0105-9

Srinivasan, R., Pepe, A., & Rodriguez, M.A. (2009). A clustering-based semi-automated technique to build cultural ontologies. *Journal of the American Society for Information Science and Technology*, 60(3): 608-620.

Star, S.L. (1999). The ethnography of infrastructure. *American Behavioral Scientist*, 43(3): 377-391.

Stokstad, E. (2011). Pioneering center ponders future as NSF pulls out. *Science*, 332(6032): 905. http://dx.doi.org/10.1126/science.332.6032.905

Strauss, A. (1988): The articulation of project work: an organizational process. *Sociological Quarterly*, 29(2): 163-178.

Strauss, A., Fagerhaugh, S., Suczek, B., & Wiener, C. (1985): *Social organization of medical work*. Chicago, IL: University of Chicago Press.

Suber, P. (2008). An open access mandate for the National Institutes of Health. *Open Medicine*, 2(2). http://www.openmedicine.ca/article/viewArticle/213/135

Suchman, L.A. (1987). *Plans and situated actions: the problem of human-machine communication*. New York: Cambridge University Press.

Suchman, L. (2002). Located accountabilities in technology production. *Scandinavian Journal of Information Systems*, 14(2): 91-105.

Suchman, L.A. (2007). *Human-machine reconfigurations: plans and situated actions, 2nd edition*. New York: Cambridge University Press.

Suchman, L. A. & Trigg, R. H. (1992). Understanding practice: video as a medium for reflection and design. In J. Greenbaum & M. Kyng (Eds.), *Design At Work: Cooperative Design of Computer Systems* (pp. 65-90). Mahwah, NJ: Lawrence Erlbaum Associates.

Sugimoto, S., Baker, T., & Weibel, S. L. (2002). Dublin Core: process and principles. In E. P. Lim, S. Foo, C. Khoo, S. Urs, T. Costantino, E. Fox & H. Chen (Eds.), *Digital Libraries: People, Knowledge, and Technology, Proceedings* (pp. 25-35). Berlin: Springer-Verlag.

Svenonius, E. (2000). *The intellectual foundation of information organization*. Cambridge, MA: MIT Press.

Swan, A. & Brown, S. (2008). *The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs.* Report to the JISC, School of Electronics & Computer Science, University of Southampton. http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinalreport.pdf

Talja, S. (1999). Analyzing qualitative interview data: the discourse analytic method. *Library & Information Science Research*, 21(4): 459–477.

Taylor, A.G. & Joudrey, J.N. (2009). *The organization of information*. 3rd ed. Westport, CT: Libraries Unlimited.

Trace, C. (2007). Information creation and the notion of membership. *Journal of Documentation*, 63(1): 142-163.

Traweek, S. (1988). *Beamtimes and lifetimes: the world of high energy physicists*. Cambridge, MA: Harvard University Press

Traweek, S. (1996). Unity, dyads, triads, quads, and complexity: cultural choreographies of science. *Social Text*, 46/47: 129-139.

Turnbull, D. (2007). Maps narratives and trails: performativity, hodology and distributed knowledges in complex adaptive systems – an approach to emergent mapping. *Geographical Research*, 45(2): 140-149. http://dx.doi.org/doi:10.1111/j.1745-5871.2007.00447.x

Turnbull, D. (2009). Working with incommensurable knowledge traditions: assemblage, diversity, emergent knowledge, narrativity, performativity, mobility and synergy. *ThoughtMesh*. http://thoughtmesh.net/publish/279.php

Uhlir, P.F. & Schröder, P. (2007). Open data for global science. *Data Science Journal*, Volume 6. http://www.jstage.jst.go.jp/article/dsj/6/0/OD36/_pdf

Unsworth, J., Courant, P., Fraser, S., Goodchild, M., Hedstrom, M., Henry, C., Kaufman, P. B., McGann, J., Rosenzweig, R. & Zuckerman, B. (2006). *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for Humanities and Social Sciences*. American Council of Learned Societies.

U.S. National Institutes of Health. (2003). *NIH Data Sharing Policy and Implementation Guidance*. http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm.

Van House, N.A. (2004). Science and technology studies and information studies. *Annual Review of Information Science and Technology*, 38: 3-86.

Vertesi, J. & Dourish, P. (2011). The value of data: considering the context of production in data economies. In *Proceedings of the ACM 2011 conference on Computer Supported Cooperative Work* (pp. 533-542), New York: ACM. http://doi.acm.org/10.1145/1958824.1958906

Wallis, J.C., Borgman, C.L., Mayernik, M.S., & Pepe, A. (2008). Moving archival practices upstream: an exploration of the life cycle of ecological sensing data in collaborative field research. *International Journal of Digital Curation,* 3(1). http://www.ijdc.net/index.php/ijdc/article/viewFile/67/46

Wallis, J.C., Borgman, C.L., Mayernik, M.S., Pepe, A., Ramanathan, N., & Hansen, M. (2007). Know thy sensor: trust, data quality, and data integrity in scientific digital libraries. In *11th European Conference on Digital Libraries* (pp. 380-391), Budapest, Hungary, Berlin: Springer.

Wallis, J.C., Mayernik, M.S., Borgman, C.L., & Pepe, A. (2010). Digital libraries for scientific data discovery and reuse: from vision to practical reality. In *Proceedings of the 10th annual joint conference on Digital libraries (JCDL '10)*. New York: ACM (pg. 333-340). http://doi.acm.org/10.1145/1816123.1816173

Wayne, L. (2005). *Institutionalize metadata before it institutionalizes you*. Federal Geographic Data Committee. http://www.fgdc.gov/metadata/documents/InstitutionalizeMeta_Nov2005.doc

Weibel, S. (1995). Metadata: the foundations of resource description. *D-Lib Magazine,* 1(1). http://www.dlib.org/dlib/July95/07weibel.html

Weick, K.E. (1995). *Sensemaking in organizations*. Thousand Oaks, CA: Sage.

Weissman, D. (2000). *A social ontology.* New Haven, CT: Yale University Press.

Wenger, E. (1998). *Communities of practice: learning, meaning, and identity*. Cambridge, UK*:* Cambridge University Press.

White, H.C. (2010). Considering personal organization: metadata practices of scientists. *Journal of Library Metadata*, 10(2): 156-172.

White, H.D. & McCain, K.W. (1998). Visualizing a discipline: an author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science,* 49(4): 327–355.

Whitlock, M.C. (2011). Data archiving in ecology and evolution: best practices. *Trends in Ecology & Evolution*, 26(2): 61-65.

Yakel, E. (2001). The social construction of accountability: radiologists and their record-keeping practices. *Information Society,* 17(4): 233–245. http://dx.doi.org/10.1080/019722401753330832

Zimmerman, A.S. (2007). Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *International Journal of Digital Libraries*, 7(1/2): 5-16.