The Role of Conceptions of Value in Data Practices: A Multi-Case Study of Three Small Teams of Ecological Scientists

by

Dharma R. Akmon

**A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2014**

Dissertation committee:

Professor Margaret L. Hedstrom, Chair
Professor George C. Alter
Associate Professor Carl Lagoze
Professor Elizabeth Yakel

For Devon, who is always there to remind me I can get up the mountain.

ACKNOWLEDGMENTS

Many people helped to make this dissertation a reality and deserve special thanks. First, I am deeply indebted to my advisor and dissertation chair, Margaret Hedstrom, whose feedback, council, and support throughout my time at the School of Information were indispensible in my development as a researcher. As I reflect on the distance between my earliest dissertation ideas and the document you see before you, I am particularly grateful to Margaret for her incisive critique throughout the dissertation process. Elizabeth Yakel has also been an encouraging source of advice and mentorship over the last several years. I also thank the other members of my committee—George Alter and Carl Lagoze—for their support of my work and generosity in providing me with insights and new perspectives for the design of my study and analysis of my findings. Additionally, I would like to thank Ann Zimmerman, who served on my committee during the proposal phase and helped me greatly in shaping the study.

The feedback and camaraderie of fellow students—both at the School of Information and at other schools—helped me develop my ideas and writing as well as provided a much-needed outlet for the ups and downs of graduate school. In particular, I need to thank Amelia Acker, Matt Burton, Eric Cook, Morgan Daniels, Kathleen Fear, and Ricky Punzalan. You all not only made graduate school a heck of a lot more fun, but you also generously offered astute critique of my work.

Of course, my family and friends have been there every step of the way in this long process, giving me encouragement and helping me maintain a healthy sense of perspective. Mom, Dad, Pat, Mike, Sonya, John, Desmond, Krista, Daniel, Jaylen, Jesse, Noah, Eli, Joel, Abbey, Heather, Erin, and Janet: I cannot thank you all enough. I must give special credit to my

mom, who instilled in me a lifelong love of knowledge and learning. My husband, Devon, was always an especially patient and reassuring presence, lifting me up when I felt discouraged and providing respite and joy when I needed it most. "Thank you" does not begin to cover how much it has meant for you to do this with me, Devon.

Lastly, this study would not have been possible without the participation of the Station and the scientists in the three teams I studied. Scientists and staff were open, tolerant of my strange questions, and patient with me constantly looking over their shoulders. Furthermore, they made data gathering thoroughly enjoyable as they welcomed me into the Station community. I also acknowledge the financial support of the National Science Foundation IGERT-09036291 (OpenData), the Station, and the University of Michigan Rackham Graduate School.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

This dissertation examines the role of conceptions of data's value in data practices. Based on a study of three small teams of scientists carrying out ecological research at a biological station, my study addresses the following main question: How do scientists conceive of the value of their data, and how do scientists enact conceptions of value in their data practices? I relied on interviews and participant observations for my study and analyzed my data through the lens of theories of value and meaning. I found that scientists were primarily concerned with data's value for their team's own, relatively narrow uses: addressing a gap in knowledge and producing the outputs that would garner them credit and prestige. When asked about their data's potential value beyond their studies, scientists regularly cited metaanalysis, cross-site comparison, and time-based studies as worthy secondary uses for data and assessed data's value according to how well they thought the data could serve those ends.

As they collected data and conducted their studies, scientists did not think about data's value beyond whether or not they were good as resources for addressing a gap in knowledge. However, when asked to make their data more openly available, researchers indicated that their decision to share was based strongly on data's value for producing publications for the team. Data that teams were still working with and planned to publish were regarded as too valuable to the team to make widely available. Conversely, when scientists thought data's publication value had been fully exploited for the team, they saw little threat in sharing. In addition to publication potential, scientists also suggested that study type influenced their decision to share data and told

me that they felt less compelled to share data from controlled studies because they assumed such data had inherently limited value.

CHAPTER 1: INTRODUCTION

In this dissertation, I contribute to a growing body of research on scientific data management by examining how scientists conceive of the value of the data they create and how those conceptions are enacted in their data practices. To date, research on data practices has focused primarily on scientists' willingness or unwillingness to share data or engage in activities—such as data documentation—that would facilitate data reuse by others. This research has revealed important tensions between what an increasing number of constituents—including the public, the National Science Foundation (2011b), and, most recently, the Office of the President (2013)—are asking of scientists and the incentive structure that undergirds much of scientific research. However, the focus on incentives and motivations leaves out other potentially important factors in the decision to share data or create data that endure past the life of a project. In particular, some researchers' findings suggest that conceptions of data's value play an important role in scientists' decisions about what to do with data. An examination of how scientists conceive of their data's value—including what, who, and how long they think their data are good for—promises to contribute to a more comprehensive understanding of scientists' data practices, including sharing data and managing them to ensure their persistence past the life of a project.

1.1 Definition of the Problem

Data, long regarded as expendable by-products of science, are increasingly positioned by the government, the public, and some scientists as valuable products of research. In recent years, they have been variously called a "vital" aspect of eScience or cyberinfrastructure (National Science Foundation Cyberinfrastructure Council, 2007, p. 2), a "collective" resource (Edwards, Jackson, Bowker, & Knobel, 2007, p. 19), an "integral part of the scientific record" (Hey, Tansley, & Tolle, 2009, p. 181), and "national and global assets" (Interagency Working Group on Digital Data to the National Science and Technology Council, 2009, p. 10). Reflecting this view of data, a number of federal funding agencies and scholarly journals have implemented policies meant to encourage data management, open access, and preservation. These, and other advocates of data sharing and archiving, are motivated by the conviction that a significant amount of potentially useful data are not being shared or managed by scientists to facilitate reuse and preservation (Arzberger et al., 2004; Committee on Issues in the Transborder Flow of Scientific Data & National Research Council, 1997; Costello, 2009; Interagency Working Group on Digital Data to the National Science and Technology Council, 2009; Nelson, 2009).

Data's value is largely taken for granted in efforts to increase their availability for secondary reuse. According to sharing and reuse proponents, openly accessible and well-managed data make possible cross-disciplinary research (W. Anderson, 2004); facilitate new scientific advancements (Committee on Issues in the Transborder Flow of Scientific Data & National Research Council, 1997; Erpanet, 2003); maximize public investment in science (Arzberger et al., 2004); enable replication and verification (Uhlir & Schröder, 2007); and allow the identification of long-term trends (Lauriault, Craig, Taylor, & Pulsifer, 2008). These arguments and the policies inspired by them have not, however, necessarily impacted how

scientists manage their data or whether or not they share their data. In fact, researchers have shown that the extent of data sharing remains minimal in many fields (Tenopir et al., 2011), despite journal requirements and funding policies (Piwowar & Chapman, 2008, 2009) and an increasingly strong social message about the importance of managing data for reuse (Tucker, 2009).

Research on scientists' data practices has sought to better understand why scientists rarely share or manage data for reuse, frequently emphasizing the conflict between the current incentive structure of science and the new demands being placed on scientists. Scientists, this research has revealed, withhold or inadequately manage data for reuse for several key reasons: documentation takes a significant amount of work that is not rewarded (Birnholtz & Bietz, 2003; Campbell et al., 2002; Louis, Jones, & Campbell, 2002); scientists are more concerned with publications (which are rewarded) as a product of their work (Borgman, Wallis, & Enyedy, 2007); and scientists fear that their data contributions will not be formally recognized (through citation or co-authorship, for example) (Louis et al., 2002) or that their data will be misused (Baker & Millerand, 2010; Cragin, Palmer, Carlson, & Witt, 2010).

Numerous disincentives negatively influence sharing or the creation of "archive-ready" data (Hedstrom & Niu, 2008); however, as the Blue Ribbon Task Force on Sustainable Digital Preservation and Access underscores, ensuring continuing access to science data encompasses not only clear incentives to act in the public interest, but also the articulation of a compelling value proposition (2008). Those who determine, through their decisions and actions, whether data are accessible over time (and this includes scientists as data creators) must, in other words, see some apparent benefit to doing so.

Empirical work on scientists' data practices suggests that conceptions of data's value are an important consideration in scientists' decisions about what to do with data. Social scientists in one study, for example, reported that they would be more likely to document and deposit their data if they thought those data "would be used and have a broader public benefit" (Hedstrom & Niu, 2008, abstract). In another study, scientists considered data shareable when they had the potential to generate new results (Cragin, Palmer, Carlson, et al., 2010). The Research Information Network revealed a similar concern with data's value among scientists across a variety of disciplines, who found it "difficult to believe" that other scientists would actually want their data (2008, p. 28). In contrast, some studies have found that scientists are *less* likely to share data that they think are of high value (Tucker, 2009) or that are hard-won or difficult to generate (Borgman, Wallis, & Enyedy, 2007).

There is strong evidence that scientists' conceptions of their data's value influence their data practices, but no research to date has studied this relationship in depth. Further, there has been very little attention paid to how scientists think about their data's value as they work and the particular purposes for data that they have in mind as they collect and work with their data. While value is an underlying concept in research that focuses on sharing and withholding, these studies seem to take data's value for granted. As a result, *not* sharing or effectively managing data for reuse is attributed almost exclusively to a mismatch in incentives, and associated implications and recommendations tend to focus on strengthening policies or removing barriers to sharing all data. However, without understanding how scientists conceive of their data's value, including differing value they assign to different data and the benefits or purposes that underlie those assessments, researchers cannot fully characterize their data practices nor can funders,

repository managers, and publishers appropriately and effectively target policies, mandates, and incentives.

1.2 Research Questions

This dissertation addresses the lack of understanding of the role of conceptions of value in scientists' data practices. Specifically, I analyze the views and practices of scientists conducting research at an ecological field station and working in small teams, who were collecting, analyzing, and otherwise working with data they created. My research answers the following question:

How do scientists conceive of the value of their data, and how do scientists enact conceptions of value in their data practices?

This question is made up of several related subsidiary questions:

- What specific uses for data are salient to scientists (e.g. as evidence of claims; resources for conducting longitudinal studies; inputs for new research questions)?
- What time spans do scientists use to think about their data's value?
- On what basis do scientists assess data's value? For example, do they consider the speed with which technological advances are expected to render data collected with a particular piece of equipment obsolete; the processing state of the data; the ease of replicating the data, etc.?
- How do scientists create data that are valuable (as construed by the scientists themselves), and what do they do in response to their notions of data's value?

1.3 A Multi-Case Study at a Field Station

In this dissertation, I focus on the views and practices of scientists who worked in three small teams and conducted research at a U.S. university-sponsored field station (referred to throughout the dissertation as "the Station"[1]). Broadly speaking, these researchers carried out ecological research. Ecology—a multidisciplinary field that includes disciplines such as wetland ecology, atmospheric science, biogeochemistry, and geoscience (among many others)—is the study of relationships of organisms to one another and with the environment. I focused my study on scientists conducting ecology-related research for two reasons.[2] First, those with a concern for advancing ecological knowledge (including the scientists themselves) have made a strong case for data integration across both time and sites to look at large-scale phenomena that are of key importance to human welfare (Michener & Brunt, 2000; M. Palmer et al., 2005; Whitlock, McPeek, Rausher, Rieseberg, & Moore, 2010). In other words, they have put forth a well-articulated value proposition that says that carefully managed, shared, and preserved data will not only advance science, but also help society respond to environmental crises. In a recent example, ecologists argued that they could have better understood the impacts of the 2010 *Deepwater Horizon* oil spill in the Gulf of Mexico had they had access to relevant planktonic, oceanographic, and atmospheric science data (among other kinds of data) (O. Reichman, Jones, & Schildhauer, 2011); and, further, that meeting mandated species recovery goals for the oil spill would depend crucially on openly accessible and well-managed data from a number of disciplines (Bjorndal et al., 2011).

---

[1] I have changed all place and person names to protect participants' identities.

[2] Many of the scientists at the Station, including my participants, labeled themselves according to a more specific disciplinary orientation than "ecology." However, when referring to my participants throughout the dissertation, I employ the terms "researcher," "scientist," and "ecologist" interchangeably.

At the same time, ecological science is primarily comprised of small science disciplines, with characteristics that make widespread agreement on what the valuable data products are especially challenging. Small science, often contrasted with big science, is generally driven by individual investigators or small teams, with data collected and analyzed independently (J. Reichman & Uhlir, 2001). In big science fields, such as physics and astronomy, scientists rely on equipment of such a scale and expense that the resulting data is often shared among many collaborators (Borgman, Wallis, & Enyedy, 2007). In small science fields, like biogeochemistry and wetland ecology, scientists tend to collect heterogeneous and non-standardized data. Traditionally, few of these data have been deposited in repositories (J. Reichman & Uhlir, 2001) and instead have been managed locally, according to data creators' own needs (Borgman, Wallis, Mayernik, & Pepe, 2007). Several researchers have argued that we need extensive studies of small science fields to better understand how scientists manage data in a local context, emphasizing that "little is understood about the diverse ways [these] scientists actually produce, manage, and use data" (Baker & Millerand, 2010, p. 115; Borgman, Wallis, Mayernik, et al., 2007).

In this dissertation, I begin to fill that gap. The field station where I carried out my study hosts a number of teams of scientists every year and has recently implemented a "data management" policy. Specifically, the Station's policy states that researchers must submit a copy of their data, along with the appropriate metadata, to the Station for deposit in the repository within a year after researchers have completed their data collection. By archiving and making the datasets available, the Station hopes to ensure long-term access to data created using Station resources and to help scientists fulfill funding mandates.

1.4 Study Overview

This dissertation is based on a study of three small teams of scientists that conducted research at the Station during the summer of 2012. Approximately 40 research projects are carried out at the Station each summer; projects are funded by numerous agencies (including the NSF, the U.S. Department of Defense, the U.S. Department of Energy, and the EPA) and are concerned with a range of phenomena. The three teams I studied varied along several important dimensions, including career stage of the primary investigator(s), the type of investigation, the time span of the project, and the funding source. Together, the three cases I selected elucidate a range of ways that scientists think about data's value and manage data, based on socially situated meanings of data that include scientists' notions about the purposes data serve; the time spans over which data will useful; who might benefit from using the data; and what is required to create valuable data.

This study relies on three main sources of data: participant observations of scientists as they collected, analyzed, and managed their data; semi-structured interviews with scientists and Station staff and faculty; and documentary sources, including relevant funding and journal policies, grant applications, data, data documentation, and discipline-specific literature on sharing and reuse. I conducted the bulk of my observations and interviews with participants over an eight-week period during which I lived at the Station along with the scientists carrying out research. I employed qualitative methods to analyze all the data I collected.

1.5 Key Terminology

Below, I provide definitions of several terms I use regularly throughout the dissertation.

Data Practices: Data practices in this study refers to the "research processes and activities related to scientists' work with data" (Cragin, Palmer, & Chao, 2010, para. 5). These practices encompass data collection, description, and analysis as well as activities like data sharing.

Data Management: I use "data management" to refer to those activities undertaken to meet the need—both long- and short-term—to access, use, and understand data. This includes, but is not limited to, activities such as documenting contextual information about data, organizing data, and storing data. Data practices include data management; data management is a more specific set of activities than data practices.

Data Sharing: The act of making data available for others' use. This can be accomplished any number of ways, including via personal contact or publically accessible databases. The chosen mode of sharing might facilitate long-term preservation (for example choosing to deposit data with a disciplinary repository that ensures longevity), but not necessarily.

Continuing Value: Archivists define continuing value as the "enduring usefulness or significance of records, based on the administrative, legal, fiscal, evidential, or historical information they contain, justifying their ongoing preservation" (Pearce-Moses, 2005). To encompass the context of scientific data, I focus on the notion of enduring usefulness. Understanding usefulness for whom, what purpose, and for how long is one of the concerns of this study.

1.6 Theoretical Foundations

In this study, I focus on scientists' conceptions of data's value to better understand how they create, manage, and work with data. The theoretical foundations for this study are constructivist; I draw on theories from sociology and the sociology of science as well as on philosophical explications of value. In employing these theories, I position scientists' conceptions of data's value as an important factor across the span of the research process. Scientists' notions about data's value arise from the meaning(s) that data have for them. These meanings are socially situated and include assumptions about the purposes for data, the traits that make data good, how long data will be valuable, and who could (or should) benefit from data's use.

1.6.1 DATA STREAM MODEL

One of the many problems with making broad claims about the value of scientific data is that there is no one, agreed upon, standard definition for data that applies across all disciplines. There are, in fact, many different ways to conceptualize what data are. The *Oxford English Dictionary* defines a datum (the singular of data) as

> a thing given or granted; something known or assumed as fact, and made the basis of reasoning or calculation; an assumption or premise from which inferences are drawn;

and data (plural) as "facts, especially numerical facts, collected together for reference or information." Open data sharing and long-term preservation proponents tend to similarly treat data as fairly self-contained and stable entities. U.S. government-sponsored reports, for example, liken data to bricks (Committee on Issues in the Transborder Flow of Scientific Data & National Research Council, 1997, p. 47) that make up the foundation of scientific knowledge or position them as "endless fuel for creativity" (Interagency Working Group on Digital Data to the National Science and Technology Council, 2009, p. 1).

Some definitions of data emphasize the varied forms that they can take:

> Scientific or technical measurements, values calculated there from, and observations or facts that can be represented by numbers, tables, graphs, models, text, or symbols and that are used as a basis for reasoning and further calculation (Committee on Issues in the Transborder Flow of Scientific Data & National Research Council, 1997, p. 197).

Others specifically emphasize data's increasingly digital format to highlight their potential for reanalysis. For example, Simberloff et al. (2005) define data as

> [. . .] any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc. (p. 9).

Some definitions of data focus on differentiating *kinds* of data. For example, level of processing is viewed as an important characteristic of data: raw/level-one data (often the most

difficult for others to understand); second-level data (generally what interpretations in papers are based on); and third-level data (compilations of data from several sources) (Committee on Issues in the Transborder Flow of Scientific Data & National Research Council, 1997). Alternatively, another classification is based on the type of phenomena data represent: observational data represent one-time phenomena; experimental data are collected through controlled experiments; computational data are generated from simulations; and reference datasets are curated collections of data (Committee on Issues in the Transborder Flow of Scientific Data & National Research Council, 1997).

These definitions of data largely obfuscate, or at least overlook, the work, judgments, and decisions involved in making data that can be used meaningfully over time and/or by others. Adopting a constructivist perspective to data, Hilgartner and Brandt-Rauf (1994) warn that we must not

> [. . .] assume that data somehow arrive on the scene in pre-packaged units that are transferable, sharable, or publishable, or that there is some discrete point in time at which data should "naturally" be transferred (pp. 361-362).

Instead, they propose a more process-based approach to understanding data-access issues, characterizing data as part of an "evolving stream." In their data stream model, Hilgartner and Brandt-Rauf describe the stream as a collection of heterogeneous elements that includes—among other things—output, equipment, know-how, and samples. Because the stream is heterogeneous, Hilgartner and Brandt-Rauf stress the importance of taking a broad view of data: one that encompasses the myriad component elements instead of only some pre-determined end product of research.

In addition to heterogeneity, the data stream model emphasizes several important qualities of data elements that are likely to bear on scientists' considerations of data's value. One

characteristic is the availability or rarity of a data element; at one end of the spectrum are commonly or widely accessible items, such as equipment; and on the other end are very rare items, such as craft knowledge of a novel data collection technique. Another important data characteristic is the factual status of data, which Hilgartner and Brandt-Rauf equate with certainty and reliability. At one end data are "so uncertain that even the scientists who produced them doubt their credibility and utility," and on the other end data are "widely regarded [as] reliable and valuable" (pp. 360-361). Time is an important factor for this characteristic, since data are constantly being interpreted and reinterpreted and their credibility and utility being reassessed. Finally, Hilgartner and Brandt-Rauf draw on Latour's (1987) concept of inscriptions to emphasize that data streams are made up of chains of products that range from relatively raw to highly refined. On the raw end of the spectrum lie first-order inscriptions, such as the output from primary inscription devices, while on the highly refined end lie extensively processed data graphs intended for publication in scientific papers. These different levels of inscription represent, they argue, not only changes in the form of data, but also alter the "purposes for which they can be used" and therefore the data's utility or value (p. 361).

The data stream model has important implications for studying scientists' conceptions of data's value and their data practices. Specifically, this model suggests that how scientists view the value of the data they create will vary throughout the research process. For example, if inscription level is an important characteristic for assessing data's value, then scientists would likely regard data collected at earlier stages of their study (i.e. rawer data) as having different worth than data produced in later stages (i.e. heavily processed data). Furthermore, Hilgartner and Brandt-Rauf's model suggests that purposes against which data's value is judged are likely to vary across the data stream. This leads to questions such as: What are the purposes that are

assessed during the research process? Which purposes are salient to scientists at different stages? How do notions about data's uses impact on the creation of what scientists would consider to be valuable data?

Lastly, the model also suggests that important decisions and valuations occur across the stream. Data preservationists have noted the path-dependent nature of digital preservation (Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2008): the decisions made early in data's life cycle (or early in the stream) play a crucial role in determining whether or not data will endure (Cedars Project Team, 2001; National Academy of Sciences, 2009). One of the aims of my study, then, is to understand not just whether scientists regard their data as having long-term value (Cragin, Palmer, Carlson, et al., 2010), but also how and why scientists create valuable objects that might persist past the life of their project.

1.6.2 VALUE, VALUATION, AND MEANING

The meaning of value is often ambiguous. Simplistically, value is defined as the worth of something and valuation as the *process* of estimating, assessing, or measuring the worth of something (Saracevic & Kantor, 1997).[3] Najder describes three main senses of value: 1). what something is worth (often this is expressed in numeric or monetary terms and is the sense of value that economists commonly refer to); 2). something to which worth is ascribed (the object is regarded as possessing value on the basis of its qualities); or 3). an ideal that causes one to judge "objects, qualities, or events as valuable" (e.g. fairness as a value) (1975, p. 42). Philosophical theories emphasize value's multi-dimensional nature, highlighting that assessments of value

[3] Value*s* (sometimes referred to as "value with a capital V") can also refer to people's and society's concepts about what is good, just, moral, or ethical. While a prominent topic of study in both philosophy and sociology, it is not the meaning of value I am concerned with in this study.

encompass aspects of the object being assessed (Beckert & Aspers, 2010) and the benefits and purposes at issue (Rescher, 1969). In other words, to understand what is meant when something is said to have high value or be valuable, we should ask: value for what and whom and on what basis?

An assumption that scientific data are valuable products of science underlies claims that scientists should manage and share their data for reuse and preservation. In many instances, value is treated as a quality inherent to data. For example, the National Research Council argues that data (seemingly all data) are valuable because they facilitate new discovery, while the National Academy of Sciences states that data's ability to ensure verification and replication make them worthy of preservation and sharing (Committee on Issues in the Transborder Flow of Scientific Data & National Research Council, 1997; National Academy of Sciences, 2009). Other organizations and researchers emphasize that data allow for endlessly exploiting research investment for new ends (Arzberger et al., 2004; Association of Research Libraries, 2006; National Academy of Sciences, 2009; Whitlock, 2011).

Some data preservation proponents assert that effective data stewardship will depend on the careful selection of data, and, as a result, have attempted to discern which kinds of data are most likely to have *long-term* value. Several have proposed that long-term value stems from the type of phenomena data represent and their resulting degree of uniqueness (W. Anderson, 2004; Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010; Simberloff et al., 2005; Steering Committee for the Study on the Long-term Retention of Selected Scientific and Technical Records of the Federal Government National Research Council, 1995b). This schema positions observational data, such as astronomical observations of galaxy formation, as having greater preservation value because they represent one-time phenomena that cannot be recreated.

Experimental data, which scientists could conceivably recreate, are viewed as less worthy of long-term preservation (Simberloff et al., 2005). Other researchers, however, have noted that it is often difficult (or even impossible) from a practical standpoint to recreate experimental data (Collins & Pinch, 1998).

In another perspective on the value of scientific data, Heidorn (2008) and Cragin et al. (2007) argue that considerable value is likely to be found in what they call "the long tail of science." These authors use "the long tail" to refer to the portion of scientific research that is carried out by individual scientists or small teams of scientists resulting in "small but numerous data collections" (C. Palmer et al., 2007, para. 2). The value of data in the tail—comprised mainly of heterogeneous datasets (many of them experimental)—could derive, in large part, from their integration across time, sites, and studies, regardless of any individual dataset's reproducibility (C. Palmer et al., 2007).

These different conceptualizations of data's value highlight value's dependence on assumptions about what and whom data are for and also suggest several characteristics of data that could be used by those who seek to determine data's sharing or preservation worth. Problematically, however, these characterizations set up normative assumptions about the way scientists *should* regard and care for the data they generate. Policy makers then treat departures from the norm (i.e. not sharing openly) as indicative of scientists' unwillingness to share or engage in data documentation, instead of as a possible expression of how *scientists* view their data's value. In the interest of better understanding scientists' data practices, my study focuses on how scientists think of their data's value as they collect and work with data and as they consider sharing or archiving those data.

The philosophical study of value (known as "axiology") is concerned with value as it

relates to ethics, or what is good, right, and moral. In characterizing what value is, philosophical theories of value generally recognize two main types: intrinsic and instrumental (also frequently called extrinsic). Intrinsic value refers to the quality of something being "worthy in and of itself" (Saracevic & Kantor, 1997, p. 529); pleasure, for example, might be regarded as having intrinsic value. Instrumental value applies to the quality of something being valued "as a means of attaining some end" that is considered worthy (Attfield, 1987, p. 39). For example, when researchers who study scientific data practices observe that data play important roles for scientists in proving hypotheses (Latour & Woolgar, 1986), creating publishable products (Borgman, Wallis, & Enyedy, 2007), and forming collaborations with others (Birnholtz & Bietz, 2003), they describe purposes or benefits that scientists expect to achieve through their data. This hints at an instrumental value for data: scientists value the data for what they expect the data will allow them to do.

Theorists argue that objects can exhibit both kinds of value: knowledge, for instance, is often regarded as having both intrinsic and instrumental value (Kirschenmann, 2001; Lemos, 1995; Saracevic & Kantor, 1997). In the scholarly research context, in particular, the accumulation of knowledge is regarded as valuable for its own sake; and at the same time knowledge can be valued for its ability to reveal a cure for a common disease (Kirschenmann, 2001).

Philosophical explorations of value provide a basis for understanding some of the dimensions along which scientists might think about their data's value. In particular, they point to the importance of considering the set of aims scientists strive toward as they collect and work with their data. Where data exhibit instrumental or use value for scientists, what specific purposes are salient to scientists? For example, do scientists value data for proving claims;

conducting longitudinal studies; or answering new research questions? Do the purposes concern only their own uses or also other scientists' uses of their data? And then what data traits are important to scientists as they think about data's value for the purposes that are salient to them? For instance, do scientists consider the speed with which technological advances are expected to render data collected with a particular piece of equipment obsolete; the processing state of the data; or the ease of replicating the data?

Philosophical characterizations, whether describing instrumental or intrinsic value, also hint at a more fundamental aspect of value put forth by Beckert and Aspers (2010): "Value is not intrinsic to the materiality of an object but rather is inseparably connected to the concept of meaning" (p. 11). Conceptions of value are tied to the characteristics of an object being evaluated, but how people valuate objects depends on the meanings that objects take on for them. Beckert and Aspers assert that these meanings arise from within social systems.

In his theory of symbolic interactionism, which largely rests on the premise that people act toward things and objects on the basis of the meanings objects have for them, Blumer (1969) similarly argues that meaning is neither a purely psychological phenomena nor something intrinsic to objects. Instead, meanings are social products that arise from ongoing interactions between people. Meanings are never stable, but rather people deal with and modify meanings through an interpretive process. Taking a similar perspective, Wenger (1998) provides a useful framework for examining meaning, which he identifies as being negotiated within communities of practice.

Wenger defines a community of practice as a group of mutually engaged participants working on a joint enterprise and utilizing a shared repertoire of resources. Participation in these communities not only shapes what we do, but also shapes who we are and how we interpret our

actions. Several studies of information use, including those on scientific data practices, have employed Wenger's communities of practice model (e.g. Birnholtz & Bietz, 2003; Talja, 2002; Zimmerman, 2003), largely because it emphasizes the socially situated nature of using, creating, and managing information resources. For this study, I use Wenger's framework to elucidate data's meanings for scientists and, hence, to better understand scientists' conceptions of data's value. The key reason I have focused my study on small teams of researchers is that the social phenomena of meaning-making and valuation are likely to be more visible than they would be were I to study the practices of individual scientists working alone.

Wenger repeatedly highlights the provisional and situated nature of meaning. In fact, he argues that in order to find the meaning of activities, objects, or concepts for individuals, we must look at the practice and process of its making or the negotiation of meaning. Wenger characterizes this negotiation (which is meant to highlight the provisional and ongoing nature of it) as continuous, active, and dynamic. It is both historical and rooted in the present context and involves interpretation and action; doing and thinking; understanding and responding. Meaning-making happens constantly, resulting in "extend[ing], redirect[ing], dismiss[ing], reinterpret[ing], modify[ing], or confirm[ing] [. . .] the histories of meanings of which they are a part" (pp. 52-53).

According to Wenger, meaning negotiation is made up of two component processes that exist in a duality with each other: participation and reification. Participation refers to the "social experience of living in the world in terms of membership in social communities and active involvement in social enterprises" that "combines doing, talking, thinking, feeling, and belonging" (Wenger, 1998, pp. 55-56). It is important to note that, in Wenger's model, this participation is *always* social, even when it does not directly involve conversation or other

interactions between people. A scientist working alone at her computer to document the details that she might need to evaluate a set of data later implicitly involves people who are not present, including the other members of her disciplinary community and the people she works with more closely.

Reification refers to "the process of giving form to our experience by producing objects that congeal this experience into 'thingness'" or the process of making something abstract more concrete (Wenger, 1998, p. 58). These reifications, according to Wenger, create points of focus around which actors can organize the negotiation of meaning. Wenger includes laws, procedures, and tools as examples of reifications. In looking at the work of scientists, reifications include data management mandates, plans, metadata standards, local workplace policies and procedures, journal publication policies, and the data themselves (to name just a few).

To understand the negotiation of meaning, then, Wenger asserts that we must look at the ways in which reification and participation interplay or come together. Looked at through this conceptual lens, scientists' data practices can be viewed as an example of negotiating meaning. These practices (which include attaching descriptions to data and arranging, processing, and transforming data) take place within a context that includes the scientists' local work setting, disciplinary expectations, work needs, training, the data themselves, past experiences with similar data, the time that they have to engage in data management practices, standards, the scientists' own planned research trajectory, who scientists are working with, and what things are going on as they engage in data practices. The scientist's role (participation) in meaning negotiation in data practices comes from her being a member of a particular community of practice and her related history of participation in that community. Data, mandates, policies, and standards also contribute to the negotiation of meaning in data practices by reflecting aspects of

practice that have been fixed or reified. The convergence of these two (participation by scientists and reification as embodied in artifacts) is where negotiation of meaning takes place; where the meanings of data and what they are good for are formulated. These meanings not only have direct bearing on scientists' conceptions of their data's value, but the data practices themselves also work to create data that are good for some purposes and not good for others.

1.7 Significance of the Study

As funders, the public, journal publishers, and scientists increasingly expect data to be managed as reusable products of research, it becomes ever more important to understand how scientists work with data and the factors underlying their practices. Much of the rhetoric on data sharing begins from the perspective that data are valuable and then asserts that scientists *should* manage data to support reuse and preservation. When scientists do not adhere to this norm, data sharing proponents often regard them as unwilling to share data or manage them for reuse. Empirical work on data practices suggests that scientists' conceptions of their data's value influence their decisions about providing access to data or making them archive-ready. Yet we have only a rudimentary understanding of how scientists think about their data's value, particularly as they engage in their day-to-day research practices.

In this study, I focus on scientists carrying out ecological research at a field station. By examining researchers' conceptions of data's value as they relate to data practices, this study offers a different perspective on how scientists manage their data and the aims of such activities. One of the primary contributions of this research is a better understanding of how scientists determine that some of their data deserve actions that facilitate the data's endurance beyond their immediate time of collection or creation.

In terms of broader impacts, my findings have the potential to help policy makers, repository managers, and others interested in ensuring data preservation and reuse in two main ways. First, a more thorough understanding of how scientists conceive of the value of their data (including value for what, whom, and how long?) has the potential to reveal possible avenues for impacting scientists' data management behavior. Second, while many data stewardship proponents have argued that selection will be a necessary component of science data preservation efforts, governmental advisory committee reports on data sharing and federal funding policy seem to still be guided by an implicit assumption that scientists should manage *all* of their data to facilitate future reuse. Such a perspective not only places an onerous responsibility on scientists, but also potentially prevents targeted investment in stewardship for the most valuable data. While this dissertation does not identify which particular kinds of data are most valuable for preservation, it does reveal how three teams of scientists think about their data's value and hence sheds light on whether, how, and why the scientists I studied came to view any of their data as a product with continuing value.

CHAPTER 2: LITERATURE REVIEW

2.1 OVERVIEW

My literature review brings together several areas of research to lay the foundation for a study on how scientists' conceptions of data's value relate to their data practices. I begin by describing the demands being placed on scientists regarding their data and the reasons for those demands, particularly as it applies to ecology and the long tail of scientific research. These demands have placed considerable attention on scientists' data practices; most notably the degree to which scientists share and withhold data and their reasons for doing so. Much of this research focuses on scientific norms, incentives, and motivations, so I review this work as it relates to scientific practice before moving on to a discussion of sharing and withholding studies. Then, I discuss the current state of knowledge regarding scientists' sharing and withholding behavior to argue that it presents a limited view of how scientists make decisions about what to do with their data. I follow this with a section devoted to research, primarily drawn from science and technology studies (STS), on the contingency of data and the various roles that data play for scientists. I conclude the chapter by reviewing the limited work that has been done to understand how scientists view the value of the data they create.

2.2 CHANGING EXPECTATIONS FOR DATA AND THE PROMISE OF THE LONG TAIL

In virtually every scientific field, scientists are generating data at an unprecedented rate, primarily using digital technologies. In climate science, for example, modeling simulations have

been a significant contributor to what has been called an "explosion in data" (Overpeck, Meehl, Bony, & Easterling, 2011, p. 701); and in genomics, genetic sequencing output has doubled approximately every nine months (Kahn, 2011). Even in small science fields, such as habitat ecology, scientists increasingly deploy sensor technologies, resulting in large volumes of digital data (Borgman, Wallis, & Enyedy, 2007). The proliferation of digital data and more pervasive networking capabilities have given scientists the ability to access and transfer data more readily, as well as to conduct analyses over much larger-scale data collections than ever before.

In many areas of research, these changes have already enabled scientific discovery that was previously inconceivable. For example, the mapping of the human genome and subsequent development of tools to analyze the resulting database, have made it possible for scientists to identify the genetic markers associated with specific diseases. In astronomy, scientists are able to use supercomputers to simulate (based on observational data) the collision of two black holes to better understand the fundamental laws of physics and perhaps even the origins of our universe (National Science Foundation Cyberinfrastructure Council, 2007; O'Hanlon, 2010; Reddy, 2012).

The promise of transformative science, however, is not limited to big science fields or petabyte-sized datasets. In fact, several researchers have begun to argue that there is considerable value in bringing together the countless small science datasets that make up a significant portion of U.S. federal research expenditure (Heidorn, 2008). Inspired by a power law distribution concept introduced by Chris Anderson (2004) to explain the business strategy of online retailers, such as Amazon and Netflix, several researchers have employed the phrase "the long tail of scientific research" to describe these many, relatively small datasets collectively (Heidorn, 2008; C. Palmer et al., 2007). Heidorn (2008) characterizes the long tail as being comprised of

heterogeneous data that are generally not well-serviced by established disciplinary repositories. The data in the tail tend to be hand-generated (as opposed to mechanized), the result of unique procedures, and subject to individual curation. As a result, the data—which Heidorn estimates are the most voluminous by the nature of the number of studies and amount of federal funding they represent—are also the least accessible to others since they are primarily managed locally.

Palmer et al. (2007) suggest that substantial value could be realized in making data in the long tail more widely available. They argue that environmental and ecological science in particular "can profit greatly from more prudent management of the data resulting from long-tail science" (para. 9). Not only does better management of data in the tail have the potential to facilitate greater access to data for a number of important environmental stakeholders, but it might also allow the kind of multi-disciplinary, cross-site integration of data that many see as key to solving contemporary scientific grand challenges, such as global climate change and natural resource management (W. Anderson, 2004; Atkins et al., 2003; C. Palmer et al., 2007).

The growing emphasis on data in virtually all scientific fields marks an important shift and challenges much of traditional scientific practice. Data are no longer assumed by those who fund their creation and collection to be expendable by-products of research, but rather, are conceived of as enduring products in their own right (Bowker, 2000, 2006). Further, whereas scientific data have traditionally been viewed as private possessions (McCain, 1991; McSherry, 2001) of limited value once papers were published (Latour & Woolgar, 1986), they are now frequently positioned, especially by research funders and journal publishers, as a collective resource with continuing value (Edwards et al., 2007). A prominent demonstration of this shift is the U.S. National Science Foundation requirement, implemented in 2011, that all grant proposals include a data management plan (National Science Foundation, 2011a). While the NSF does not

require scientists to deposit their data in a repository (and, in fact, the NSF has not yet indicated whether or how it will monitor and enforce adherence to the data management plans), it does "expect" researchers to

> [. . .] share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants (National Science Foundation, 2011a).

The promised potential of data repositories, however, demands far more than that scientists just make their data available at the conclusion of a project (which presents plenty of challenges in itself). If data are to be fuel for the scientific revolution enabled by the digital age, scientists, as data creators, must manage their data in such a way that they are usable for others. Scientists are expected, in other words, to consider their data's potential value for purposes beyond their immediate needs and engage in activities that promote the data's capacity for reuse.

Scientists, it has been noted, may have little incentive to think beyond their own immediate needs (Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2008) and tend to have a necessarily narrow conception of the possible uses of their data (Steering Committee for the Study on the Long-Term Retention of Selected Scientific and Technical Records of the Federal Government National Research Council, 1995a). Yet they represent crucially important decision-makers in enabling the continuing viability of the data they produce. Decisions that data creators make at the point of data creation and shortly thereafter have long-standing influence on what can be preserved (National Academy of Sciences, 2009). As the Cedars Project Report emphasizes,

> [. . .] the way a digital object is created influences how (or indeed whether) it can be preserved. Likewise, decisions taken at the start-point of preservation can impact on future access (Cedars Project Team, 2001, p. 68).

There are many concrete examples in the data curation literature that serve as proof of the impact of scientists' actions on data preservation and reuse. Gutman et al. (2004) report that the Inter-university Consortium for Political and Social Research (ICPSR) data archives determines which data to accession based in large part upon the format of those data and the level and quality of descriptive information provided by the scientist(s) who created the data. In her study of the experiences of ecologists in reusing data created by others scientists, Zimmerman (2003) found that documentation of research methods, which can only be adequately captured by the scientists who carried out those methods, was used extensively as a tool for assessing the quality and applicability of data for particular purposes. Further emphasizing the crucial role of documentation in secondary data interpretation, Wallis et al. (2008) identified nine different stages at which scientists make decisions that ultimately affect data. This includes how numerical data are derived from other kinds of data and how data are normalized based on calibrations. In the absence of this knowledge—often taken for granted by data creators—it is difficult if not impossible to use the data meaningfully.

Without attention to preservation starting at the point of data's creation and beyond, by the time that many digital data arrive at repositories it is often too late to properly ensure their preservation (Wallis et al., 2008). Regardless of who *should* be responsible for making decisions about which data are most worthy of curation, scientists are already, without question, key determinates of the ability to preserve data (Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2008). It is not surprising then that researchers have focused heavily in recent years on the views and data practices of scientists, paying close attention to the factors that underlie sharing and withholding behavior. Much of this research examines the incentives,

norms, and motivations that operate in the scientific context. Before moving on to studies of data sharing and withholding, I discuss this set of factors.

2.3 Incentives, Norms, and Motivations in Science

Many scientists and researchers who study the challenge of data sharing and reuse argue that scientists do not share or create data that will endure past the life of their projects because the academic research incentive structure does not reward such activities (e.g. Association of Research Libraries, 2006; Borgman, 2007). At the same time, several data sharing and preservation proponents use the supposed scientific norm of openness to argue that scientists are obligated to create data as a publicly accessible resource (e.g. Arzberger et al., 2004; Fienberg, Martin, & Straf, 1985; National Academy of Sciences, 2009). To embark on a study of the data practices of scientists, then, one must understand the basic norms, incentives, and motivations at play in the scientific context.

Robert K. Merton, an early sociologist of science, devoted much of his work to understanding the normative structure of science and described the nature of the "ethos of science" through four "institutional imperatives" (Merton, 1973, p. 273). Data sharing and stewardship proponents commonly employ the imperative that Merton called "communism," or the common ownership of goods, to argue that scientists should make their data available to others. In short, according to Merton, property rights in science are kept at a minimum so that more knowledge can be generated from the resulting commons. The maintenance and efficiency of the institution of science is tied to the commons that is the result of scholarly publication.

Although data sharing proponents often use his ideas to argue that science is fundamentally open (e.g. Campbell, Weissman, Causino, & Blumenthal, 2000; Fienberg et al.,

1985), Merton made no claim that maintenance of the commons is necessarily an individual scientist's motivation to openly disseminate her research. Instead, he considered communism (and the other imperatives) to be an institutional norm. Scientists are rewarded for adherence to the norm by the recognition and esteem they receive as a result of publication. The incentive structure, in other words, rewards behavior that adheres, at least partially and in specific circumstances, to openness.

Similarly, Hagstrom (1965), aiming to explain why scientists would openly share the products of their work without direct monetary compensation, characterized the scholarly publication system as a gift exchange. Scientists gift manuscripts to the scientific community—primarily through scholarly publication—and in return they (ideally) accrue status in the form of citation and emulation. The gift exchange, then, is "an exchange of social recognition for information" where one of the few exclusive rights the scientist retains to the information is the right to the recognition for his or her contribution (p. 13). In fact, contribution to the communal system is virtually the *only* means by which a scientist can gain credibility and recognition. According to Hagstrom, such a system creates a moral culture to which scientists not only outwardly adhere, but also internally conform.

Several scholars have criticized the focus on norms to explain scientists' motivation to expend considerable effort on making their discoveries openly available. In their anthropological study of "Laboratory Life," Latour and Woolgar (1986) take issue with Hagstrom's information-recognition-exchange theory. The scientists they studied hardly ever used norms to talk about their behavior. Instead, scientists described their motivations in primarily quasi-economic terms, using economic and business metaphors and characterizing their actions in relation to investment and return. Credit, Latour and Woolgar argue, is not actually a gift given in exchange for

contribution, but a *commodity* that can be exchanged, shared, stolen, accumulated, or wasted.

While credit is an important commodity to scientists, Latour and Woolgar emphasize that *credibility*, or the "attribute of generally being believed in," is scientists' primary motivation because credibility allows them the ability to do science (Oxford English Dictionary as cited by Latour & Woolgar, 1986, p. 194). Scientists' behavior would be more properly understood as that of an investor in capital, where they work to build up their stockpile of credibility. Scientists can then invest that credibility for yet more, or different kinds of, credibility.

Perhaps even more significant to understanding the motivations of scientists, however, Latour and Woolgar's framing of scientific knowledge exchange as a contribution market reveals the demand side of this exchange. While Hagstrom and Merton largely overlook the reason scientists value others' scientific knowledge in the first place, Latour and Woolgar emphasize that information is valued in this market for its utility to generate fresh information.

An important implication of the research outlined thus far, whether scientific publication represents a gift exchange or a commodity market for credibility, is that the system creates a competitive environment in which scientists vie for priority in making discovery and expect community recognition of their contributions. The publication system not only creates competition among scientists to be the first to publish a new finding, but also works as a metric for distributing rewards (e.g. tenure, attractive employment prospects, and grant funding) (National Research Council, 2003). This competitive aspect of publishing indicates that, despite the tendency of scientists to make their findings widely available, scientists do so strategically through a system that rewards them. One might expect, then, that scientists would be reluctant to make their data openly available. At the very least, it seems likely that scientists would protect data and findings until they have turned them into the products for which they are rewarded.

In the next section, I detail research on data sharing and withholding to better understand the specific reasons scientists have for protecting their data and the various factors they consider as they contemplate making data openly available. I argue that while such research has revealed significant aspects of scientists' data practices, it has neglected other potentially important factors.

2.4 Studies of Data Sharing and Withholding

The increasing emphasis on data as a potentially reusable (but perhaps under-shared and/or poorly managed) product of science has resulted in numerous studies that examine scientists' data practices, or the "research processes and activities related to scientists' work with data" (Cragin, Palmer, & Chao, 2010, para. 6). Researchers have studied the practices of scientists as both data *creators* and data *reusers*. While the development and maintenance of data sharing and preservation infrastructures will depend on understanding the needs and practices of secondary reuse (Niu, 2009; Zimmerman, 2003), knowledge that clarifies how scientists work with their data, and particularly those activities that impact future reuse, is equally important. My primary focus is on the latter, and as a result I concentrate on studies of scientists as data creators.

Research on the data sharing and withholding behavior of scientists can be grouped into two main kinds: quantitative studies on rates of data sharing and withholding and the attitudes of scientists regarding these practices; and qualitative studies that examine scientists' sharing and withholding behavior within the context of their day-to-day research activities.

2.4.1 Quantitative Studies on Data Sharing Practices and Attitudes

For all the promised potential of data as a renewable resource for scientific research, the numerous policies implemented to influence scientists' behavior, and an increasingly strong social message about managing data for reuse, many fields of science continue to struggle to build repositories of well-curated collections of data (Borgman, Wallis, & Enyedy, 2007; Interagency Working Group on Digital Data to the National Science and Technology Council, 2009; Nelson, 2009). Data reuse proponents often attribute empty archives to scientists' unwillingness to share data or to take the time and effort required to create data that can be used by others. As a result, several studies have examined rates of, or opinions about, sharing and withholding in an attempt to determine which specific factors are associated with these behaviors and views.

These studies are primarily quantitative and rely on self-reported survey data and measures of repository participation to examine what scientists in various disciplines think about sharing; what their experiences are with sharing and withholding; what the predictors are of sharing and withholding; and to what extent scientists share and withhold. Many of these studies focus their data collection on a specific discipline or resource (e.g. genetics or a specific large-scale database).

Several studies have examined scientists' sharing and withholding behavior and how it relates to journal and funder policies, journal impact factor, and scientists' experience by conducting quantitative analysis of submission rates to various data repositories. In one such study of six journals with explicit policies about sharing genetic sequence information, Noor, Zimmerman, and Teeter (2006) found that no journal had complete compliance with their requirements to deposit sequence data in a public repository. The study revealed that between

3% and 15% of the published studies never submitted DNA sequences; the researchers speculate that this may be the result of journals' lax or non-existent enforcement of their own policies. As Noor et al. point out, this study's lack of direct interaction with the scientists meant that the researchers could not ultimately determine the scientists' reasons for not submitting sequences.

In a similar study, Piwowar and Chapman (2008) found that journal data sharing policy strength had a positive association with data sharing prevalence. Yet even the journals with the strongest policies saw relatively low repository submission rates at 29%. Interested in determining what other factors might be positively associated with data repository submission, Piwowar and Chapman (2009) conducted a small pilot study to look at the relationship between data sharing and funder and publisher requirements, journal impact factor, and investigator experience and impact. Their preliminary findings suggest that the impact factor of a journal and authors' experience were most strongly related to higher rates of data sharing.

The strength of studies that analyze repository submission rates is that they reveal scientists' *actual* data sharing behavior (at least sharing that occurred through the repositories the researchers studied) instead of relying on what scientists say they do. Taken together, this research demonstrates that journal policies, journal impact factor, and investigator experience do positively affect repository submission rates, but on their own do not result in complete (or even overwhelming) compliance. These studies, however, do *not* reveal scientists' reasons for making or not making their data available, nor do they expose whether and why scientists consider some data more appropriate to submit to a repository than others. Perhaps more importantly, these studies are narrowly descriptive, lacking a strong conceptual foundation for explaining the findings or applying them beyond the scope of the studies' limited disciplinary or repository focus.

Several survey studies of scientists' data sharing practices and views exhibit a similar lack of theoretical explanation, but better address the reasons that scientists have for withholding data. Overall, these studies—most of them of geneticists and various other life scientists—show that data withholding is fairly common, but not necessarily rampant. The most common reasons scientists gave for withholding their data were to protect their and their students' ability to fully exploit data first (Blumenthal, Campbell, Anderson, Causino, & Louis, 1997; Campbell et al., 2002; Louis et al., 2002); to avoid the high cost of making data available (Blumenthal et al., 1997; Campbell et al., 2002); and because they were concerned that their contribution would not be reciprocated or recognized by other scientists (Campbell et al., 2002; Louis et al., 2002).

Still, these survey studies do not fully reveal how scientists go about creating data that will have use beyond their original purpose nor do they help us understand the factors that go into a scientist's judgment that a particular set of data is worth the time and effort involved in making it available to others or managing it for continued access. At a more fundamental level, in approaching the study of data practices with the normative assumption that data *should* be shared, these studies (both the survey and the repository submission rate studies) tend to treat any deviation from the ideal as "withholding" due to ineffective punishments and rewards, instead of as possible indications of how scientists view their data's value. As a result, this research is limited in what it can reveal about scientists' data practices and the factors that underlie them.

In an attempt to more fully understand why scientists do not share data or manage them for others' reuse, some researchers have employed qualitative methods such as interviews and observations. These studies help to enhance what we know about scientists' views on data sharing as well as how scientists' data practices are situated within their research activities.

2.4.2 Qualitative Studies of Data Sharing Practices

Characterizations of the life-cycle of data (Borgman, Wallis, Mayernik, et al., 2007; Wallis et al., 2007), data curation (Digital Curation Centre, n.d.), and scientific research (Humphrey, 2006) have been important for identifying the stages that data go through from conception to possible preservation action and the implications of decisions at each stage of this life-cycle on reuse and preservation. All of these depictions place heavy emphasis on the influence that actions taken early in the life-cycle have on preservation and reuse. The actions of data creators throughout their research process are, in other words, vital in determining what data remain viable over time and the purposes to which they can be put. Unsurprisingly, then, many researchers have examined scientists as they carry out their activities to better understand how they work with and manage data, particularly as it relates to making them available to other scientists.

In their study of scientists in earthquake engineering, HIV/AIDS research, and space physics, Birnholtz and Bietz (2003) found that scientists had a number of reasons for not sharing data: they felt it was in their best interest to get as much as possible out of the data themselves first; data documentation required a lot of work for which they were not rewarded; and it was difficult for them to make explicit the tacit knowledge important to understanding the data. Further, the scientists generally thought there should be a period of time in which they were the only ones who could access data they collected. Especially interesting was Birnholtz and Bietz's finding that scientists highly valued data sharing as a means of forming new collaborations.

McCain (1991) examined what factors affected scientists' sharing behavior. Based on data she gathered through unstructured interviews with over twenty geneticists, McCain's scientists reported that they often shared information with those who directly requested it. The

scientists, however, described a number of what they felt were legitimate reasons for refusing access. Scientists thought denying a request to share data was warranted when sharing would threaten the interests of their graduate students; when the proposed work would duplicate their own; or when they lacked the time and money to make the data shareable. McCain's study suggests that, like the Birnholtz and Bietz study, scientists appreciate some degree of personal interaction with potential users of their data when making sharing decisions.

Borgman, Wallis, and Enyedy (2007) studied the data practices of multi-disciplinary teams of ecological scientists. They found that scientists embraced the concept of sharing, but that in practice "little sharing actually occur[ed]" (p. 17). Furthermore, the scientists were more focused on publications as an output of their work than they were on shareable data. Borgman, Wallis, and Enyedy identified several more reasons for the lack of attention scientists paid to creating shareable data: they were not rewarded for data management; making data shareable required significant effort; and they saw no need to use others' data or to share their own data. When asked about the kinds of data they would be willing to share, the scientists said they were much more willing to share data that were already published and least willing to share those that they planned to publish. Some of the scientists also noted that they would be less willing to share data that were difficult to collect.

Tucker (2009) focused her data sharing study on scientists' motivations and emotions related to sharing. Her case study of a large-scale database project at the National Cancer Institute (NCI) revealed that the individual motivations for data sharing often did not align with the stated motivations of the project. For instance, while the stated goal of the NCI cancer biomedical informatics grid was to create a shared repository to aid in scientific discovery, the scientists valued things like research grants or contracts, publications, professional recognition,

being the first to solve a problem (because it advances reputation), patents, and tenure. She points out that while proponents of the grid may think that advanced technology necessitates and enables increased sharing, many of the individual scientists do not have a desire or see a need to share. Because scientists were largely rewarded for individual achievement and data were the means of securing rewards, many of the scientists wanted to extract as much value out of the data as possible before sharing. Far from casting the scientists' self-centered motivations as negative, however, Tucker observes that these scientists' jobs were highly dependent on their own self-interest. Scholarly publication has the potential to garner public recognition, which then strengthens a scientist's ability to procure grant funding. Many of her interviewees described this grant funding as the ultimate goal, since it is what kept them gainfully employed and allowed them to earn a decent salary and conduct their research.

It is clear from these qualitative studies of data sharing behavior that data sharing represents a complex decision based on numerous factors, including how much access to provide, when to provide it, and to whom to provide it. Consequently, the decision to share is frequently not a binary one (share or withhold) but a multi-faceted one that includes non-release, delayed release, and isolated release (Hilgartner, 1997). Taken together, research on data sharing and withholding indicates that many scientists withhold data at least sometimes, despite journal and funding agency policies and pleas to uphold the Mertonian norms of sharing in science. Further, withholding stems from a number of factors, including an incentive structure that values publication over data sharing or data management; scientists' perception that sharing is not necessary; and scientists' concern that their data will be used in ways they do not like or that their sharing will not be reciprocated or credited. These studies show some of the ways in which scientists value data for their own uses, at least in the short-term, and even clarify some of the

factors involved in the decision to share. Many of these studies focus on the issue of sharing vs. withholding; however, this downplays other factors that may be important to understanding what scientists do with data (including making those data available to others). Studies that examine the nature of data and the roles that data play for scientists help to illuminate some other factors.

2.5 The Roles of Data for Scientists

Traditionally, in many scientific fields, once a scientist felt she had fully exploited her data for her own purposes, she tended to treat them as secondary products of research. But technological developments in recent years have led many to characterize this practice as a squandering of resources (particularly when the data are produced using federal dollars) and to emphasize the potential of data to answer many different research questions (e.g. Arzberger et al., 2004; Interagency Working Group on Digital Data to the National Science and Technology Council, 2009). Problematically, however, data sharing and preservation proponents often treat data as inherently shareable and reusable. The following excerpt from a National Research Council report encapsulates such a perspective:

> Data in science are universal—they have the same validity for scientists everywhere. The atomic mass of iron, the structure of DNA, and the amount of rainfall in Manaus in 1972 are facts independent of the political views of their user, the time at which we determine them (apart from the evolving, improving accuracy of the determinations), or the user's location (Committee on Issues in the Transborder Flow of Scientific Data & National Research Council, 1997, p. 48).

Such characterizations of data ignore their complexity and gloss over the amount of work involved in creating data that can be reused by others or that will have longevity. Examinations of data from the STS community offer a more complicated view of data; one that says data are rarely universal and always depend on theoretical and methodological assumptions.

Most STS-oriented definitions emphasize the contingency of data. Borgman (2007), for

example, characterizes data as "reinterpretable representations of information" and asserts that whether data are considered observations or contextual information depends on the use and user of the data (p. 120). The importance of contextual information—made tangible through metadata—in making meaningful use of data reveals data's dependence on instruments, techniques, and theory. Data, according to this perspective, are not just the measurements, but also all the relevant contextual factors that led to the creation of those data. As Bowker argues, there is never such a thing as *pure* data that can "stand outside of time" (Bowker, 2006, p. 177). Even supposedly "raw data" are cooked; the outcome of complex sets of assumptions about the level of precision required and what counts as "noise," to name just two (Gitelman, 2013).

Hilgartner and Brandt-Rauf (1994) characterize data as an evolving stream rather than easily defined end products or isolated entities and, like other researchers, draw attention to the heterogeneous elements that make up data. In particular, they emphasize that instruments, techniques, and written inscriptions, and both data inputs and outputs are important components of scientific data.

Several researchers have examined the various roles that data play for scientists. These researchers are less concerned with defining data than they are in uncovering what scientists use data to do. An especially influential conception of data in STS research is that data are laboratory inscriptions; surrogates for some aspect(s) of objects (Latour, 1987; Latour & Woolgar, 1986). Latour observes that a guinea pig cannot by itself tell a scientist much until it is turned into text or data. The data, unlike the guinea pig they are derived from, can be "combine[d], compare[d], summarize[d] [. . .] and manipulate[d]" to create knowledge (Van House, 2004, p. 14). Further, the *guinea pig as data* can be "accessed, transferred, calculated, processed and analyzed in seconds" (Hine, 2006, p. 131).

In addition to serving as surrogates for objects, data also play a significant role as evidence for or against a particular hypothesis or theory (Latour & Woolgar, 1986). According to Latour and Woolgar, data are most effective in their persuasive role when all signs of their production are hidden. While Amman and Knorr Cetina (1988) caution against equating data with evidence, they acknowledge data's role in creating evidence.

Other researchers have emphasized data's role for scientists as material goods or commodities. Federal funders and data sharing and preservation proponents have argued that data generated with federal dollars are, by definition, public goods (e.g. Arzberger et al., 2004; Interagency Working Group on Digital Data to the National Science and Technology Council, 2009; J. Reichman & Uhlir, 2001). And, in some cases, scientists seem to agree with this principle, believing that data are meant to be shared (Bietz & Lee, 2009; Vertesi & Dourish, 2011). However, in many instances, scientists treat data as though they are private goods to be used by them, as they see fit (Tucker, 2009; Vertesi & Dourish, 2011).

As a product of a significant investment of resources and effort (Borgman, 2007; Tucker, 2009), data represent potential payoff in terms of material or professional gain (Tucker, 2009) or credibility (Latour & Woolgar, 1986). They are assets that scientists are often driven to fully exploit (Birnholtz & Bietz, 2003) and strategically control to gain competitive edge or to use in exchange for other resources (Hilgartner, 1997; Hilgartner & Brandt-Rauf, 1994). Because of their potential market value, scientists can use data as bargaining chips in forming collaborations. Data can help a scientist gain entry into a collaboration, but if those data are unique she is also in a better position to define the terms of any collaboration that involves those data (Hilgartner, 1997). Going further, Birnholtz and Bietz observe that a scientist with unique and valued data and the tacit knowledge to make use of them can effectively become a gatekeeper for others in

her field of study (2003).

As this research shows, even while funders increasingly expect data to be "end[s] in themselves" (i.e. reusable products of scientific research), data are, for scientists, "means to ends" (e.g. publishing papers, gaining credibility, etc.) (Hine, 2006, p. 11). Because of data's contingency on theory, instruments, and techniques, assumptions about what data are *for* necessarily undergird scientists' data practices. We know that for data to have continuing value to others, they must be accompanied by metadata that describe their context of creation (Borgman, Wallis, & Enyedy, 2007; Zimmerman, 2003). However, choices about which contextual information to document will always constrain possible uses for those data. Further, data may be surrogates for objects, but how and what they represent involves choices about equipment, theory, and assumptions that are often taken for granted by the scientists who created the data.

Much of the current research on data practices in science focuses on incentives for motivating scientists to share data or manage data in ways that bolster longevity, with relatively little attention paid to how and why data are (or are not) viewed as objects of continuing value. Karasti et al.'s (2006) study of data stewardship at the Long Term Ecological Research (LTER) network represents one notable exception, however it focused primarily on how information managers created long-term data, not on the views and behaviors of scientists. As a result, we know a fair amount about the motivational obstacles that prevent scientists from distributing their data widely, but lack a clear understanding of how scientists evaluate data's value and whether and in what ways the evaluation influences their data practices.

Empirical work suggests scientists' conceptions of data's value influence what they do with data. In a survey study, for example, social scientists indicated that they would be more likely to create archive-ready data if they thought those data would be of use to others (Hedstrom

& Niu, 2008). In another study, scientists regarded data that were expected to have the greatest potential for generating new results as the most shareable (Cragin, Palmer, Carlson, et al., 2010). These findings conflict with other studies, however, that revealed scientists are *less* likely to share data that they think are of high value (Tucker, 2009) or that are hard-won or difficult to generate (Borgman, Wallis, & Enyedy, 2007). Adding additional complexity to understanding scientists' sharing and withholding practices, Vertesi and Dourish (2011) found that the shareability of data depended on the context of data's production: "Data [. . .] is not a shared resource simply because it is data: it is a shared resource because it is crafted that way from the outset" (p. 540).

2.6 Studies that Address Scientists' Conceptions of Data's Value

Despite the emphasis that is placed on the value of scientific data, researchers interested in scientific data practices have largely ignored scientists' conceptions of data's value, with a few exceptions. In one study—a survey of scientists from four UK universities—researchers sought to understand scientists' long-term storage needs (Beagrie, Beagrie, & Rowlands, 2009). Beagrie et al. reported that scientists indicated a strong need for long-term (>5 years) and medium term (1-5 years) data curation and preservation services. The study showed less demand for short-term data storage (1-12 months). Additionally, scientists in this study said that they expected almost half of their data to have a useful life of less than ten years, while they expected 27% of their data to have an indefinite period of value. This research, however, does not specify value to what purpose nor does it reveal why some data were considered to be more valuable for long-term use than others.

In another study, also aimed at understanding the services scientists need from institutional repositories, scientists in several domains indicated the importance of long preservation periods for their data (Cragin, Palmer, Carlson, et al., 2010). The largest number (13 of 20) indicated that their data would remain valuable for reuse for a minimum of ten years; four indicated reuse value extended indefinitely, while the rest estimated data would be valuable for reuse for between three and ten years. In this study, scientists generally considered observational data to have very long-term value, but the study design (interviews with twenty individual scientists, picked from a range of disciplines, with only one or two scientists representing each discipline) makes it difficult to draw conclusions about what factors were important in scientists' consideration of their data's continuing value.

A recent ethnographic study of space mission scientists' data practices found that datasets have "fundamentally different values" depending on their context of production (Vertesi & Dourish, 2011, p. 540). Specifically, Vertesi and Dourish found that whether or not scientists thought of their data as a public good, to be shared, or a private one, to be protected and exploited for their own purposes, depended on how those data were crafted in practice from the outset of scientific work. An important point worth noting is that the differences in the assignment of value of data as public or private goods did not depend exclusively on disciplinary boundaries. While a good deal of research on scientific data practices focuses on elucidating disciplinary differences, Vertesi and Dourish's study reveals that how data are valued goes deeper than disciplinary orientation and depends a great deal on the local context of data production.

While the Vertesi and Dourish study demonstrates how data's value as a shareable public good is constructed in practice and specifically how value depends on assumptions made early in

data's life, the study does not deal with how scientists might assign differing value to different data and the specific dimensions of value that inform that assessment. The survey and small interview studies do somewhat explore scientists' conceptions of their data's continuing value, but not in a way that makes clear how scientists arrive at these judgments or the way value judgments are reflected in their data practices.

Valuation is at the heart of ensuring continuing access to any materials. The Blue Ribbon Task Force on Sustainable Digital Preservation and Access (2010) broadly defines the value of preserved data as the benefits that are gained as a result of having access to information in the future, but the specifics that such a broad definition leaves out comprise several important qualifying dimensions of value. Specific dimensions include: value for what purposes? For whom? For how long? Further, what factors determine selectors' assessment of the value of data? —The difficulty of recreating data? Data's expected level of usage? Data's source of creation or collection? These are some of the dimensions of value I explored as I studied scientists' conceptions of their data's value.

2.7 Conclusion

As scientific data are increasingly expected to serve as products of science, it becomes ever more important to understand scientists' data practices. While a good deal of research has been conducted that reveals the motivations and incentives underlying sharing and withholding behavior, such work neglects other factors that might influence decisions to make data widely available or to create data that endure beyond their immediate collection. Scientists' conceptions of their data's value has emerged from several data practices studies as an important influence on what scientists do with data, but results are conflicting and do not show how or why scientists

come to view some of their data as having longer term value than others. In order to fully understand scientists' data practices, an explanatory framework that takes into account scientists' conceptions of data's value must be developed. In the next chapter I describe a study designed to address this need.

CHAPTER 3: RESEARCH DESIGN

In this chapter, I describe my research design, including the rationale for it, the site of the study, the case selection rationale, as well as the kinds of data I collected and the methods I used to analyze those data. I briefly introduce the cases that are the focus of this study: three small teams of scientists (from several ecology-related disciplines) that carried out work at a university-run field station (what I refer to as "the Station" throughout the dissertation) during the summer of 2012. In Chapter 4, I provide a more detailed description of the teams I studied and give examples of their valuation activities before I present my findings in Chapters 5 and 6.

3.1 RATIONALE FOR A QUALITATIVE, MULTI-CASE STUDY

This dissertation explores scientific data practices through the lens of scientists' conceptions of their data's value. I employed a qualitative, embedded, multi-case study approach to answer the following overarching research question:

How do scientists conceive of the value of their data, and how do scientists enact conceptions of value in their data practices?

This question is made up of several related subsidiary questions:

- What specific uses for data are salient to scientists (e.g. as evidence of claims; resources for conducting longitudinal studies; inputs for new research questions)?
- What time spans do scientists use to think about their data's value?
- On what basis do scientists assess data's value? For example, do they consider the speed

with which technological advances are expected to render data collected with a particular piece of equipment obsolete; the processing state of the data; the ease of replicating the data, etc.?

- How do scientists create data that are valuable (as construed by the scientists themselves), and what do they do in response to their notions of data's value?

My research questions focus on understanding practices that are situated within daily activities and are framed by a theoretical orientation that emphasizes the meanings data have for scientists. Therefore, like several other studies on scientific data practices (e.g. Akmon, Zimmerman, Daniels, & Hedstrom, 2011; Birnholtz & Bietz, 2003; Borgman, Wallis, & Enyedy, 2007), this study relies on qualitative data. I gathered data through participant observations of scientists while they collected and worked with data; and semi-structured interviews with scientists about their experiences working with data and their impressions of their data's value. More specifically, I designed this research as a qualitative, embedded, multi-case study (Yin, 1994) of three small teams of scientists who carried out research and lived at the Station during the summer of 2012.

The case study research strategy emphasizes "contemporary phenomenon within its real-life context," bounded by time and space (Yin, 1994, p. 13). This research approach, which includes a variety of data collection strategies, is particularly effective for addressing *how* questions. A multiple-case study design allows one to study the specifics of several cases in order to generate a richer understanding of the main phenomenon of interest. For this reason, the approach is often considered more rigorous and the findings more compelling than would be those from a single case, and it allows for identification of patterns and themes that cut across cases (Miles & Huberman, 1994; Yin, 2009). In studying multiple cases, my aim is not to make

generalizations to all scientists or all ecologists. As methodologists emphasize, case studies (even multiple case studies) do not facilitate generalization to some larger population. Instead, I intend to use the three cases to illuminate, in a deeply contextual sense, *these* scientists' conceptions of data's value and how conceptions of value were reflected in what they did with data.

3.2 Study Site

This study centers on the data practices and views of three small teams of scientists engaged in research at a university-run, field station in the United States. The Station was founded in the early 20th century to support ecological research and education and continues to serve both researchers and students today. Every year, approximately 110 scientists (primarily working in small teams), 30 teaching faculty and assistants, and 100 undergraduate students live, work, and study at the Station. As a temporary living facility, the Station offers a cafeteria, dormitories, shared rustic cabins, and small family houses on its campus, which lies on a lake in a remote and sparsely populated location.

Scientists come from around the U.S. to take advantage of the Station's research facilities and field sites, and, along with the students and teaching faculty, they live onsite as they conduct their research. With 10,000 acres of land, the Station encompasses a number of habitats of interest to researchers, including forests of aspen, pine, northern hardwoods, and conifer swamps; fields and meadows; pine plains; wetlands; and rivers and streams. In addition to the various sites for conducting field research, the Station also offers specialized research facilities, such as an artificial stream lab and atmospheric towers; a fee-based analytical lab equipped to perform analyses of air, soil, and water samples; and laboratory space and equipment that scientists can use to process different kinds of samples and perform their own analyses.

The Station hosts a range of scientists, representing a variety of ecology-related disciplines. During the summer of 2012, this included atmospheric science, forestry, biology, zoology, ornithology, biogeochemistry, and wetland ecology, to name a few. As is common in ecological research (Michener & Brunt, 2000), small teams of scientists carry out the majority of funded research at the Station. During the 2012 field season (from May to August), the Station was home to 19 funded research projects. Of these, 14 were carried out by teams of 2 to 12 people. The following organizations funded 2012 Station projects: the U.S. National Science Foundation (NSF), the Department of Energy (DoE), the Environmental Protection Agency (EPA), the Defense Advanced Research Projects Agency (DARPA), the National Aeronautics and Space Administration (NASA), the Fish and Wildlife Service (FWS), the National Oceanic and Atmospheric Administration (NOAA), the State Department of Natural Resources (DNR); as well as several North American universities.

The Station offered a number of compelling advantages for conducting my study. First, as just outlined, it attracts scientists who come from a range of specialties, rely on funding from a diverse mix of organizations, and collect a variety of different types of data. This presented a unique chance to study conceptions of value from a number of perspectives. Second, the fact that the scientists resided (as I did) at the same location where they conducted their research offered a rare level of access to them and their work. Not only did I interview and observe scientists while they carried out their research, but I also ate meals and socialized with them on a daily basis. This greatly enhanced my ability to understand and observe the real-life contexts in which data practices and evaluation of data occur; a key component of qualitative inquiry (Lincoln & Guba, 1985). Finally, in the last couple of years the Station has begun implementing policies and building an infrastructure for data sharing and archiving. While still in its infancy, this local

policy environment and infrastructure to support long-term data archiving presented the opportunity to study a diverse set of scientists working within a shared, local context in which staff and administrators were making a case for the value of data beyond individual projects.

The Station implemented its current data management policies in 2010. The reasons were twofold: address a local problem of data inaccessibility; and be ready to meet anticipated funding mandates that would require attention to data management. At the time of this study, the policy required scientists to submit a "properly documented" set of data collected using Station resources to the Station within a year following the completion of a dataset. Scientists were permitted to deposit the data with a different public repository, but, if they made use of this option, they had to submit a metadata file (essentially a project abstract) to the Station so that staff could provide a link from the project's description to the dataset.

The Station employed one information manager who was responsible for maintaining a publically accessible database of current and past research descriptions, a list of publications that resulted from Station research, and datasets generated using Station facilities. In the summer of 2012, the information manager was beginning to approach faculty researchers to identify datasets and work with them to archive data. The information manager, however, had no role in project-level data management, and, aside from local server space, the Station did not provide infrastructure for managing data during the active portion of their lifecycle. It was up to researchers and teams of researchers to devise methods of managing data for their own uses.

The scientists I studied collected and worked with data for their research within the local setting I just described. This overarching, shared context for the scientists' work included the Station's data sharing and archiving mandate, a set of potential resources for conducting research, and a span of time for carrying out work and living at the Station that meant they ate meals

together, socialized, and generally were members of the Station community. At the same time, teams of scientists engaged in independent projects, with needs and a context that was unique to their projects. The following section describes my study's participants and my participant selection process and rationale.

3.3 Study Participants

3.3.1 Rationale and Method of Participant Selection

Of critical importance to successfully carrying out case study research is appropriately defining the unit of analysis, or the case, to be studied. This study focuses on the views and activities of scientists, with each of the three teams I studied constituting a case and individual scientists comprising embedded units within each case (Yin, 2009). Methodologists offer several strategies for selecting cases for a multi-case study. Making a comparison to experimental research design, Yin (2009) views replication as the main strength of multiple case studies. As a result, he argues cases should be selected based on either literal replication (similar results are expected) or theoretical replication (contrasting results are expected but for "anticipatable reasons") (Yin, 2009, p. 54). Stake (2006) recommends that cases should be selected for their anticipated ability to help the researcher better understand the phenomenon in question. He presents three main criteria for case selection: relevant to the phenomena or thing studied; representative of diversity across contexts; and provides opportunities to learn about complexity and contexts. Both Stake and Yin emphasize that a sampling logic to case selection is inappropriate. In other words, case study research, even those involving multiple cases, should never strive to represent an entire population through case selection.

The time period of greatest research activity (i.e. the greatest number of research projects underway) at the Station occurs each year during what Station staff call the "full summer session." In 2012, the full summer session took place between June 21 and August 21. For this reason, I conducted most data gathering activities during that same time period.[4] In 2012, eight teams conducted research at the Station during the full summer session. When I arrived at the Station, I had reviewed descriptions of the eight projects hosted there, but I had not made any prior selection of cases. During the first week at the Station, I scheduled brief (approximately 15-minute) meetings with PIs from each of the eight projects. The goal in these meetings was to learn more about each of the projects, including its composition, the types of data collected, the length of the project, and any data management mandates required by the project's funders. I also assessed the group's willingness to participate in my study. I quickly ruled out two groups. One was unsuitable because studying the group would have required extended travel away from the Station; I excluded another because the PI was unable to participate in the screening interview.

From the remaining six groups, I selected three that contrasted along four main dimensions: career stage of the PI(s), the methodological approach to research, the length of the study, and the funding source for the study. I chose these characteristics as selection criteria based on prior research that suggests these factors play a role in scientists' sharing and archiving activities; and on assumptions and claims that have been made by data sharing and preservation proponents. While I covered this literature in detail earlier in the dissertation, I briefly cite it again here as justification for selecting cases based on these four dimensions.

[4] One of the teams split up their data collection activities between sites at the Station and near their home institution. I interviewed and observed them primarily at the Station, but also gathered additional data from them in the fall, after I left the Station.

Piwowar and Chapman (2009) found "investigator experience" to be positively associated with data sharing behavior. They noted in their analysis that they were unsure whether the influence of investigator experience had more to do with tenure status, age, or previous experience sharing data. However, the fact that data play such a strong role in creating the products on which tenure decisions are in large part based suggested that I should study groups that differed according to the career stage of project PIs. Two of the teams I selected were led by senior PIs (all had tenure and had been publishing articles for at least 20 years); one team was led by two junior faculty PIs, who did not yet have tenure and who had been publishing for less than 10 years.

Some data sharing proponents have argued that assessments of data's long-term value should be based on data type. In making this argument, they claim that observational data have longer-term value than experimental data by virtue of their uniqueness and irreproducibility (Simberloff et al., 2005). Further, empirical work suggests that scientists assess the length of their data's value on the basis of whether those data are observational in nature or drawn from more highly controlled experiments (Cragin, Palmer, Carlson, et al., 2010). This led me to consider that a team's methodological approach to gathering data might influence the scientists' notions of data's value. None of the teams I selected carried out purely observational work, where no manipulation of a study system was involved. One team, however, relied on data collected from a field study (some aspects of the environment were intentionally manipulated, and then differences between manipulated and control plots were observed); one team relied on a highly controlled experiment; and one team collected data from modeling simulations and a mesocosm experiment (mesocosm studies are more controlled than field studies, yet less controlled than laboratory studies).

The teams I selected also varied according to the length of their studies, with the presupposition that longer-term studies might engender a longer-term view of the value of data. Two of the teams were engaged in three-year studies, and the other team was carrying out a study lasting only one field season (approximately eight weeks).

Lastly, I selected teams that varied according to their funding source. Several federal funders, including the NSF, NIH, and NASA, have implemented data sharing and/or data management plan requirements under the implicit assumption (or hope) that such requirements will influence scientists' data management and sharing behavior. Preliminary research indicates that funder requirements have very minor influence on whether scientists share or manage their data for reuse and archiving (Piwowar & Chapman, 2009). However, I felt it was worth selecting projects supported by funders that required data management or sharing planning as well as those that did not to see if such requirements came up in scientists' characterization of data's value or their data sharing and archiving plans. One of the team's projects was subject to data sharing plan mandates required by NASA; while two of the teams were subject to no requirement to write a data management plan or archive data (aside, of course, from the Station requirement, which all Station projects were subject to).

Prior to working with a group, I obtained a signed consent form from each group member. The consent form informed the participants of my study and its exemption from IRB oversight and also stipulated that I would conceal identifying information in products of the research. I also let participants know that they could end their participation in the study at any time. I briefly describe each team below (I describe the teams in more detail in Chapter 4).

3.3.2 Brief Description of Each Team

The Invasives Management (IM) Team was at the Station in its second year of a three-year, EPA-funded study to test environmentally sustainable methods of restoring wetland biodiversity. More specifically, this team was engaged in a heavily applied field experiment to study the effectiveness of different methods of removal of invasive *Typha*,[5] which is highly destructive to coastal wetlands; and to assess the potential of using the removed biomass as a form of clean, renewable energy. The team was comprised of seven people, most of whom had been coming to the Station for several years (and in some cases decades) to conduct research. The grant proposal lists three PIs. One of the PIs was a tenured, full professor and vice provost at her university. While she played a large role in the study's conception, she had relatively little involvement in collecting and working with data, visiting the Station only briefly on two occasions during the summer. Another PI, an assistant professor, served in an advisory capacity for the project. During the time of my study, he had no involvement in the day-to-day activities of the project and was instead teaching courses at the Station. I did not include him in observations or interviews. The project was primarily led by a non-tenure-track, research faculty PI. The other members consisted of one post-doctoral fellow who was moving on to a tenure-track position in the fall at another university; one master's student in her final year of study; one undergraduate; and one hired research assistant (Table 3.1).

[5] *Typha* is the genus name for several species of plant commonly known as "cattail."

Table 3.1: Invasives Management Team Members

Name[6]	Career Stage	Project Role
Evelyn	Tenured full professor	Primary investigator
Phil	Non-tenured assistant professor	Primary investigator
Matt	Research faculty	Primary investigator
Amy	Post-doctoral fellow	Post-doctoral researcher
Renee	Graduate student (master's)	Graduate researcher
Dylan	Post-bachelor's degree research staff	Research assistant
Brooke	Undergraduate student	Undergraduate researcher

Aside from the Station's requirement to submit completed datasets to the Station's repository within a year of the project's conclusion, the team was not subject to any other data management or sharing mandates. The EPA does not currently require data sharing or data management plans; however, the team stipulated in the grant application that they would deliver a completed database in the final report of the project to the funder.

The Nutrient Uptake in Streams (NUS) Team, led by two junior, non-tenured PIs, was at the Station to carry out a controlled experiment on the effect of leaf litter on nutrient (in particular nitrate, ammonium, and phosphorus) uptake in streams. The research was funded by a subaward of a university-administered NSF ADVANCE Grant[7] program at one of the PIs universities. The team was comprised of five people: the two PIs, from separate institutions, who had a history of working together on other projects; two undergraduate students, who had never before conducted research; and one master's student, working relatively independently on a sub-project (Table 3.2). This was the first time that any of the members in the NUS Team conducted research at the Station.

[6] Pseudonyms are used to conceal participants' identities.
[7] NSF ADVANCE Grants support women conducting research in science and engineering fields. See http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5383 for more information. Retrieved November 8, 2012.

Table 3.2: Nutrient Uptake in Streams Team Members

Name	Career Stage	Project Role
Elizabeth	Non-tenured assistant professor	Primary investigator
Jessica	Non-tenured assistant professor	Primary investigator
Tina	Graduate student (master's)	Graduate researcher
Carolyn	Undergraduate student	Undergraduate researcher
Janet	Undergraduate student	Undergraduate researcher

In terms of data mandates relevant to this team's data, because the ADVANCE grant was awarded to Jessica's university in 2008, NSF's data management plan requirement (implemented in early 2011) did not apply to this team's work. Furthermore, the university that administered awards from this grant to individual researchers did not require any data management plan nor did it require grantees to share or archive data.

The Invasion Dynamics and Modeling (ID&M) Team was at the Station in the second year of a three-year, mesocosm study. A mesocosm is an enclosure that scientists use to bring the environment into semi-controlled conditions. In this case, each mesocosm consisted of a metal tank measuring six feet in diameter, which the researchers filled with soil and populated with various types of plants. Similar to the IM Team, this group was studying wetland invasion, however the scientists were focusing on understanding how nutrient dynamics affect the ability of two problematic invasive plants—*Typha* and *Phragmites*—to succeed in wetlands. Additionally, the ID&M Team was working to further develop and test a computational ecosystem model. The researchers intended their model to be used to generate hypotheses about potential futures in wetlands given particular sets of circumstances (such as the level of nutrient runoff into a wetland).

The team was made up of five scientists: two co-PIs who were tenured professors; two post-doctoral fellows; and one lab assistant hired for the summer (Table 3.3). All but one of the

members of the ID&M Team had conducted research at the Station prior to the summer of 2012; the team's PIs had been coming to the Station for over ten years to conduct research.

Table 3.3: Invasion Dynamics & Modeling Team Members

Name	Career Stage	Project Role
Kate	Tenured full professor	Co-primary investigator
Mark	Tenured associate professor	Co-primary investigator
Ethan	Post-doctoral fellow	Post-doctoral researcher
Gabe	Post-doctoral fellow	Post-doctoral researcher
Chad	Post-bachelor's degree research staff	Research assistant

The ID&M Team's project was part of a larger NASA-funded project (awarded in 2010) with PIs from two other universities. Scientists at the other universities were developing models of their own—using satellite imagery—that were intended to depict land use changes and hydrological flows in the state. The project was a collaborative one in which scientists from each institution planned to share their data with the larger group. At the time of my study, however, work proceeded for the most part independently at the different institutions. My observations and interviews included only members who were part of the team conducting work at the Station during summer 2012.

NASA requires applicants to include a data sharing plan in their proposals (National Aeronautics and Space Administration, 2009), and this team's proposal specified that the researchers would make the project's data available to the scientific community (either on a website or through a NASA-developed repository) and create a public website to host invasion maps as well as a user-friendly version of the ecosystem model.

Before moving on to the next section where I describe the data I collected and my methods of analysis, I have summarized the three teams that are the focus of this dissertation in Table 3.4.

Table 3.4: Summary of the Three Teams Studied

	PI(s) Career Level	Study Inquiry Method(s)	Funder Data Mandates	Project Length
NUS	2 non-tenured, assistant professors	Controlled experiment	No (NSF Subaward)	8 weeks
IM	1 tenured full professor, 1 non-tenured assistant professor, and 1 non-tenure-track research faculty	Field experiment and mesocosm experiment	No (EPA)	3 years
ID&M	1 tenured full professor and 1 tenured associate professor	Mesocosm experiment and modeling	Yes (NASA)	3 years

3.4 Data Collection

Scientists from each of the three teams represent the main source of data for the study. The data come primarily from semi-structured interviews with scientists and participant observations of scientists as they collected and worked with their data. I also interviewed key Station staff: the director, information manager, and resident biologist. In addition to interviews and observations, this study also relies on various documentary sources, including grant applications, data, and metadata. I gathered most of the data for this research from June 25 to August 17, 2012 when I was living at the Station.

3.4.1 Participant Observations

Lincoln and Guba argue that observations

> maximize the inquirer's ability to grasp the motives, beliefs, concerns, interests, [and] unconscious behaviors [. . .] to see the world as his subjects see it [. . .] to capture phenomenon in and on its own terms [. . .] (1985, p. 273).

My interest in understanding the meanings that data have for scientists and their data practices made observations of scientists as they worked with data a crucially important part of

my research design. I conducted observations as a participant observer; I helped to record data, used instruments to get automated readings, measured and counted plants, sorted invertebrates from water samples, set up experimental stream channels, prepared soil samples for analysis, and assisted with carrying and setting up equipment and maintaining study setups. The length of the participant observation sessions ranged from a couple of hours to a two-day stay with scientists at a more remote field site. Most of the observations began and concluded with the workday for the researchers (typically from about 8am to 5pm).

Observations focused on activities aimed at creating good data (as conceived by the scientists) that would be accessible and understandable past the immediate point of the data's creation. I was particularly concerned with uncovering when and how scientists' conceptions of the value of data influenced how they collected, documented, managed, and worked with data. Based on Hilgartner and Brandt-Rauf's (1994) data stream model, as well as subsequent research on the life cycle of data by Wallis et al. (2008), I expected data valuation to occur across the research process, including during data collection, processing, and analysis. The scientists I studied were primarily engaged in data collection; though I also observed some processing and analysis activities in all of the groups.

My initial observations were relatively unstructured and concerned with clarifying my understanding of participants' work practices at a general level. As the study progressed and I understood more about how they created, worked with, and managed data, my observations became more focused on understanding the influences that the scientists' conceptions of data's value had on how they collected data and organized data for future use. For example, my early observations revealed that preformatted datasheets were a key tool for ensuring the most important data were captured and that data collection was consistent across the time span of the

study. In later observations, I focused significant attention on how the sheets were created, altered, and maintained.

The main information that I captured in the observations was:

- The points in the research process at which scientists assessed data's value
- The activities that scientists engaged in to make data that were useful, accessible, and understandable past data's immediate point of collection
- How scientists worked to create data intended for particular purposes and uses and what the salient uses were

Altogether, I conducted more than 130 hours of observation, divided roughly equally between the three groups (Table 3.5). The timing of my observations of each group was highly contingent on what the group was doing. For example, while the NUS Team was collecting and working with data the entire time I was at the Station, the IM Team had off-weeks where team members were mostly gathering and organizing equipment. Members of the ID&M Team only came to the Station some weeks that I was there, as they split their time between the Station and another study site. My goal was to spend a total of two working weeks with each of the groups: one week each for three weeks, and then a repetition of the first three weeks. I allowed a week in between the cycles to go through my data and consider how I would focus observations in the final three weeks.

Table 3.5: Observation and Interview Time

Team	Observations	Semi-Structured Interviews	n
NUS	43 hours	4 hours	5
IM	48 hours	5 hours	6
ID&M	41 hours	7 hours	5
Station Staff	n/a	2 hours	3
Total	**132**	**18 hours**	**19**

I was able to adhere to this plan with two main deviations. I took the off-days earlier, and completed observations and interviews with ID&M Team members at their home institution in the fall (Table 3.6).

Table 3.6: Timeline of Observations

	Week 1	**Week 2**	**Week 3**	**Week 4**	**Week 5**	**Week 6**	**Week 7**	**Week 8**
Recruiting								
NUS Team								
IM Team								
ID&M Team								

I captured observations with field notes that I recorded in a notebook while I was with the scientists. At the end of every day, I transcribed and expanded on my notes in a Word document. My observational notes make up 58 single-spaced pages of text. I also took numerous photographs of the participants as they collected and worked with samples and data.

3.4.2 Semi-Structured Interviews with Scientists

Qualitative interviews represent an effective method for understanding processes and their antecedents; generating holistic descriptions that represent the complexity of real life events; and learning how people interpret events (Weiss, 1994). My interviews with scientists in this study served several purposes: an orientation to the type of research they conducted and their planned work at the Station; a description, in the scientists' own words, about the purposes they conceived for their data and the aims their data practices strived to serve; clarification of my observations of scientists as they worked; and member checking.

I interviewed all of the members of each team that I studied, except for one of the PIs for the IM Team, who was unavailable and had little involvement with the team's data. I interviewed each scientist at least once; some of them I interviewed two or three times to accommodate their

schedules and to allow me the opportunity to ask follow-up questions. Total interview time with each participant varied from 15 minutes to two hours. I spent a total of 16 hours interviewing scientists (Table 3.5). With the permission of participants, I recorded these interviews and submitted them to Scribie[8] for transcription. I proofread all of the transcripts as I listened to the recordings. These transcripts make up 336 single-spaced pages of text, which I uploaded into NVivo9 for analysis.

The interviews with researchers served as a significant resource for understanding scientists' data practices as they conceived them, as well as their own interpretations of the uses for data and what was needed to make data that served those uses. Further, I used the interviews to gain insights into how scientists thought about their data's potential continuing value to others, including how long data would be valuable and for what purposes.

In initial interviews with scientists, I used a semi-structured interview guide. In the interviews that followed, I based the interview questions on previous answers and things I learned or observed as I studied the group. I asked scientists a range of questions about (See Appendix 1 for Interview Guide):

- Their scientific work (both past work and planned future work) and the kinds of data they collected
- The process by which they collected and managed data
- The uses they anticipated for their data
- Their experience making data available for others' uses and their plans for making their current project's data available to others
- The characteristics that made for what the scientists considered to be good data

[8] Scribie, http://scribie.com/. Retrieved June 19, 2013.

3.4.3 Semi-Structured Interviews with Station Staff

I interviewed Station staff (by which I mean those responsible for the operation of the Station as a whole, including the data repository) to understand the Station context, including the data management resources offered to Station scientists, the messages communicated to them about the value of their data to others, and the staff's view on the kinds of data that were good candidates for archiving. These interviews focused on a varied set of questions depending on whom I interviewed (for example, the questions I asked the information manager differed from those I asked the director). The information I collected from these interviews more specifically was:

- The requirements the Station placed on scientists with regards to their data
- The training Station staff provided scientists in data collection and management
- The kinds of data Station staff thought belonged in the repository and what they thought the data would be valuable for
- Level of participation and compliance by scientists in contributing data

In total, I spent two hours interviewing three Station staff members. Transcripts of these interviews total 24 pages of text.

3.4.4 Documentary Sources

To understand the contextual factors that influenced the scientists' conceptions of value as well as to more fully understand scientists' data practices, I collected a range of documentary sources. These included relevant journal policies, funding policies, and grant proposals. I also collected datasheets, samples of participants' data and metadata, grant proposals, and data quality control plans.

3.5 Data Analysis

My data were in the form of text: transcripts of interviews with participants, observational field notes, and documents that I gathered from the participants. My analysis of these data, including coding, began with the first data I collected and ran throughout the project. Early analysis helped me to target observations and formulate additional interview questions.

I used the qualitative data management tool, NVivo9, to facilitate coding and analysis. My strategy was to use an inductive, open coding approach to analyzing the interview data with scientists and staff (Miles & Huberman, 1994) and the observations of scientists to identify important themes and concepts that emerged from the data. While I approached my first coding efforts with a provisional list of codes I developed based on the dissertation's research questions and theoretical framework, I developed and altered this coding scheme as themes emerged during data collection and analysis.

I created document summary forms to capture the significant and most important aspects of the documentary sources (e.g. grant applications, data collection sheets, and spreadsheets of data) I collected (Miles & Huberman, 1994). I was mostly interested in these sources as a means of understanding the context in which the participants carried out data practices, not as a major data point in and of themselves. With regard to the data and metadata samples, I compared differences in data's documentation across different kinds of data to determine how conceptions of value were translated into practices.

In addition, during my data analysis I regularly wrote memos that reflected on emergent themes, concepts, and relationships that arose. These were an effective tool for consolidating data at a more conceptual level as well as for identifying areas of the study and its conceptual framework that needed to be altered as the study progressed (Miles & Huberman, 1994).

Lastly, I conducted both within and across-case analysis to uncover similarities and differences across the three cases as well as across scientists within the cases (Yin, 2009). In doing so, I sought to generate locally valid explanations for how scientists valued and managed their data to serve particular uses (Maxwell, 1996).

3.6 VALIDITY

The concept of validity has its roots in the positivist tradition of inquiry and refers to the quality of accurately depicting reality (Creswell, 1994). Researchers who work within a qualitative paradigm often take issue with the application of conventional standards of validity to more naturalistic research, arguing that reality is contingent on the context(s) in which it is observed (Erlandson, Harris, Skipper, & Allen, 1993). As a result, some methodologists propose replacing the term "valid" with "trustworthy" and argue that trustworthy research is built on credibility attained primarily through prolonged engagement, triangulation between sources and methods, and member checks (Creswell, 1994; Lincoln & Guba, 1985). I used each of these strategies to ensure a fair and accurate account.

An important consideration in selecting the Station as the site of this study was that it presented the opportunity for intense, prolonged engagement with participants. As a facility where researchers both lived and worked, the Station allowed me to become, in many ways, part of the community of those I was studying. My status as a non-ecological scientist studying data practices clearly set me apart from other scientists at the Station; I was, however, part of the day-to-day life there. Like the other researchers, I lived in a Station cabin, ate all my meals in the community cafeteria, and attended social events and research talks. I quickly found that researchers welcomed my willingness to serve as an assistant, a role that allowed me "work with

people day in and day out" (Fetterman, 1989 as quoted by Creswell, 1994, p. 201), furthering my ability to gain an insider perspective on the scientists' views and practices.

I triangulated data that I gathered across sources and methods. I paid careful attention to discrepancies and confirmations across interviews and observations and across participants in a team. When I found discrepancies, I investigated the circumstances that might explain the discrepancy and followed up with participants as needed.

Lastly, I conducted member checks of my findings and interpretations as I gathered data and wrote up my analysis. As I collected observational notes and reviewed interview transcripts, I made note of events or statements that were unclear to me and later asked participants for clarification. Additionally, as I began to formulate interpretations of the data, I checked with participants to ensure that the interpretations were valid, and then I amended them as necessary. I conducted member checking both in face-to-face interactions with participants as well by submitting passages of some of my findings text to participants for review. For instance, after I wrote a draft description of a team's study, I emailed the text to participants, requesting that they check for accuracy. Participants returned the passage with comments that I then incorporated into revisions.

3.7 LIMITATIONS

Despite careful attention to the design of a study and the interpretation of results, researchers face limitations in any research inquiry. As is common in qualitative studies, the relatively small number of people I could interview and observe limits the generalizability of my findings. I focused my inquiry on the practices and viewpoints of 16 researchers working in three small teams carrying out work at the Station. I selected teams that differed across important

dimensions such as length of the study, funding source, and the career level of PIs in order to look at conceptions of value and their relationship to data practices across a variety of specific contexts at the Station. However, I did not employ this scheme to generate findings that would be applicable to all scientists everywhere or even all ecologists. Rather, I sought to understand the contextual factors at play as scientists considered data's value and as they engaged in data practices.

A second limitation of this research is its reliance on studying scientists during what they called "the field season." Several of the participants characterized the field season as an intense period of data collection, during which they were able to devote relatively little attention to data analysis. With the exception of the NUS Team, whose short study timeline necessitated early and frequent data analysis, I did not observe much data analysis work. I addressed this limitation in a couple of ways. First, I asked participants questions to elicit descriptions of the parts of the research process I was missing. Second, I had several participants walk me through how they planned to analyze data. Still, as my interviews with scientists made clear, important valuation activities occur as scientists analyze their data and even afterward. Scientists' conceptions of their data's value undergo revision and refinement as their studies progress; we can expect that an important point of revision might occur as they analyze data and gain a better understanding of what data show.

3.8 Data Presentation Conventions

To illustrate for readers how scientists understood their activities and conceptualized the value of their data, I quote the participants extensively throughout the dissertation. In doing so, I employ several conventions. I use an ellipsis without brackets to indicate an unfinished statement

or transition in thought by a participant. For example: "So this is like a . . . And we've talked about a controlled experiment for a couple of years." I employ an ellipsis with brackets ([. . .]) to indicate that I have omitted a portion of the speaker's words. I have been careful in these instances to ensure that the omitted text does not change the meaning of the speaker's utterance. Lastly, I also use square brackets to indicate text that I have added to clarify meaning or to hide identifying information such as location names.

I use pseudonyms for all participants in this study. To avoid potential confusion about what team a participant belongs to and/or what position s/he holds on the team, I employ a two-part abbreviation for both facets. For example: "So this is like a . . . And we've talked about a controlled experiment for a couple of years" (Elizabeth, NUS-PI). The first part of the abbreviation stands for the team name, while the second stands for the participant's position within the team. Throughout the dissertation, NUS stands for Nutrient Uptake in Streams, IM for Invasives Management, and ID&M for Invasion Dynamics and Modeling. PI stands for Primary Investigator, PD for post-doctoral researcher, GR for graduate researcher, UR for undergraduate researcher, and RA for research assistant. Three of the participants were Station staff and, hence, had no team affiliation: the station director, the resident biologist, and the information manager. In the dissertation I refer to each of them as such, without abbreviation.

In the next chapter I describe the teams I studied in more detail and depict data valuation activities through three vignettes before moving on to my findings in Chapters 5 and 6.

CHAPTER 4: DETAILED TEAM DESCRIPTIONS AND DATA VALUATION VIGNETTES

Before moving on to my findings, which I present in Chapters 5 and 6, I describe the teams that I studied in detail. In the last chapter, I briefly presented each team's composition and salient characteristics; here I delineate the teams' study designs and the main data they collected. Every team's description is paired with a data valuation vignette. These vignettes highlight the role of data valuation throughout scientists' research processes and demonstrate valuation's influence on how scientists created, managed, and worked with their data. My purpose in presenting the vignettes before discussing the findings is to give more tangible examples of how scientists' interpretations of data's value (including what and whom data are for and how long they might remain valuable) were woven into their work with data as well as demonstrate some of the ways valuations shaped decisions about what to do with data.

4.1 THE INVASIVES MANAGEMENT TEAM

4.1.1 TEAM DESCRIPTION

The IM Team was at the Station to study the effect of different *Typha* removal techniques on wetland restoration and the feasibility of using the removed material as biofuel. At the center of the team's study was an experimental restoration project at three different coastal marshes, each with a different level of anthropogenic disturbance. The team had constructed 24 four-meter by four-meter plots at each of the three sites in 2011 and randomly assigned plots to one of four mechanical harvesting treatments: above ground removal, where researchers removed all *Typha*

biomass and standing litter above the soil surface; below ground removal, where they removed rhizomes and roots and all biomass as well as standing dead litter above the soil surface; mow, where they cut all *Typha* biomass and standing litter above the soil surface and left the material within the plot; and control plots, where they removed nothing.

The team gathered several kinds of data and samples from the plots in order to measure the effects of the treatments and to establish *Typha's* potential as biofuel. During the time of my observations and interviews, the team was collecting biodiversity data and samples that would help them assess the effects of the management treatments. Using four one-meter-square subplot markers (each subplot was also called a "quadrate") within each of the plots, the researchers took several different measurements that would later be used to determine biodiversity (Figure 4.1). They visually estimated the percentages of the subplots that were vegetated, bare ground, detritus, and *Typha*. They measured the heights of each *Typha* stem and the water depth and organic matter depth of each subplot. Additionally, they identified all the plant species in the subplots and estimated a percent cover for each species. The subplot level data would later be averaged together for each plot. In terms of samples, IM Team scientists collected one soil core from each plot, which they would later dry to measure the root biomass. All of these data were to contribute to a time series made up of data collected prior to the treatments in 2011 and data to be collected in years two (2012) and three (2013) of the project.

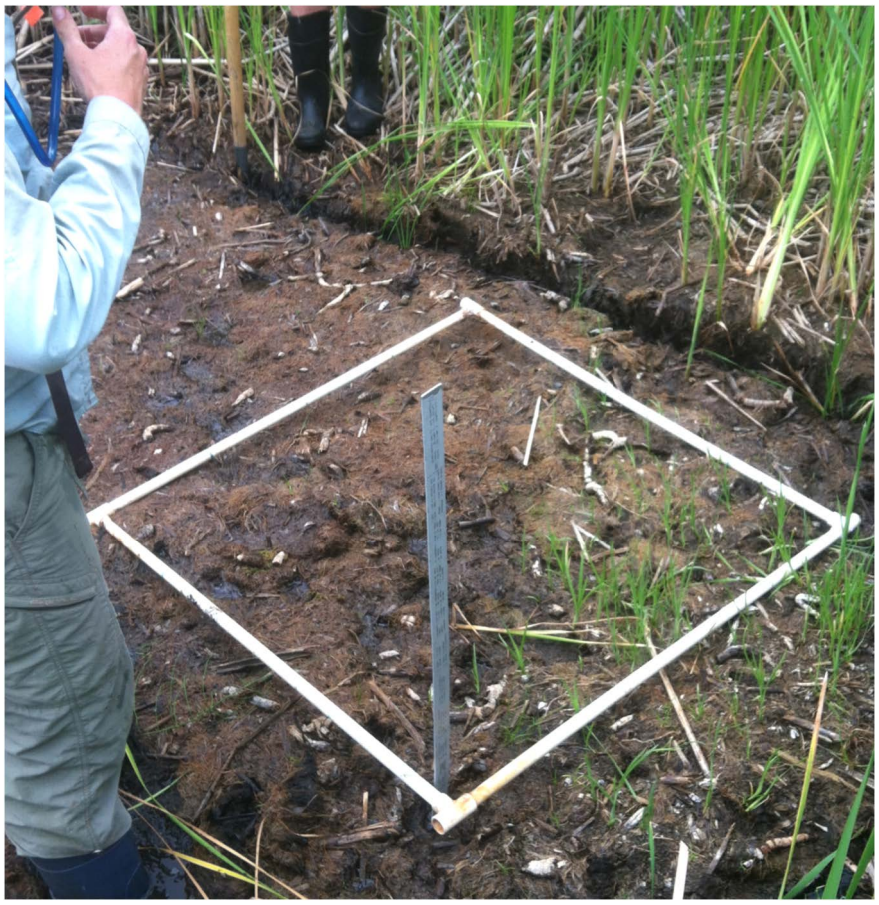


Figure 4.1: IM Team subplot (i.e. quadrate) within a plot. The ruler was used to measure organic matter depth.

While not formally part of grant as written, the IM Team complemented the field study I just described with other, more controlled methods of study. One method consisted of a set of 40 mesocosms that Evelyn (IM-PI) and her team built outside on a hill at the Station in 2002. Ecologists commonly use mesocosms to bring the natural environment into more controlled conditions, helping to counteract the high variability and complexity of factors in the field. In this instance, a mesocosm consisted of a one-meter deep, two-meter long, one-meter wide aluminum tank, filled with soil and populated with plants pulled from a local marsh to mimic the native plant community in that marsh. Additionally, each of the tanks was divided in two sides, with one side receiving a high-water treatment and the other a low-water treatment. After Evelyn

established the native communities in the tanks, she "invaded" half of the tanks (20) with *Typha*. In this study, the team planned to pair some of the data gathered from the experimental field plots with data collected in the mesocosms under the high and low water treatments just mentioned.

The IM Team was also conducting a seed bank study to account for some of the effects that fluctuations in water levels might have on seedling emergence in the field plots. Because the researchers were interested in the restoration potential of wetlands, it was important for them to know what viable seeds were available in the wetlands they were studying. They were collecting data on plant growth at the field sites, but high water conditions could have yielded little plant growth in the years they were studying the sites, hence giving them limited (if any) information about the viability of the existing seed bank at the plots. By removing soil plugs from each of the plots, subjecting them to different water level treatments, and incubating them, the team could assess this restorability potential without the water level being a confounding factor.

Renee (IM-GR), the graduate student working on the team, participated in the plot construction and data collection activities I just described. Additionally, she was collecting data for her master's thesis (with assistance from other members of the team). Her thesis was concerned with the effects of the different *Typha* removal techniques on wetland macroinvertebrate (e.g. leeches and snails) populations. She collected samples of macroinvertebrates from a subsection of each of the field plots so that she could later identify, count, and measure the species to come up with a species-specific biomass number. She also collected information on the vegetation in the plots by identifying plants and either counting stems or estimating percent coverage. In addition, at each subplot she measured the dissolved oxygen content of the water, the temperature of the water, and water depth. Renee conducted

sample and data collection three times over the course of the summer at plots in two of the team's three field sites (once in June, July, and August).

All of the raw data that this team collected at the time of my study was recorded on paper, mostly using paper templates that the team created. All the team members participated in later transferring the data to Excel files to facilitate calculation, graphing, and the creation of statistical measures.

4.1.2 DATA VALUATION VIGNETTE

It is early August, and the IM Team is in its last few weeks of data collection at the Station for the 2012 field season. This is year two of the team's three-year, EPA-funded project to examine the effect of different invasive *Typha* removal techniques on restoring wetland biodiversity. Today, we are at one of the team's three field sites: a marshy wetland on the edge of an inland lake. This site is more than two hours' drive from the Station, so we are camping overnight nearby to maximize the time available for data and sample collection activities. There are six of us on this trip: Matt (IM-PI), Amy (IM-PD), Renee (IM-GR), Dylan (IM-RA), Brooke (IM-UR), and me. Upon arrival at the entry point for the wetland, we load all of the team's sampling and data collection equipment onto "the Argo" (an amphibious, all terrain, off-road vehicle) and climb aboard for a wet and bumpy ride to the first of 24, four-by-four-meter treatment plots the researchers had constructed and treated the previous summer.

In this particular plot, the researchers removed *Typha* by taking away all of the rhizomes and roots of the plant as well as all of its standing litter above the soil surface. As we unload the equipment, the researchers look around at what is coming up in the plot. "This is diverse!" Dylan (IM-RA) remarks as he notices a variety of native plants growing where *Typha* once crowded out

almost everything. Matt (IM-PI) responds, "This is what it's all about. This is money; we can probably use these data in a paper."

Long before they can write a paper, however, they must collect the samples and data. The scientists have a routine that I am by now well acquainted with (and even part of), as this is the third time I have accompanied and assisted the team in the field. Matt (IM-PI) begins to manually "edge" the plot, to maintain its borders and prevent *Typha* from encroaching once again. Amy (IM-PD) and Dylan (IM-RA) pair up to take plant survey data of the plot. To do this, they will sample four, one-meter subplots within the larger plot using a square the researchers have constructed from PVC pipe (Figure 4.1). The team has several such squares, enabling the researchers to break into smaller teams to collect data simultaneously.

Renee (IM-GR) is primarily working on her own sampling and data collection. For her master's thesis, she is examining how wetland macroinvertebrates respond to the *Typha* removal treatments. I partner with Renee and serve as her data recorder, as Amy (IM-PD) and Dylan (IM-RA) begin to call out their sample plot data to Brooke (IM-UR), who has a clipboard with the team's paper datasheets. Renee is studying the same treatments as the rest of the team and even collecting many of the same kinds of data, but she conducts her data collection and sampling activities largely separate from the efforts of the rest of the team. Not only do graduate research expectations demand that she collect her own data for her thesis, but the scientists' assumptions about *what data need to be good for* also differ in important ways between the two projects. As we are about to see, this difference in assumptions affects the data that the researchers collect and how they collect them.

An important component of both the IM Team's main project and Renee's master's project consists of identifying plants and quantifying their amount within a sample area of a plot.

For the main study, which focuses specifically on native vegetation's response to *Typha* removal, the researchers must properly identify the genus and species of all the plants they see in the subplots and estimate what percentage of the subplot each species covers (they call this variable "percent cover"). They plan to use these data to calculate measures of biodiversity, such as a diversity index and floristic quality.

Renee (IM-GR) also has to identify plants and quantify their amount, but she is doing so to calculate a diversity index and to characterize the context of her study. She hopes to determine whether or not plant diversity correlates with macroinvertebrate diversity. These different purposes mean that Renee's data are not interchangeable with that of the rest of the team's, a fact demonstrated when Amy (IM-PD) has trouble identifying a plant. "What did you call this?" she asks Renee, who has just gathered her own data from the subplot Amy and Dylan (IM-RA) now work on. "I called it '*Cicuda*,' but it doesn't really matter for me." While important, the species names of the plants are not crucial to the questions Renee is trying to answer in her thesis, which revolve around macroinvertebrate response to invasive *Typha* removal. Technically, she explains to me later, she could just call each species "1," "2," "3," etc., "but that wouldn't be good science," nor an appropriate way of characterizing her plots when she describes where she conducted her study. As a result, she names them, but does not expend a lot of effort on ensuring that she has correctly identified every plant because she can afford to be less accurate.

For the larger project, however, it is crucial that the researchers record the correct plant species name, because calculations like floristic quality rely on species-specific measures. To ensure that they capture that information accurately, the researchers carry a plant identification book in the field. If they are unable to identify a plant in the field, they take a small sample with them so that they can "key it out" later at the lab with a microscope.

When Renee (IM-GR) qualifies her identification of a plant with "but it doesn't really matter for me," she acknowledges that what are good enough data for her are likely not good enough—at least in this case—for the larger project. Still unable to confidently identify the plant in question, Amy (IM-PD) clips a small sample, places it in a plastic bag, and packs it away for later identification.

4.2 The Nutrient Uptake in Streams Team

4.2.1 Team Description

The NUS Team was at the Station to study the effect of leaf litter on nutrient uptake in streams. The PIs designed an eight-week-long experiment that would utilize the Station's artificial stream lab to test the nutrient uptake mechanism. The team's overarching hypothesis, formulated from observations the PIs made together in a field study, was that uptake—in this case, the amount of nutrients consumed from the water column by microbes—would depend on the nitrogen to phosphorus ratio of the leaves that the microbes live on.

The Station's stream lab (an outdoor facility) took in water from a nearby stream and pumped it out into several areas of the lab, where researchers could then use the water to conduct various studies. The NUS Team made use of a large concrete pad and four tanks that were fed with stream water. Each of the tanks had eight taps that could be turned on or off. Using plastic roof gutters, the team built and attached 20-meter long artificial stream channels underneath each of the eight taps on the four tanks (Figure 4.2).

Figure 4.2: Header tank, taps, and two channels for the NUS Team's artificial stream setup.

The researchers randomly assigned each of the channels belonging to a header tank to a treatment of maple leaves, cottonwood leaves, a mix of maple and cottonwood leaves, or a control that had no leaves at all. If, for example, a channel was assigned a maple treatment, that channel was filled with 300 grams of maple leaves that the researchers gathered and brought with them to the Station. They planned to collect water samples and analyze them from each replicate periodically throughout the summer, because they hypothesized that leaf decomposition would affect the relative nutrient uptake from the channels.

NUS Team researchers gathered most of the data for this project from samples of water they collected from the artificial stream channels. From the samples, the scientists measured concentrations of ammonium and phosphate using equipment they brought with them to the Station. They also needed to measure the concentration of nitrate in samples of water, but

because that equipment was not available for them to use at the Station, they would wait until they returned to their home institutions to analyze most of the nitrate samples (they paid to have the Station's analytical lab analyze a small portion of their nitrate samples so that they could check that their setup was appropriate). The data on the nutrient concentrations of water samples would later be used, along with other important data (e.g. the length of the channels), to calculate a "nutrient uptake rate," or the rate at which the microbes on the leaves consumed nutrients from the stream.

Every day, the team spent the morning collecting water samples from the eight channels beneath one of the tanks. Later in the day, once all the samples were collected, the undergraduate researchers ran nutrient analyses on the samples. The researchers recorded all of the raw data by hand onto paper templates. They transferred most of these data shortly thereafter (within a few days) to Excel spreadsheets, designed by the PIs and formatted with formulas for calculating other measures.

4.2.2 DATA VALUATION VIGNETTE

It is late July, the fifth week of the NUS Team's planned eight-week project at the Station to study the effect of leaf litter on nutrient uptake in streams. The researchers are gathered in a room they have occupied this summer in the main laboratory. Carolyn (NUS-UR) and Janet (NUS-UR) clean equipment and prepare water samples they gathered earlier in the day for analysis; Tina (NUS-GR) takes notes on a separate, but related, experiment she has been responsible for running; and Elizabeth (NUS-PI) and Jessica (NUS-PI) look at some of the data that the undergraduates have entered into preformatted Excel spreadsheets. It is a fairly normal afternoon for these researchers save for one thing: as they carry out these activities they are also

discussing whether they should continue with the project or pack up early and leave the Station to return home.

This team's project has been beset by difficulties from the beginning. The setup of the experiment, which relies on 32 artificial streams the researchers had to construct from vinyl gutters and fill with leaves they shipped to the Station, was more challenging and labor intensive than they anticipated. The time the researchers spent constructing and adjusting the setup ultimately cut into time that they expected to be producing data. Thankfully, they have been analyzing samples and looking at their data on a daily basis to, as Elizabeth (NUS-PI) explained, "inform what we do [. . .] as we go through the process." However, each day spent adjusting their experiment instead of collecting "real data" meant fewer data points on the graph they wanted to generate. At some point, this loss of data points would threaten to diminish the potential impact of what the researchers wanted to show.

The cost of conducting the study has been relatively high for the researchers. In stark contrast to the other two teams I am studying, these researchers work on their project every day of the week (including Saturdays and Sundays). Their days start at 9am and, for the PIs, frequently end at 11pm, with much of that time spent exposed in the sun on hot and humid summer days. Additionally, their work has had the usual inconveniences of working in a place that is unfamiliar (NUS scientists, unlike most at the Station, had never conducted research at this Station) and remote: Internet service is spotty, the researchers have to go to another floor of the lab to get the deionized water they need to rinse sample bottles, and some of their sample analyses will have to wait until they return to their home institutions because the equipment is not available to them at the Station.

If they can get a good, publishable set of data (especially one compelling enough to land an article in *Ecology*, a high-impact journal) it will all be worth it. The problem—and the reason for the current meeting—is that, despite numerous adjustments to the experimental setup, the researchers are not getting the data that they need to show that the uptake of nitrogen and phosphorus is affected by leaf litter. Elizabeth (NUS-PI) and Jessica (NUS-PI) are struggling to figure out why these data do not support what they observed in the real streams they studied several summers ago. Is the setup too artificial? Is the vinyl gutter material influencing nutrient uptake? Did the researchers kill the microbes on the leaves when they switched from stream water to colder groundwater (an effort to reduce algae and other fine particulate matter that they suspected was influencing nutrient uptake)? Regardless, these are not the data that the team intended to capture. Now, the most salient question for the team is, "Should we continue?" and to answer that question the researchers generate a list of "costs" and "benefits" of continuing the project or stopping it (Figures 4.3 and 4.4).

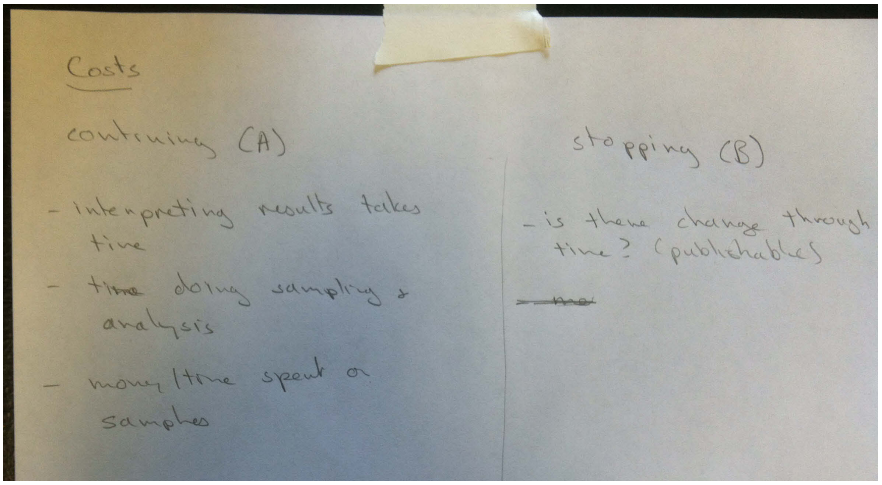


Figure 4.3: NUS Team list of anticipated costs of continuing or stopping the project.

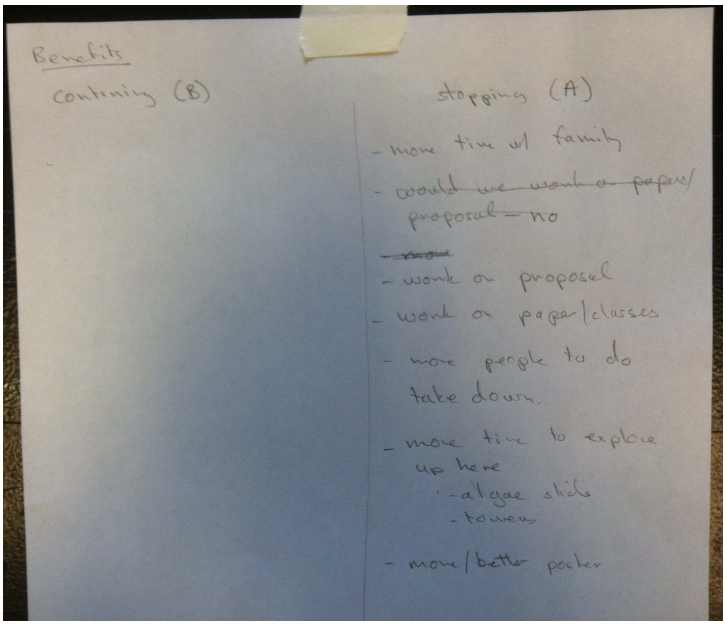


Figure 4.4: NUS Team list of anticipated benefits of continuing or stopping the project.

The only benefit the researchers can conceive of in staying is that they *might* generate data that are publishable, but that seems unlikely to the PIs now. If they stop the project, the researchers can spend time with their families, engage in leisure activities around the Station, and/or work on other things (like grant proposals, classes, student posters, and writing up previous research). After looking at the two lists Elizabeth (NUS-PI) remarks,

> It seems pretty clear . . . I think our energies might be better used elsewhere. It's pretty obvious this data isn't measuring up in terms of value to other things.

The other researchers agree, and they begin to disassemble their stream channels the next day. All of the NUS Team researchers will have left the Station by the following week, with what they consider to be, at most, preliminary data for a future grant proposal.

4.3 The Invasion Dynamics and Modeling Team

4.3.1 Detailed Team Description

The ID&M Team was studying how nutrient dynamics affect the ability of invasives to succeed in wetlands and was also developing and testing a computational model for ecosystem modeling. A major component of the team's study was a mesocosm experiment that the researchers were carrying out at two sites: one at the Station and another approximately 250 miles south of the Station. In 2011, the ID&M Team scientists installed 50, six-foot-diameter, steel tanks into the ground at each of these sites. They filled the tanks with gravel and soil, installed water input and output systems into each of them, and planted half with native wetland vegetation (Figure 4.5). They left almost half (24) of the tanks unvegetated (not planted with native plants) at the time of my study; the team planned to invade 48 of the tanks (empty and those populated with natives) with *Typha* and *Phragmites* (also known as "common reed") in 2013.

Figure 4.5: ID&M Team's mesocosm tanks at the Station before researchers populated them with plants (Courtesy of the ID&M Team).

Two tanks at each site would serve as controls, with no vegetation. Additionally, the team assigned different treatments of fertilizer to the tanks.

Using the mesocosms, the researchers intended to study how nutrients affect the probability and severity of invasion and how climate influences that relationship. Because none of the tanks had yet been invaded at the time of my study, the scientists were collecting what they called "baseline data" from the mesocosms. These data consisted of measurements taken by one of the postdocs and the research assistant on plant community and soil properties from each of the planted tanks. They counted and measured stems of each of the four native species they planted in the tanks. They planned to later convert these data into a biomass measurement, which is a measure of plant productivity. They also collected soil samples periodically over the field season from the tanks and processed them for analysis by the Station's analytic lab. Using the data they got back from the lab, the researchers would calculate a nitrogen mineralization rate (the rate at which nitrogen is converted to a form that plants can use) for the soil. Lastly, the team was collecting data on the chemistry of the water going into and coming out of the mesocosms.

In conjunction with the mesocosm study, the ID&M Team was also developing and testing a computational ecosystem model. Mark (ID&M-PI), a tenured professor, led this effort with assistance from Gabe (ID&M-PD), a post-doctoral researcher. Mark had already developed a version of the model when the team applied for NASA-funding. During my study, Mark and Gabe were running beta tests of the model to uncover bugs in the model and were also preliminarily assessing how well the model's predictions matched early observations in the mesocosms. Most of the data generated from the model were the result of exercises that Mark and Gabe ran using different sets of input parameters.

As with the other two teams, the majority of ID&M's raw mesocosm data were recorded using pencil and preformatted templates. The research assistant transferred these data to Excel spreadsheets where the post-docs prepared the data for analysis by running various transformations and calculations. The modeling data were much more voluminous (millions of cells in a spreadsheet) than the data generated from the mesocosms. Each time Mark (ID&M-PI) or Gabe (ID&M-PD) ran the model, it output a text file containing a set of numeric data characterizing the ecosystem and its changes over time within the exercise. They would then import these data into Excel and Stata, a statistical software package, in order to organize the data, check for errors, create graphs, and look for patterns.

4.3.2 DATA VALUATION VIGNETTE

It is early evening, after dinner, and I have been invited to observe a meeting between the Station's information manager and Ethan (ID&M-PD), a post-doctoral researcher for the ID&M Team. The information manager has been scheduling meetings with researchers at the Station this summer to begin to try to identify the datasets that researchers are developing and to find out which datasets they would be comfortable archiving at the Station. This is the second year of the ID&M Team's project, and the researchers are just now beginning to collect some data from the mesocosms they installed last summer at the Station and another site. Ethan anticipates that he will be the team member in the best position to provide data documentation, but also knows he will likely have departed the project when the data are ready to be archived. He has come to the meeting to find out what things the Station might want archived and what is required from the team to facilitate archiving the data. He has brought with him a list of datasets that the team has

begun to assemble so far. We sit at a picnic table outside, along the lake, and I listen as Ethan and the information manager talk.

The information manager begins the meeting by checking to make sure that Ethan (ID&M-PD) understands the Station's data policies. While Ethan recalls reading the policy online, it was long enough ago that he does not remember the details. The information manager explains that researchers using Station resources are required to submit their data to the Station within a year after ending their projects. He emphasizes that this includes data themselves and the documentation that makes data understandable to other scientists.

Ethan (ID&M-PD) begins to recite the list of datasets that his team is collecting to find out which of them the Station would be interested in archiving and what, more specifically, would be needed to archive them. In discussing what data the team might archive at the conclusion of the project, the information manager and Ethan talk about what data would be "useful" or "important" for other scientists to have. The information manager tells Ethan that he thinks some of the water data that ID&M researchers are collecting might be particularly useful to other scientists who conduct research at the Station. Specifically, he thinks that data on the chemistry of the water flowing into the mesocosm tanks have the potential to be useful to others because they represent something "more global." Many other Station researchers use the same water in their own studies, so those data have possible broad applicability. Ethan remarks in surprise, "That's true. I hadn't thought of it that way." In fact, Ethan came to the meeting thinking of those data as having a rather short-term and limited value compared to some of the other data his team was collecting.

As I would later learn, the water chemistry data were just one example of several datasets ID&M Team researchers were compiling that they considered to be "small, ancillary datasets"

they would use to interpret other data. The researchers were collecting water chemistry data because they anticipated a difference between the water at the Station and the water at their other site. They did not want this difference to "mess up" their data, so they accounted for the possible difference by measuring it. Differences in the water chemistry, in other words, might be a way to explain differences that the researchers could see in their data between the two sites.

ID&M Team researchers considered the water chemistry data, along with several other peripheral datasets, to have only short-term, within-team value. In the following interview excerpt, Kate (ID&M-PI) describes the assumptions that undergirded her ideas about what makes data worth archiving or sharing.

> Interviewer: Are there any data that you *don't* think are really worth . . . that you're planning to collect that you *don't* really think would be worth archiving or making available?
>
> Kate: There is some sort of what we would call "small, ancillary datasets"; just things that we figure are going to help us interpret, but they're pretty small. Like tomorrow and Friday, Ethan and Chad are going to go out and get a whole lot of temperature data. That's not something we're going to archive. That's something—
>
> Interviewer: Because that's to measure the water flow through the tank?
>
> Kate: Yeah, that's for us to try to figure out what is the . . . That's something that we might use to illustrate or to demonstrate that our treatments work in a way we'd like them to work, but they're not data that anybody else could ever use. The plant data is definitely . . . The long-term vegetation data, that's something that we will use and will add to every year. The changes in the soil nutrients, we would archive. The temperature data just sticks out to me. I mean, another little piece . . . We have some baseline data on the water input comparing what happens at . . . the water quality going into the mesocosms at the Station versus the [other site]. They're actually different. We're trying not to think about it, because there's not much we could do about it. It just will go into our interpretation. Will that be archived? I don't know.

Without the information manager's suggestion that the water chemistry data might be more "broadly useful," these researchers assume almost the exact opposite; that data they are just using to ensure their experiments are set up right and make sure they can account for anomalies

would have no value past the team's own very narrow use. It is still early enough in the ID&M Team's data collection that the information manager suggests that he and Ethan (ID&M-PD) meet again in the winter, when the researchers have more data. The two part, still undecided about what data will be archived, but Ethan leaves this meeting with a different perspective on what data might be worth archiving at the Station.

4.4 SUMMARY

The three vignettes I presented capture points in the research process where scientists' conceptions of value played an important role in determining whether and how data were captured, as well as how they were managed for future use. Assumptions about what data were for—their purposes—undergirded scientists' valuation of data as well as the creation of valuable data. In the IM Team, researchers ostensibly looking at some of the same variables created data that differed in their level of accuracy. Renee's (IM-GR) planned use for plant species identification demanded somewhat less accuracy than the rest of the team's use for the same data. It was more important that Renee capture the number of different species than to get those exact species right. The other researchers, however, planned to calculate a measure that required species-specific information. As a result, they took extra steps—consulting a plant identification book and taking samples back to the lab when necessary—to ensure the accuracy of these data.

In the second vignette, which recounted an important meeting for NUS Team members, the researchers decided to end their project because the data they were generating were not effectively capturing the mechanism they intended to study, despite repeated attempts to alter the experimental setup. The researchers embarked on the study with the expectation that—as long as everything went according to plan—they would have data worthy of publication in a top-tier

ecological science journal. For the two PIs, who were both junior faculty members, getting data that could result in such a publication was especially important for bolstering their cases for tenure.

The stakes for the NUS Team were high and their time to collect good data relatively short, both of which were reflected in the PIs regular and frequent assessment of the quality of their data. As they conducted the study and reviewed their data, the PIs contemplated the worth of the data against the costs of being at the Station, including the opportunity costs associated with working on this project instead of other things (grant proposals, writing up data from previous studies) that may have been more worthwhile. With three weeks left in their study, the researchers thought the most they would be able to use the data for was as preliminary data in a grant proposal and in the posters the undergraduates were required to present by their university as part of their summer research experience. Those were not end purposes that were compelling enough to justify continuing the project at the Station.

In the last vignette, I described assumptions that ID&M Team researchers had about the limited value of "peripheral" datasets. The scientists were not necessarily interested in these data for analyses that would directly answer their research questions; rather, the researchers were gathering such data to set up their mesocosm experiment and account for possible anomalies in their findings. The scientists' assumption that those kinds of data had little, if any, longer-term value or value to researchers outside of the ID&M Team, was challenged by the information manager when Ethan (ID&M-PD) met with him. In discussing what ID&M data might be worth archiving, the information manager thought that water chemistry data would be particularly useful to other scientists that also relied on the same water in their own studies. Such a difference in conceptions about who and what and for how long data might be valuable for highlights how

scientists' relatively narrow focus can influence their conceptions of value as well as what they plan to do with data. Furthermore, it demonstrates how someone with a more global perspective, such as the Station's information manager, might be able to influence scientists' conceptions of their data's secondary or longer-term value.

These themes will reemerge in the next two chapters, which together describe my study's findings. In Chapter 5, I focus on scientists' conceptions of data's value, while in Chapter 6 I delineate how conceptions of value were enacted in scientists' data practices. I conclude the dissertation in Chapter 7 with a discussion of the implications of my findings and suggestions for future research.

CHAPTER 5: SCIENTISTS' CONCEPTIONS OF DATA'S VALUE

5.1 OVERVIEW

Scientists' conceptions of data's value were woven into virtually every aspect of their work, from project planning and data collection to documentation, analysis, and sharing and archiving plans and activities. Their conceptions of value encompassed not only assessments about whether data were good or bad but also assumptions about how long data would be of value, what purposes and benefits data could serve, and who could or should reap the benefits of data's value. As suggested in previous research on the roles of data in science (Birnholtz & Bietz, 2003; Latour, 1987; Latour & Woolgar, 1986), I found that the instrumental value of data was particularly salient for scientists: that whether scientists considered data valuable and to what degree depended largely on how well they thought data would serve some end that they considered worthy. Furthermore, my study confirms other researchers' finding that scientists are concerned with reward-based ends, such as peer-reviewed publications (Borgman, Wallis, & Enyedy, 2006), increased credibility (Latour & Woolgar, 1986), and tenure (Tucker, 2009). However, these ends do not fully account for scientists' conceptions of value or their data practices, especially given that scientists' notions of data's value were based in large part on the type of data being considered.

Scientists made it clear that, first and foremost, they were carrying out their studies to produce data for addressing the specific gap in knowledge they had defined. Good, valuable, or useful data (scientists used these terms interchangeably) were data that helped the scientists

answer their research questions and/or test their hypotheses, and, therefore, fit-to-study was an important basis for assessments of data's utility or value across the data stream. However, scientists also made assumptions about the time span of their data's value and who or what purposes data would be valuable *for* on the basis of data type, as conceived by the scientists working with the data. Type designations, such as "field data," "experimental data," "unpublished data," "publishable data," "raw data," and "derived data" revealed the numerous meanings data had for scientists—both in terms of what the data represented and the purposes to which they could be put—that shaped which particular data scientists thought should be shared and/or archived. Scientists' focus on data's publication status and potential and data's processing state affirms the data stream model's suggestion that how scientists view the value of their data—and hence suitability for sharing—varies according to data's level of refinement and where scientists are in their research process. But my findings also extend the model by uncovering an additional property—the potential of data to be supplemented, extended, or combined with other data —that is important to scientists as they consider the value of their data to others.

I present the results of my study in two chapters. This chapter focuses on elucidating scientists' conceptions of data's value. It begins with a description of what scientists emphasized was the key purpose of their data: filling a specific gap in knowledge. Then, I describe the prerequisites data needed to fulfill in order for scientists to consider them of value for this primary purpose or to be what they considered "good data." In the second half of the chapter, I delineate the three main ways in which scientists characterized or classified their data when considering or talking about data's value. This typology reveals the bases that are important to scientists as they assess the value of data in the data stream.

5.2 What Data are For: Addressing a Gap in Knowledge

The scientists I studied measured the nutrient content of water and soil samples, the lengths of various macroinvertebrates, and the amount of dissolved oxygen in wetland sample plots. They counted plant stems and recorded their heights, took soil temperature readings, and weighed soil samples. They calculated biomass, nutrient uptake length, and plant diversity indices. However, these activities were motivated by purposes that extended well beyond collecting or calculating any of the individual measures themselves, a point highlighted by a researcher from the IM Team.

> [. . .] if you think about it short-term it almost kind of seems meaningless. Like sometimes I actually find myself getting caught up in that. I'm like, "Does it really matter what this exact sedge is?" Like if it's *Juncus balticus* or *Juncus nodosus*, does it matter? But if you think about it long-term, it's not just about that. [. . .] It's not about the little identifying plants. *(Brooke, IM-UR)*

At the most basic level, the teams were at the Station to examine a specifically defined gap in knowledge. What that particular gap was and what the teams hoped addressing it would accomplish informed scientists' overarching expectations for the outcomes of their work at the Station and their understanding of the kinds of data they needed to fulfill those expectations. Each of the teams came to the Station with clearly delineated research questions and/or hypotheses that framed virtually all of their work activities. Successfully answering their own research questions was a key criterion against which scientists judged the success of their work and, hence, the quality and value of their data. Next, I describe the knowledge gap each team was at the Station to explore, highlighting how it shaped researchers' understanding of what data they needed.

5.2.1 The IM Team: Plant Community Response to Wetland Restoration

The IM Team was working to find a successful and economically sustainable method of removing invasive *Typha* (cattail) from wetlands. IM Team scientists frequently emphasized the applied, rather than basic, nature of their research at the Station:

> [. . .] it's not basic. It's applied. It's applied to try to solve a problem. But there is a lot of basic in there, because we don't know anything about this kind of stuff. We don't know anything about harvesting and how much it increases biodiversity. There's definitely basic stuff that you can present and publish. It's just that, for me, it's sort of driven by the bigger question of solving a problem. *(Evelyn, IM-PI)*

The problem of figuring out how to sustainably and affordably remove this invasive species to support greater biodiversity is an important one, especially for land managers who are responsible for mitigating plant and animal invasion. *Typha*, which grows prolifically via underground rhizomes, can quickly form a dense layer of litter that dramatically affects both plant and animal diversity, reducing a wetland to essentially a monoculture. Land managers do not currently have effective, economically sustainable, and non-harmful methods of removing this invasive plant. Herbicides release toxic chemicals, and burning the *Typha* releases greenhouse gases. Mechanical removal methods, such as mowing and digging below ground, are much less harmful to wetlands, but they are labor-intensive and, hence, expensive. IM Team researchers aimed to learn how different *Typha* removal techniques affected wetland biodiversity and whether removal could be made more economically viable by converting the *Typha* into biofuel that land managers could sell to recoup some of the removal costs.

When I asked IM Team scientists to describe the goals of the project, they frequently emphasized the applied ends. The researchers said they wanted to "transform the way people manage their wetlands" (Amy, IM-PD) and "make management more sustainable" (Dylan, IM-

RA). As a result, IM Team researchers thought it critical that their work yielded results that land managers could use and understand.

> We want this data to be used by not only invasive plant researchers, but [also by] land managers; the people who deal with *Typha* on a daily basis, people who have *Typha* control or people that are interested in controlling cattails where they are. They should be able to read these papers and be like, "This mowing technique at this time of year seems to be effective in eliminating cattails." *(Dylan, IM-RA)*

> We're trying to make it a little bit more applied, so that it can help affect change and bring solutions rather than just being really eccentric and kind of tedious little work that you publish and ten people read your papers. *(Evelyn, IM-PI)*

The team's focus on what the scientists characterized as an applied problem and outcomes that could be implemented in real-world land management settings shaped their notions about what their data needed to be good for. Most importantly, the IM Team's data needed to demonstrate the biodiversity effects of different *Typha* removal methods and lead to results that could be applied by managers. This meant the data had to span enough time to convincingly show trends, represent phenomena in a natural setting, and be built on measurements that land managers could and would use. I describe each in turn.

The changes that IM Team researchers were studying occurred over a span of several years. Phenomena such as native plant resurgence (including the species that emerged, their diversity, and how much of the plots those species covered) and *Typha* resurgence (including how quickly it came back and what portion of the plots it covered) take years to show anything meaningful for scientists and land managers. Evelyn (IM-PI) explained that the longer-term the dataset, the better.

> [. . .] because if you have long-term datasets, you can really . . . When your questions are like mine, which is, "What is the human impact on these natural ecosystems?" the longer-term your database is, the more powerful your description of human impact will be. If you just show one-year impact it's kind of like, "Well, that's not very helpful." *(Evelyn, IM-PI)*

The particular project the team was working on as I conducted my study was funded for three years; while the researchers described this as not ideal, it was a sufficient timeframe to study the response to *Typha* removal.

> Dylan (IM-RA): [. . .] we have a management technique, and we want to find out if our management's working. So we need to go out and collect data over a course of a couple of years. This is a three-year project. This is year two, and then we're going to be collecting the same data year three. In those three years, we can compare changes in species composition, *Typha* density, soil, [and] organic matter. Ideally—a project like this—we would have decades of data, and we just—
>
> Interviewer: Because time is such an important factor?
>
> Dylan: Right. Time is really important, but we don't really have that luxury.

Not only was it important that data spanned several years, but because IM Team scientists were focused on studying *Typha* management techniques, the data needed to be drawn from a minimally controlled setting. The main portion of the team's study focused on a field manipulation at three different wetland sites. In year one, the researchers cordoned off 24 plots at each of three wetlands (for a total of 72 plots) and assigned each of the plots to one of four treatments: above-ground removal, below-ground removal, mow, or control. The researchers referred to their study design as a "field experiment." Unlike other kinds of experimental designs (e.g. the mesocosm experiments conducted by the IM Team and the ID&M Team and the artificial stream experiment carried out by the NUS Team), in field experiments scientists control relatively few variables. In this case, the only manipulation the scientists performed was *Typha* removal. To study the influence of two other variables the team posited were important in determining biodiversity response (wetness and *Typha* age), the scientists also assigned plots according to the age of the *Typha* stand and relative wetness of the wetland.

Scientists use field experiments to study phenomena in a natural environment, with the variabilities of the natural world kept relatively intact. This lends results a high degree of validity for application to the real world. At the same time, the variability that is an inherent trait of ecosystems (spatial, temporal, etc.) threatens to complicate findings by obscuring phenomena of interest. For example, the IM Team wanted to evaluate the wetlands' restoration potential, which would be indicated by how many species grew after *Typha* removal. However, because seedling resurgence depended not only on the seeds in the soil, but also on water level, an exceptionally wet year could have resulted in data that did not accurately capture wetland restorability. In Matt's (IM-PI) words,

> In the field there's this—particularly, the water level variability . . . it's out of our control. Many wetland plants species will only germinate [. . .] when the water levels are at the right level. So we weren't sure . . . I mean, had this been a really high-water a year, we might have gotten no germination in any of our treatments, which would have—if that were the only data we were collecting—would have not have represented the actual potential for restorability. And so what we did is we collected soil plugs from each of the plots, and then we have them in environmental chambers. We're looking at what plant species grow out of the sediments, which is sort of a proxy for restorability.

Because of the importance of water level on seed germination, the researchers carried out a small-scale experiment whereby they could control the water level. Doing so would allow them to reliably assess the quality of the existing seed bank in the wetlands they were studying, regardless of the water variability in the field.

Lastly, the scientists' interest in the results of their work having real-world application meant that their data had to be the result of "quick" and "simple" methods of measuring key variables that would be more useable to managers who, as Matt (IM-PI) explained, would not take a "painstaking amount of time" to carry out in their work. For example, numbers that were the result of the time-consuming process of counting plant stems to arrive at measures of

abundance were described to me as "less transferable" to the applied science community than percent cover measurements.

> A lot of applied scientists and managers understand percent cover, so it is something that is pretty transferable outside of [basic] research. *(Matt, IM-PI)*

The IM Team's questions depended on understanding phenomena that occur over years, and the team sought findings that could be applied to real-world management situations. As a result, valuable data to the team were data collected over several years, drawn mostly from a relatively uncontrolled system, and based on simple and easy-to-attain measures.

5.2.2 The NUS Team: Leaf Litter's Effect on Nutrient Uptake in Streams

In stark contrast to the IM Team, the NUS Team aimed to reveal a narrow mechanism (the reliance of nutrient uptake in streams on the chemical composition of leaf litter), and this undergirded an entirely different conception of valuable data. The NUS Team came to the Station to study how leaf litter affects nutrient uptake in streams. As human-driven nutrient runoff continues to plague many streams (contributing to algal growth and decreased water quality), results from this work could potentially suggest methods of mitigating the problem, such as planting specific kinds of trees along streams affected by nutrient disturbances. The application of the research, however, was less the focus of NUS Team's current project than was understanding what the PIs hypothesized was an important factor in nutrient uptake in streams. Based on fieldwork that the PIs previously conducted on real streams, the team hypothesized that: 1.) Nutrient uptake in a stream would depend on the ratio of nitrogen to phosphorus on the leaves in the stream; and 2.) As the leaves decomposed, the carbon to nitrogen and carbon to phosphorus ratios of the leaves would increase and nutrient uptake in the different leaf treatments

would become more similar to one another. The PIs characterized their study as designed specifically to address *only* these hypotheses and the questions related to them.

> [. . .] the *whole* experiment was designed around two questions, and you don't have other sorts of variabilities. You don't have differences in ambient concentrations or differences in site or differences in channel dynamics that . . . You know, they're all . . . It's all for the exact same thing . . . all designed just to answer two questions. *(Jessica, NUS-PI)*

The primary goal (supporting hypotheses that the scientists developed on the basis of their own fieldwork) of the NUS Team was, in other words, purposefully narrow. While data from an earlier field study indicated that leaf litter was a significant driver of nutrient uptake in streams, the scientists thought a controlled experiment would better allow the team to isolate and test the particular mechanism and test their hypotheses.

> [. . .] we have some mathematical equations that describe the patterns that we think are going on. I think our steps right now are really just sort of . . . you can think of it as parameterizing those equations, like testing. If you actually you do have this, does it actually work out the way the math says it should work out? *(Elizabeth, NUS-PI)*

In order to demonstrate the relationship between leaf litter and nutrient uptake, the team needed to remove other potential confounding variables from their study system. For example, the researchers controlled the quality of the leaf litter in their channels; the only leaves in the channels were the leaves they placed there themselves. Most importantly, NUS Team scientists worked to keep out of their system any other factors that could potentially influence nutrient uptake. They took several steps to accomplish this level of control. For instance, they covered their channels with a blue tarp to inhibit potential algae growth. The researchers attached nylon mesh to the output valves of the water tanks to filter out fine particulate matter (like bits of leaves of unknown origin and other organic material) that might be brought in from the stream. Lastly, every night they used paintbrushes to manually remove algae from the channels.

If everything went according to plan, the narrowness of this team's questions and goals (to show a specific, isolated mechanism) and the relatively artificial setup the researchers constructed would yield a set of data that showed just one thing: nutrient uptake in streams as a function of nutrient balance of leaf litter. Data's value to these scientists, in other words, was contingent upon the data showing a specific relationship.

5.2.3 The ID&M Team: Mechanisms of Wetland Plant Invasion

The ID&M Team's work was part of a larger, NASA-funded, multi-disciplinary and multi-institutional project that aimed to understand the mechanisms of wetland ecosystem functioning and to quantify the potential impacts of anthropogenic changes (such as those from agriculture and land development) on wetlands. At the time of my study, there was minimal collaboration across partner sites, and work proceeded relatively independently at each institution. The overarching goals of the ID&M Team, then, revolved around these researchers' own portion of the larger plan: uncovering the mechanisms that control plant invasion into wetlands and refining and testing a computational model. Specifically, ID&M Team researchers aimed to identify the traits of invasive plants that were most important in those plants' ability to invade a wetland and to better understand the effect of different levels of nutrients, climate, and native vegetation on invasion success. Additionally, the researchers wanted to understand the consequences of invasion, particularly how invasion changes nutrient cycling.

Hypotheses about these dynamics or mechanisms came from field studies that Kate (ID&M-PI) previously conducted and from the Wetland Ecosystem Model:[9] a computational model developed by Mark (ID&M-PI)—an ecosystem modeler—in conjunction with Kate. This

[9] Pseudonym used.

model simulated plant community dynamics and the biogeochemistry of nitrogen and carbon-nitrogen interactions in wetlands. ID&M Team researchers planned to test quantitative relationships between invader size, nutrient availability, and competitiveness by experimentally manipulating the soil fertility of the wetland mesocosms they constructed at two different sites. In addition to using the results they gathered from the mesocosm experiment to test their hypotheses, the researchers planned to use the data to further parameterize the model to enhance its power as a hypothesis generator. To accomplish their goals, the researchers needed data that were drawn from a relatively controlled system, that could be used to represent change over time, and that were comparable to data used and generated by the Wetland Ecosystem Model.

This team's study centered on a mesocosm experiment that the researchers set up in the summer of 2011. Mesocosms are constructed enclosures that bring the natural environment into semi-controlled conditions that scientists can then manipulate for study. This experimental approach (which IM Team researchers also used as a supplement to their field study) allows scientists a level of control that is not possible in a typical field study and at the same time enables scientists to maintain a degree of realism that is not normally seen in a more tightly controlled laboratory experiment. As Kate (ID&M-PI) explained,

> You're doing [mesocosm experiments] because you're trying to make it somewhat realistic, but maintain much more control than you can do in the field. [. . .] For me the important control is usually what the community is going to be. I can start the community the way I want it and then look at particular components that are going to be varied. Whereas if you go out in the field you've got the community and the environment co-varying, and you can't pull out one or the other except by really huge sample sizes. You can isolate, you can impose the same experimental treatment, but you've just got a huge variation in the background [. . .]

ID&M Team scientists built each mesocosm from a 6-foot diameter, metal tank. They placed the tanks (50 in total) in the ground at two study sites, filled them with soil, and installed water input and output systems in each of them. The end result was that each tank became what

Ethan (ID&M-PD) called "[its] own six-foot diameter wetland." In 2012, the researchers planted half of the mesocosms with four different species of native vegetation (they left half empty of vegetation) and treated the mesocosms with varying levels of fertilizer. In 2013, once the native plants were well established, the researchers planned to add invasive plants to all of the tanks at each site.

In order to study the mechanisms that were ID&M Team's focus, the scientists deliberately constructed the mesocosms to isolate the variables of interest and remove other factors that might increase the variability of the team's data. For example, the soil (including its starting nutrient level) was exactly the same in every tank, because the only soil in them was what the researchers added. The researchers also placed each of the two sets of mesocosms in an exposed location without the threat of shade that might cause variability in light exposure. In this way, the researchers could be assured that all of the tanks would get the same amount of sunlight. Additionally, the team populated half of the tanks with four native species of plant, and each of the tanks had the same number and composition of species (the team, however, expected this composition to change in response to invasion and fertilizer treatments).

While the team required data that reflected the relative isolation and control over variables, as Kate (ID&M-PI) emphasized the researchers also wanted to "make it somewhat realistic." ID&M Team scientists, in other words, felt that they needed data from a system that was, at least to some extent, natural. The mesocosm approach provided the team a means of exerting control while also maintaining the necessary level of naturalness.

The research problem, as defined by ID&M Team, also required measurements that the scientists could use to see changes that occur over a few years' time. ID&M Team researchers, like the researchers on the IM Team, were interested in phenomena that needed at least a couple

seasons' growing time to develop and demonstrate their consequences. For example, the researchers wanted to understand how native vegetation hindered or competed with the growth of invasive species. The seedlings—both native and invasive species—that the researchers planted in the tanks required time to become established. Furthermore, the scientists' questions depended on seeing how factors like biomass, nutrient cycling, and nutrient availability changed over time and in response to variables like *Typha* size, which the scientists expected to change over several seasons. Ethan (ID&M-PD) described some of the data collection activities they planned over the span of the project.

> [. . .] we'll track [basic plant nutrients] over time to see if that changes as we fertilize and as the vegetation kind of fills out. That's what we've been doing this year. [. . .] Next spring, we'll be planting the invasives into all of the tanks. [. . .] then we'll be monitoring the rate of growth of those invasives over the next few seasons to see how they respond to nutrients and how they interact with the native plant community.

The term of ID&M Team's NASA funding was three years (with the possibility of a one-year, no-cost extension): the researchers spent 2011—the first year of their study—constructing the mesocosms. They began data collection in 2012; 2012 data would represent a baseline state for their mesocosms, before the researchers populated them with invasive species. Later, using data they collected periodically over each of the field seasons (approximately March to October, depending on weather), the scientists would analyze data to see changes that occurred in the mesocosms over the years of their study.

Lastly, ID&M Team scientists needed the mesocosm data not only to help them answer research questions related to the mechanisms that control invasion, but also to develop and test the Wetland Ecosystem Model. Mark (ID&M-PI), the lead on the modeling portion of this team's study, developed the initial iteration of the model using data from other field studies and

mesocosm experiments. He planned to use the ID&M Team mesocosm data to further parameterize the model and test its predictions. As Mark explained,

> The model presents sort of a hypothesis about relationships between parameters, and the field study[10] allows us to test whether the results in the field lend credence to a hypothesis about the relationships between parameters as expressed in the model.

The collaboration between ID&M Team researchers who were working on the mesocosm portion of the study (Kate (PI), Ethan (PD), and Chad (RA)) and ID&M researchers whose work was focused on model development (Mark (PI) and Gabe (PD)) yielded advantages to both the modeling and the mesocosm work, influencing the design of the mesocosm experiment and the data scientists collected from the mesocosms.

> This is a great example of a very integrated project where the fieldwork directly informs the model, the development of the model, the parameterization of the model, and then the model directly informs the fieldwork. *(Mark, ID&M-PI)*

Unlike modeling efforts in which modelers rely on data that have already been published, in the ID&M's project the modelers' needs were taken into account as the researchers collected the mesocosm data.

> Mark (ID&M-PI): I think the advantages go two ways. I have a lot of input into how the field study's getting conducted, which is great. And then [. . .] the people that are doing the fieldwork are sensitive to "What are the things that we need for the model?" So there's a real two-way.
>
> Interviewer: What are some differences that play out in how the data are collected that you might not see in publications that *aren't* working with modelers?
>
> Mark: That's a tough one. Well, people might not collect all the things you need for a budget, I guess, would be one example. In this case, we want to be able to do a nitrogen budget, so you need to know what's going in, you need to know what's coming out. [. . .]

[10] That Mark characterized the mesocosms as being "field" in the context of talking about the computational model demonstrates the importance of the relative naturalness of a study system to the meaning of data to scientists. While a mesocosm experiment is not a field study, mesocosms are far less controlled than a computation model; in other words, mesocosm experiments are relatively natural (i.e. "field") compared to computational modeling. I cover this topic in a later section of the chapter.

> There are sort of key pools that are integrative pools that are really good for either developing a model or testing a model, like the nitrogen and foliage is a key one. It's like an integrator. There are ten different processes that control that, and if you measure that one thing it's a really good link between all the different processes happening in the field and all the different processes happening in the model. They're all linked together in this one thing: the nitrogen concentration and foliage. I guess if a group was not working with a modeler, they might not realize that and might not measure that.

The ID&M Team's project depended on data that could be used to reveal the mechanisms that control plant invasion into wetlands and refining and testing the Wetland Ecosystem Model. For data to be valuable for the team, then, they had to be drawn from a relatively controlled system; to cover several years' time; and be comparable that to the data used and generated by the Wetland Ecosystem Model.

5.2.4 SUMMARY

The IM Team aimed to generate results that showed change over time and could be applied to a real world management setting; the NUS Team sought to reveal a narrowly defined mechanism that would contribute basic knowledge of what drives nutrient uptake in streams; and the ID&M Team was working toward understanding the mechanisms of wetland invasion and refining and testing a computational model that could be used to generate probabilities of invasion risk. The overarching purpose of addressing a gap in knowledge framed scientists' notions about what data needed to accomplish, shaping how they designed their studies and assessed the success of their work at the Station. Data's value was always assessed in relationship to how well the data served the purpose of answering a team's questions. Answering a team's research questions was, in other words, a salient purpose throughout the entire data stream, from project planning and study setup to data collection and analysis.

By emphasizing the role of each team's research questions in how scientists assessed the success of their work and conceived of data's value, I do not mean to suggest that the individuals that made up each of these teams were without their own, more personal reasons for participating in their team's research projects. Rather, I argue that answering research questions represented a prominent shared meaning for data around which the teams organized their work at the Station.

Personal goals did come up in the interviews with scientists. These goals included things like to earn money while formulating new career plans (Chad, ID&M-RA); gain research experience and build a relationship with professors (Carolyn and Janet, NUS-UR); publish a piece that would strengthen a tenure case (Elizabeth and Jessica, NUS-PI); and stay connected with research, thereby gaining more legitimacy as a change agent on university-wide environmental policy (Evelyn, IM-PI). However, researchers did not refer to these individual goals when I asked what their data were for or why their team was doing the research. Instead, the researchers—no matter their status on the team—began by describing the shared meaning of data within the group: as resources for exploring a gap in knowledge as defined by the team. That was the meaning that was first and foremost in the minds of scientists as they assessed the value of their data, made decisions that impacted on producing valuable data, and judged the success of their work.

5.3 The Prerequisites for "Good Data"

With data's meaning as resources for addressing a gap in knowledge at the forefront of their minds, scientists enumerated several prerequisites that data needed to meet in order for them to consider the data of value or what they also referred to as "good" or "useful." In delineating their data stream model, Hilgartner and Brandt-Rauf (1994) highlight the importance

of reliability to scientists' sense of data's utility. While the scientists in my study regarded data's reliability—or trustworthiness—as crucial to their value, they also stressed that data needed to be comparable to one another and relevant to their team's specific research questions in order to be of value. I describe each of these traits in more detail, providing examples of some of their instantiations in the different teams.

Scientists emphasized that good data were comparable to one another and, indeed, data that were not comparable to one another were of little, if any value, to them. Data's comparability referred to their equivalence—within a study—across data collectors, time, and individual plots or sites. Often, participants described this comparability as a result of what they termed "consistency" in data collection.

> You have to make sure that the data that's being collected is consistent among individuals who are collecting it and is consistent through time, even the same person. *(Gabe, ID&M-PD)*

In other words, the plant count data that Chad (ID&M-RA) collected from the mesocosms needed to be comparable to the plant count data that the other team members collected from the mesocosms. Furthermore, the plant count data that Chad collected from mesocosm 1 had to be comparable to plant count data he collected from mesocosm 2; and the data he collected from mesocosm 1 today needed to be comparable to data he collected from mesocosm 1 the following summer.

Renee (IM-GR) described consistency as the key difference between good and bad data and emphasized that even when numbers were "slightly off," if they were consistently off, it was often acceptable, particularly for research questions where relative difference was more important than exact numbers. Describing what made data good, Renee said,

> I always refer back to consistency. As long as you're being consistent, that's good enough for me, I suppose.

Matt (IM-PI) agreed, explaining during a day of field data collection, "Relative difference is as important as the numbers a lot of the time."

In addition to "consistent," scientists also used the terms "repeatable," "replicable," and "sameness" to emphasize the qualities that made data comparable. In the following excerpts scientists from two different teams highlight repeatability and replicability.

> Trying to put in checks and balances to make that sure everyone's collecting the same kinds of data and it's repeatable, so having clear standards and checking in with each other and making sure everybody's doing it right. *(Amy, IM-PD)*

> Then you're checking the data to make sure [. . .] that you can replicate it. *(Gabe, ID&M-PD)*

If, for example, Dylan's (IM-RA) percent coverage assessment for *Typha* in a subplot was close to Amy's (IM-PD) assessments from the same subplot, IM Team researchers felt reasonably assured that they were collecting comparable data across the team.

Scientists described challenges in ecological research that made it difficult to collect comparable data. Specifically, natural variabilities in ecological systems could stymie efforts to use the same methods to collect samples or measure variables. Such a situation created considerable frustration for Renee (IM-GR) in her first year of data collection for her master's thesis: "Last year's data were a clusterfuck. I don't know how else to describe it."

During the summer of 2011, Renee (IM-GR) gathered her samples and data from wetland sites that had markedly different water levels from one another. The difference in water levels made it impossible for her to use the same method of collecting macroinvertebrate samples from each of the sites. Additionally, many of her sites had virtually no standing water, and since her dissolved oxygen (DO) meter required at least two inches of standing water to get a reading, she

was unable to collect DO data for all of her sites. Both Evelyn (IM-PI) and Renee likened the data that resulted from 2011's collection to "apples and oranges."

> Renee had a shitty situation last summer. She was trying to collect aquatic invertebrates, and then she gets to this one part of the wetland that's dried up. So she's like, "What do I do? I've got to collect." She's trying to go down into the soil to see if maybe there's water down there and there's still some of these little guys down there, and it's such a different method. It's like comparing apples and oranges. You don't feel good about your data. *(Evelyn, IM-PI)*

> It was just like trying to compare apples to oranges, and it was a third of my plots. Then, because the water levels were dropping too, some of the plots I collected with the stovepipe in June and then in August [I] had to take a soil core. It wasn't comparing directly throughout the summer or across plots or across sites, and it was just kind of a big hassle. *(Renee, IM-GR)*

The issue of comparability was an important factor in scientists' notions of data's value. As I discuss later in this chapter, data's degree of comparability impacted not only whether scientists thought data were good or bad for their own studies, but also undergirded their assumptions about who could derive value from data, for what purposes, and for how long.

In addition to being comparable, to be what scientists would call "good," data also had to be trustworthy or—in the language of the data stream model—reliable. As Gabe (ID&M-PD) asserted, "Good data is data that you can trust." Scientists deemed data trustworthy when they had "high confidence" in them; in other words, when they felt assured the data accurately represented what they were supposed to represent.

Scientists relied on their own prior knowledge and experience to assess whether data were trustworthy. This knowledge and experience helped them to formulate expectations for what the data should look like or what a reasonable numeric range for the data would be. Several interview excerpts demonstrate scientists' reliance on their own expectations to determine whether they should trust the data:

> You're checking to make sure that all counts are right or within an area that you're expecting [. . .] *(Gabe, ID&M-PI)*

> It requires constant vigilance as you look at analyses along the way of making sure something makes sense. *(Kate, ID&M-PI)*

> Jessica (NUS-PI): [. . .] uptake lengths that sort of make sense. They shouldn't be less than a meter, and they shouldn't be 300 meters long.
>
> Interviewer: So knowing what your end result number should fall between?
>
> Jessica: Yeah, and we sort of have . . . We've been doing this long enough from real streams that we sort of have a sense. Our discharge is 100 mils per second, which you can't find in a real stream. I mean that wouldn't be above ground. Most of the time, you're working in much bigger streams. We work in streams that are a minimum of probably ten liters per second, and their uptake lengths are about 100 meters. So if we get an uptake length of ten, that's really not that surprising because our discharge is so low. And we've packed them with leaves. [. . .] We just want a number that's realistic. [. . .] I know we've packed it with leaves, but before it was *so* short it just wasn't realistic for . . . Microbes on leaf litter are not that nutrient starved, so we knew that when we were getting four centimeters and four millimeters that just wasn't right. We've been doing it from real streams long enough that we know [. . .] the uptake length should be about ten, and otherwise . . . If it's a lot longer for nitrate, that wouldn't surprise me because that tends to be the least biologically demanded nutrient [. . .]. But the rest of them should be probably between 10 and 15, you know, basically.

As my exchange with Jessica (NUS-PI) demonstrates, if data deviated too far from scientists' expectations, their confidence in the data was threatened. Consequently, they would try to identify the cause of anomalies. "Erratic numbers" or "outliers" that did not jibe with the rest of a team's data or fit scientists' expectations, could, for example, indicate a basic fault in the study design or setup—such as was the case for the NUS Team—or anomalies could indicate natural variability in an ecosystem. However, anomalous data could also be the result of the peculiarities of one "funky" mesocosm that scientists knew to be problematic or events that happened during data collection.

> We've got one mesocosm up there that leaks, and it's always giving us erratic numbers, and it's a pain in the ass. We're always trying to fix it. There's another one that has all

> these ferns in it, these terrestrial ferns that have come in, and that always gives us erratic numbers. *(Evelyn, IM-PI)*

Lastly, scientists were emphatic that data had to help answer the specific questions that the team set out to answer in their study in order for them to consider the data good. Data could be of high quality—comparable and trustworthy—but if they did not help scientists answer their research questions, they were not "good" for the team. As Amy (IM-PD) explained,

> For any science question the question needs to be clear, and in order to address that question you need to collect data that directly address that question. So good or bad data I think depends on the question you're asking.

Ethan (ID&M) similarly emphasized that good data were data that helped his team "to answer the questions that we're trying to get at." If—as with the photographs he periodically took of the mesocosms (at the request of their partners at a different institution)—data could not be used to answer any of the IM Team's questions, Ethan did not consider them good data.

> One thing that we've been doing is just occasionally taking pictures of all the [mesocosms]. And there's an *enormous* amount of data in those pictures, but I don't really know how to use those data. You could look at the total amount of greenness in the tank, but there are different species in the tank, and really what we're interested in is understanding how the different species interplay with each other. And you can't tell that from the picture in any sort of automated way. So those wouldn't be very . . . I don't think of them as good data. [. . .] What I think of as good data is something that we can use to answer the questions that we're trying to get at; that provide useful information to answering those questions. [. . .] relevant to answering the questions. When I think of good data, I think of data that answer the primary questions that are in my head.

Scientists, in other words, did not consider data to be good unless they were useful to their teams; and that usefulness was highly dependent on whether data addressed scientists' *specific* research questions.

The NUS Team's work presents an especially striking example of the way in which data's value was assessed on the basis of whether or not scientists could use them to answer their research questions, primarily because the researchers were challenged in their efforts to produce

"good" data. NUS Team researchers' alterations of their experimental setup and their later decision to abandon their project at the Station stemmed from the scientists' assessment that the data they were collecting could not be used to answer any of their research questions, as Jessica (NUS-PI) explained.

> The problem with [the data we were getting before] is that that wasn't the microbes on the leaves that was [taking up nutrients], it was algae and microbes and all that fine particulate organic matter, so that's why we had to switch to groundwater last Wednesday. Because that's not our question. [. . .] that's not what we're interested in. That wasn't the whole point of why we built all these experimental channels: to grow algae and fine particulate organic matter of unknown C to P to N ratios. *(Jessica, NUS-PI)*

The problem for the team was *not* that the data did not support their hypotheses; rather, the data did not address their hypotheses and associated research questions.

The assertion that data must be comparable, trustworthy, and relevant to scientists' specific research questions in order for scientists to consider them good might appear, on its face, obvious. However, what I have highlighted so far in this chapter is scientists' recognition that data's value is relative and depends on a match between what scientists need the data to do and what the data are actually capable of doing. Scientists regarded data primarily as resources for addressing a problem as laid out in their study designs; and the particular questions they set out to answer and/or the hypotheses they aimed to test necessarily limited what they considered to be good data: data that were worth the considerable time and effort required to produce them.

Scientists' conceptions of data's value, however, encompassed more than whether data were good or bad; of value or not of value. Scientists' value conceptions also included assumptions about how long data would be valuable, who could or should realize value from data, and the kinds of uses data could be put to. In the next part of this chapter, I delve more deeply into these nuances of scientists' value conceptions by describing how scientists used data type as a basis for data valuation.

5.4 Scientists' Use of Data Type to Make Value Assessments

In making value statements about data, the scientists I studied frequently made distinctions between what they considered to be different types of data. Scientists' characterization of different data types was made on the basis of the kind of study system from which they collected data (experimental and modeling vs. field data), data's publication status and potential (published vs. unpublished; and publishable vs. unpublishable), and data's level of processing (raw vs. derived data). For the scientists, each type of data carried with it particular assumptions about data's value, including what the data could be used for, by whom, and for how long. For example, scientists explained to me that "experimental" data were likely to be of less value to other scientists and questions and would probably of value for a shorter duration of time than "field" data.

In this section I detail the dimensions along which scientists characterized their data as they considered data's value. In each sub-section I focus on one dimension (e.g. the study system that data represent), detailing the typology that resulted from it (e.g. field data vs. experimental or computational data). In doing so, I describe the salient aspects that played into scientists' characterization of data according to that dimension and delineate, as applicable, the related assumptions about what scientists thought the data would be valuable for, how long they thought the data would be valuable, and who should or could reap value from the data.

5.4.1 The Study System that Data Represent

> *When it comes to field data, really anyone can be interested in that. Anyone who has an interest in, let's say, time effects on species composition or a [land] manager. I can see a manager in [. . .] one of the places we're going to work at being like, "What was this place like 15 years ago? Because it's so much different now" using that dataset. [. . .] But things like experimental data like the mesocosms . . . meh, I don't know. It's tough to see*

> *how it's going to be that much more useful past the point of what the people who were doing that experiment really want the data for.* (Gabe, ID&M-PD)

In the passage above, Gabe (ID&M-PD) indicates that "field data" and "observational data" (in this instance he uses the terms synonymously) have virtually unlimited value in terms of the time period over which they might be valuable and who might find them valuable. Gabe contrasts field data with "experimental data," which he asserts will likely only have value for the specific question(s) that resulted in their generation and the individuals or team of individuals who collected them. Gabe's characterization of data's value according to the type of study system the data represented was indicative of a view shared across researchers and the three teams: experimental data (and—in the case of the ID&M Team—modeling data) were finite and exhaustible resources, and field data were potentially expansive and inexhaustible resources. This presumption was based on what scientists described as the artificial and ephemeral nature of controlled studies and the more natural and enduring qualities of field study systems.

Whether researchers appended data with the qualifiers "experimental," "mesocosm," or "modeling," they contended that data collected from contrived or controlled study systems held more limited value across time and scientists than did data collected from natural field systems. For example, Matt (IM-PI) thought that the data from his team's mesocosm study would likely be of little value to other scientists, explaining that his team was "squeezing as much out of those data as we can" and that the number of additional questions one could ask using the same data was "minimal." Similarly, Mark (ID&M-PI) doubted that other scientists would find his modeling data useful: "I don't see what they would have to gain. [. . .] I don't think other people would see [reusing the modeling data] as being a valuable thing."

Scientists viewed the systems that were the basis of controlled experiments as purposefully constructed, transitory resources for addressing their specific research questions,

and hence thought data they generated from such systems were valuable only within the context of the particular study and to the team. The researchers described the systems (whether vinyl gutter streams or wetland mesocosms) as "designed for questions" and frequently contrasted them with "field" or more natural ecosystems, such as the marshes where IM Team researchers carried out their wetland restoration study.

Because experimental and computational data were collected from systems that scientists deliberately constructed to answer specific questions, they viewed such data as valid only from within the bounds of the system. For example, Gabe (ID&M-PI) described model-generated data as having a relatively "short shelf life," because as the model developed and changed the data would become, in important ways, outdated. The data might continue to serve as a record of what the team learned from a particular iteration of the model, but they would have no relevance outside of that iteration's bounds.

> The Wetland Ecosystem Model though, on the other hand, is different [than a field study] because the things you produce are important and the results are important, and that publication will be important, and that won't change. But as you . . . Let's say as you change the model itself and as you increase its realism maybe and other things, that could force some of the other data . . . I don't want to say "obsolete," because it's still important within the bounds that it gave it. But I think you're going to be more interested with the more complex, maybe more descriptive model. So I can see maybe the Wetland Ecosystem Model stuff having the shortest shelf life, but the results from the papers that come out of that, not having a shelf life. That's going to be important because the model's doing a certain thing at this certain time. But as the model hopefully progresses and gets more complicated, maybe has different aspects of it that are incorporated, like light competition or something . . . You might not be as interested with the old data anymore. *(Gabe, ID&M-PI)*

Similarly highlighting the deliberately constructed aspects of controlled studies and the resultant limitations on the value of data produced from them, both of the PIs from the NUS Team repeatedly stressed that the data they were generating from "artificial stream channels" would have less value to others than data they collected in "real streams."

> Interviewer: So do you think that you could actually see going back and using these actual data to answer new questions?
>
> Jessica (NUS-PI): Probably not for this summer; not in this sort of experimental setup. It's really designed for two questions, and so I probably wouldn't . . . I could see that with field data, but not with these experimental data.
>
> Interviewer: And why is that? What's the key difference?
>
> Jessica: I think with an experimental setup, I feel like the *whole* experiment was designed around two questions, and you don't have other sorts of variabilities. You don't have differences in ambient concentrations, or differences in site, or differences in channel dynamics that . . . you know they're all . . . it's all for the exact same thing; all designed just to answer two questions.

Researchers' descriptions of the limited value of data from controlled experiments also alluded to the systems' transitory, or ephemeral, qualities. Unlike field systems, such as streams and wetlands, scientists viewed controlled experiment systems as relatively short-lived and unlikely to persist after the team was finished with them. One way to conceptualize such systems—which were important components of each team's data stream—was that they were rare; accessible only to the team and generally only for the term of a particular team's study. The length of time scientists planned to conduct studies using their experimental setups varied. For example, NUS Team researchers planned to run their experiment for approximately eight weeks, while ID&M Team researchers thought they would keep asking questions using the mesocosms for approximately ten years, contingent on continued funding. In both cases, however, scientists viewed their systems as subject to disassembly or, at the very least, benign neglect once they had exhausted them for their own questions. Both Kate (ID&M-PI) and Matt (IM-PI) described the contrast between the value of field and mesocosm data by pointing to the transitory nature of mesocosms.

> We know when cattails first came in there, because [Person Name] was taking her class out to [Marsh Location Name]. The class data on what that vegetation was . . . that's a *really* useful thing. The mesocosm data, in a way, is less useful in that—now that I think

> about it, is less useful in that regard, because it's *not* some site that somebody could go back to. It's most likely [. . .] once we finish with the mesocosms that they're not going to be really usable for anybody else anymore. *(Kate, ID&M-PI)*

> It seems to make a lot of sense that data that are collected about the ecosystems on the Station—these long-term changes in ecological conditions on the site—that seems like that could be really valuable for the Station, but these mesocosms are this sort of ephemeral thing. We put them in, they're changing really rapidly; most people take them out. *(Matt, IM-PI)*

Because the systems that were the basis of more controlled experiments were ephemeral, other scientists would not have the opportunity to conduct their own studies using those systems and therefore would be unable to expand on the data with additional variables of interest or to compare differences across time. Again, Kate (ID&M) contrasted the ephemerality of the mesocosms with the longevity of field sites and pointed out the implications:

> They're not in the field. They're mesocosms, and we're not going to keep them, and therefore, there's no basis for keeping them going indefinitely. At some point, they'll just [. . .] have so many artifacts because they've been growing in this little thing for so long that we'll say, "It's just no good anymore." We will stop collecting data. Whereas for field data, you can always go back to it. Unless you destroy the plots, you can always go back to them. And even with a big gap, you can still say something interesting. I think for the mesocosm data it's much more likely that you've got a fixed period and that once you've analyzed that in a lot of ways, you can run out of interesting things to do, whereas you could always add a longer time period for some of these long-term data if it's set up well.

In addition to the artificial, or contrived, and ephemeral qualities of controlled study systems, several scientists pointed to experimental data's lack of utility for metaanalysis studies as a limiting factor in the value of their data. Metaanalysis, a method of inquiry in which scientists bring together data from several studies to answer questions, is increasingly common in ecological science (Gurevitch, Curtis, & Jones, 2001). For example, ecologists employed metaanalysis to examine coral decline in the Caribbean from 1975 to 2000 (Gardner, Cote, Gill, Grant, & Watkinson, 2003). Combining data from 263 sites and 65 different studies, scientists were able to tell a bigger story than any one of the studies could tell individually and were also

able to make a more compelling case that their findings were not the result of natural ecosystem variability.

Several of the researchers I studied argued that other scientists would be unlikely to use experimental or modeling data to conduct such metaanalyses and, in doing so, contrasted them with field data. For example, Mark (ID&M-PI) asserted that his modeling data would be of little interest to others because it was inconceivable to him that the data would be used in any kind of "cross-project metaanalysis."

> [. . .] if you generate field data and made that available through a website I could see that being useful to other researchers. Other researchers might do a data mining exercise where they try to gather field data from fifty different field projects from around the world and do some new kind of analysis. That I could see being very useful. But I wouldn't see someone using my model for that kind of cross-project metaanalysis.

Similarly, Elizabeth (NUS-PI) viewed field data from "real streams" as good candidates for synthesis across studies, but she saw very little potential for synthesis between her team's experimental data and others' data. In the excerpt that follows, she explains why.

> I think it's the sort of real life study system. If someone was going to do a synthesis paper, they're going to synthesize nutrient uptake in streams, and they are not going to include our little channels because little channels aren't real streams. And so I think, while that data might be potentially the same as the stream, our numbers themselves . . . they might differ in magnitude to what you see in a stream. That difference in magnitude is okay [for our study], because we are comparing channels to channels in our particular project. But it is not going to be okay if you are doing a synthesis and comparing streams to streams. If our numbers are different, you would just say, "Well of course they are different; they were done in channels."

Notice that Elizabeth did *not* assert that other scientists would be uninterested in combining the NUS Team's experimental data with their own experimental data. Rather, as she considered whether or not someone would use her data to create a synthesis paper, she immediately focused on the potential combination of her experimental data with others' field data. I would learn in later interviews and observations with the team that, as far as the NUS Team PIs knew, no one

else was conducting experiments that would yield similar data to what the NUS Team researchers were generating that summer. In other words, there were no data that would be suitable candidates for synthesis because there were no comparable data. The same prerequisite to scientists' notions of good data for their studies colored their assumptions about who else might derive value from their data for their own purposes.

As the interview excerpts I have provided thus far make clear, in considering their experimental data's value, scientists frequently positioned data from controlled study systems in contrast to data from field systems. They described experimental and modeling data's value as limited: not of much potential use to other scientists and sometimes not even valuable for much longer past the point of the data's publication. Scientists described field data's value, on the other hand, as expansive, imagining that the data could be used by others and would remain valuable across a virtually unlimited amount of time.

When researchers talked about field data, they frequently emphasized the realness of the systems from which they collected the data. For example, NUS Team scientists characterized field data they collected in a previous study as data from "real streams," "real-life systems," and "natural systems." IM and ID&M Team scientists called field data "data from ecosystems" or "data from the forest." Sometimes scientists highlighted field systems' realness by simply qualifying data according a site name (e.g. Jackson Marsh data) as opposed to the experimental equipment (e.g. "data from the cattle tanks," "channel data," or "data from these gutters"). That field data were collected from actual ecosystems, as opposed to constructed imitations, meant that they captured information about some aspect of an ecosystem at a specific place (e.g. macroinvertebrate biomass in Marsh X) and time (e.g. the summer of 2012). Highlighting field data's status as representations of an ecosystem (or aspects of an ecosystem) at a particular place

and time, scientists regarded such data as a historical snapshot. As the Station director explained to me, "It's basic data. It's almost museum-type data." That field data were tied to an ecosystem at a specific place and time meant, in other words, that the data were unique (representing something that could not be recaptured). Furthermore, as representations of real ecosystems, scientists could conceive of other scientists taking their own interest in the study system. For example, at least one of the marshes that IM Team researchers were studying had served as the site for research that Kate (ID&M-PI) and other Station scientists had conducted in previous years.

Another trait of real ecosystems differentiated them from their constructed counterparts: scientists expected them to endure into the conceivable future. Scientists assumed that unless they "became shopping mall[s]" (Gabe, ID&M-PD), their field study sites would remain accessible to future study. In contrast to controlled experiment systems, scientists viewed field systems as a more communal resource, not "personal resources" (Dylan, IM-RA) for particular scientists or teams of scientists. Field systems were—in contrast to controlled experiment systems—common resources, likely to be utilized by other scientists, rather than rare resources created and utilized only by the team that constructed them. This meant that other scientists would have the ability to revisit the site, and this played into scientists' notions about who could potentially reap value from their data and for how long. For example, IM Team researchers created a GIS (Geographic Information System) dataset that depicted the age of the *Typha* stands in one of the marshes they were studying. The researchers found that *Typha* stand age was an important determinate of whether a wetland could be restored. But, as Matt (IM-PI) explained, the data his team collected on *Typha* stand age in that particular wetland could be of value to many other scientists interested in questions that were beyond the scope of his team's research.

> [. . .] the spatial data particularly—it could be interesting to people, because we have these very clearly aged cattail stands there. If anyone else wanted to look at some other environmental factor or whatever . . . maybe if somebody was interested in insects, and they could go out and use our maps as a way if they wanted to look at differences in the insect communities with different ages. Or pick your taxa or whatever. I could see that being a useful starting place for additional research that's kind of beyond the scope of what we're doing [. . .]

The expected endurance of the sites that comprised field studies was important in scientists' assumption that field data had broader, longer-term value. However, endurance is not synonymous with permanence. Ecosystems are dynamic, and it is this dynamism (e.g. an ecosystem's reaction to specific disturbances) that underlies much of ecological research, also playing a large role in scientists' view that their field data could be of interest to others for a long time. In the following interview excerpt, Gabe (ID&M-PD) describes why field data have a "very, very long shelf life."

> Gabe (ID&M-PD): I think the field data could [have] a very, very long shelf life. The stuff that we haven't even really—not the mesocosms. Because if you really think about it, if you have this—
>
> Interviewer: Oh, the stuff you haven't collected yet?[11]
>
> Gabe: Yeah. That kind of stuff's going to be important, because you can monitor that forever. As long as [the site] doesn't become a shopping mall. [. . .] You can measure and track species composition throughout time. The data we have when we do species composition and estimate biomass; that could be good forever. Not forever, but for a very long time. Because anyone could be interested . . . Like, wow, twenty years from now, in the future, someone will look back and [be] like, "Twenty years ago, this was here. Now this new species is here. That's interesting. Let's try to figure it out." So *that* kind of data is useful for a long time.

The participants often referenced the study of change over time as a key way in which other scientists might find their field data valuable. It was the study of ecosystem change over

[11] The ID&M Team originally planned to collect field data in addition to the mesocosm data. At the time of my study, the PIs were no longer sure that a field study would be feasible under their current grant.

time that Jessica (NUS-PI) explained made her field data—but not her team's experimental data—good candidates for deposit in a repository.

> Interviewer: Do you think that [the data your team is collecting this summer] would be worth long-term preservation in a repository? [. . .] Would it make sense from your perspective?
>
> Jessica (NUS-PI): This is totally [. . .] the [. . .] ecologist in me or whatever . . . is that yes, my field data, because field systems change over time and are subject to environmental perturbations. And these are gutters.

Similarly, Dylan (IM-RA) described the "long shelf life" of field data by way of recounting his own team's experience several years ago using data that one of their team members (Phil, IM-PI) had previously collected at the same marshes.

> One of our studies, we were working with Phil, and he had data that he had collected in the 1980s. We compared that—because he was surveying the same marshes as us—compared that with what we were doing in 2007 and came up with a really good, publishable manuscript that showed a lot of stark contrasts between before and after *Typha* invasions in wetlands in the same areas that we have surveyed. So [field data] has a really long shelf life.

The IM Team researchers found value in Phil's old field data, because those data (in conjunction with data they gathered themselves) helped the team to show changes in the local ecosystem that were associated with *Typha* invasion.

When I asked Renee (IM-GR) if she thought her data on wetland macroinvertebrates might be useful to others, she also emphasized the opportunity to study change over time.

> [Location Name] wetlands are really kind of an endangered ecosystem, but we just don't know that much about them. So if people were doing work in these wetlands, my data could be used as kind of a baseline of, "This is what was here when the water levels were this high." And maybe, if the water levels have changed, things have changed, or this new invasive has come in, or this invasive has disappeared, how has it changed from this baseline survey and stuff?

It is important to emphasize that the long-term comparisons that scientists mentioned when arguing that field data might be useful to others and for a long span of time did not

necessarily depend on regular monitoring throughout time. Even with a gap of several years in which no data were gathered about a particular aspect of an ecosystem, a scientist could conceivably return to a site and study differences that were worthwhile. The director offered a compelling example of this type of potential for field data.

> Station Director: There are datasets that we're developing now that I would anticipate would be useful for long-term studies. For example, [Personal Name]'s work with earthworms. She studied the spatial distribution of different earthworm communities, all of which are invasive in [this area]. And she's going to write a paper, but we want to be careful that we can use those data to lay the groundwork for another study that might be done in 10 years or 20 years or 50 years, because these organisms have only been [here] for 50, maybe 70 . . . we're not quite sure how long. We know they weren't there 100 years ago. We want to make sure that her dataset is available for long-term studies. Even for her. I mean, she's young. She'll finish her PhD this spring. She might want to come back with students in 20 years and say, "Hey, these are the plots I sampled. Let's see if these things are still here or if their abundances have changed."
>
> Interviewer: What do you see them being valuable for? In this broad sense, what are the kinds of things you can see—
>
> Station Director: Change detection; detecting changes in the abundances and the composition of biological communities; changes in resource use patterns; changes in the functioning of ecosystems; changes in biogeography of local species, you know, what their distributions are; monitoring of invasives—not only the presence and abundance of invasives, but the potential effects on ecosystems and human communities.
>
> Interviewer: Okay, so kind of change over time?
>
> Station Director: Right, change over time and space.

As the director noted, field data are not only valuable for studying change over time, but also for making comparisons with other, similar ecosystems. For example, a scientist could be interested in differences in nutrient uptake between streams in areas with a lot of agriculture (e.g. sites characterized by high levels of nutrient runoff) and relatively unimpacted headwater streams. Or, perhaps a scientist would be interested in comparing similarities and differences across many streams in agriculturally influenced areas. In both cases, one team's field data could serve as a component for an analysis that told a "bigger story" than any individual dataset could

tell on its own. Describing who might be interested in his team's field data, Dylan (IM-RA) argued it was their field data's capability to aid in cross-site comparison that made them potentially valuable to other scientists.

> Certainly, anybody who's studying *Typha glauca*, anybody who's studying invasive plants, anybody who's studying biogeochemistry . . . You can look at data from a variety of different sites if you've got your own project and you're using similar methods to study a similar system, you can compare it with ours. It's definitely useful.

Similarly, Matt (IM-PI) thought a "bigger story" could be told if his team's field data were combined with another team's field data.

> I think some of the other marsh vegetation data would be really valuable to a very small subset of other scientists as almost like a monitoring dataset. [. . .] We've collected similar veg. data over many years and it's changing with time, and if you were to take all of the data we've collected from the [Location Name] and [Location Name] and put it in the context of other people's data, you could create a bigger story, I think, with those data.

And, while her team was not currently collecting field data, Elizabeth (NUS-PI) observed that were they conducting their study in "real streams," their data might be useful for creating a synthesis paper.

> If we were doing something like this in a stream, or like what we did last year, I think the useful life of that data is a lot longer because I feel like, pretty soon—in the next five years—someone is going to do a big synthesis paper, and then they'll be reusing our data or this type of data [we collected in real streams] in new ways.

Scientists drew stark comparisons between the value of their experimental and computational data and the value of field data, making it clear that, in their view, field data had broader and longer-term value than did data from controlled study systems. The most fundamental difference that the scientists described was the potential to expand on the data with additional variables or data from other studies or points in time. This property of data streams or data stream elements is notably absent from Hilgartner and Brandt-Rauf's model, which instead emphasizes reliability, processing state, and rarity in data valuation and access decisions. The

rarity of the study system somewhat explains scientists' view of controlled experiment data as having less value to others than field data. If the system was inaccessible to others, how could scientists possibly add to the data drawn from it? However, scientists were much more attuned to data's expansibility and what they saw as their resultant ability to answer new research questions when they talked about differences between field and experimental data's value than they were to data elements' rarity.

In the next section, I detail how scientists used data's publications status and potential as a way of characterizing differences in their data's value.

5.4.2 Data's Publication Status and Potential

> *[The Information Manager] was just trying to identify some datasets we felt comfortable sharing, essentially. Then we came to . . . We decided on one particular dataset that we'd actually already published. [. . .] It was published, and we didn't feel like there was any other real use for it. I mean, there's probably some more information you could get out of it, but, for the most part, we did what we thought we could do with the data.* (Matt, IM-PI)

Another way that researchers conceived of differences in types of data was according to data's ability to serve as resources for generating peer-reviewed publications. As a result, they made value distinctions according to data's publication potential as well as their publication status. This way of thinking of data's value was especially salient among junior faculty and post-doctoral researchers, who had tenure review or a competitive job market looming in their futures, but was also prevalent among tenured faculty and non-tenure-track researchers who understood publication as the primary output of their team's work.

In the excerpt that opens this section, Matt (IM-PI) describes a "published" dataset that, in the view of his team, had little remaining potential for generating more publications. Scientists' frequent use of the descriptors "published" and "unpublished" to describe differences

in data's value indicates that data's publication *status* was the primary factor in publication-related value conceptions. Scientists certainly wanted their teams to be the first to publish from the datasets their team had produced and regarded their unpublished data as an unexploited resource; hence unpublished data were viewed as particularly valuable to the researchers. However, as Matt's statement above exemplifies, scientists were also concerned with data's publication *potential*. Did the data hold potential for the researchers to generate publications (i.e. were the data "publishable"?)? Or had the team already exhausted most of that value (i.e. were the data "published-out" for the team?)?

In my observations of each team's work, publication was a regular topic of conversation among researchers as they collected data and discussed preliminary results. In these discussions, the researchers speculated on whether data would be good enough for publication, or, in other words, whether the data were publishable. IM Team researchers, for example, talked about their data's potential for publication before they had even transformed what they saw at their field sites into raw data. Pleased with the variety and number of native plants the researchers saw in their treatment plots, Matt (IM-PI) was quick to make an assessment of the (not-yet-collected) data's utility for generating publications: "This is money; we can probably use these data in a paper." In other words, scientists projected forward to data stream elements they had yet to produce—publications in this instance—to assess the value of elements in earlier stages. Mark (ID&M-PI) similarly explained that data's publication potential loomed large as he conducted model runs.

> As I'm generating the data, we're thinking about what paper we're going to write, we're thinking about what changes we might want to make [to the model] and what some of the next steps are. We're definitely thinking about that as we're generating the data.

Primarily because they were analyzing data as they collected them, the NUS Team researchers' assessment of data's value on the basis of how well data would serve the team in

authoring publications was especially striking. Elizabeth (NUS-PI) and Jessica (NUS-PI) came to the Station expecting their work to yield an article worthy of publication in *Ecology*, one of the most highly regarded ecological journals. As the project progressed, the PIs periodically discussed—between themselves and with me—whether or not their data were good enough for the publication they envisioned. On one such occasion, Elizabeth described what she referred to as the "costs" of being at the Station that summer; namely that she was going without summer pay and was unable to work on other projects. When she asked, rhetorically, "Is it really worth it?" I suggested that when their experiment began to generate more data the researchers would see a payoff to all their hard work. Her response highlighted the importance of publishability to scientists' notions of data's value:

> Unless it's crappy data. Unless it doesn't work. Then we have a crappy paper. Is it really worth it for a crappy paper? I don't think so.

The need to generate publishable data in order to justify the NUS Team's time at the Station was a prominent theme in my interviews with Elizabeth (NUS-PI) and Jessica (NUS-PI). Elizabeth explained that the value of the data was directly connected to the publication outcome of the project.

> Elizabeth (NUS-PI): Part of the value of this data is very much tied to the publication outcome of it in terms of thinking about the costs and benefits and my energy invested in this project.
>
> Interviewer: So the best outcome in terms of the investment of your time in creating these data would be?
>
> Elizabeth: Publication.

In a separate interview, Jessica (NUS-PI) made the same point more starkly, asserting that a project that did not generate publishable data was not worth the team's time, particularly given the fact that she and Elizabeth (NUS-PI) were not yet tenured.

> Jessica (NUS-PI): [Elizabeth and I are] not at a place in our careers where we can spend the summer futzing around with an experimental setup that doesn't work. That's not worth our time. It's not worth being away from your family for that. It's not worth . . . I mean it's too much work to just—
>
> Interviewer: If you're not going to get—
>
> Jessica: If you're not going to get a publication out of it, it's not worth your time, right?

As I have just described, scientists made assessments of data's value, in part, on the basis of whether they thought the data were worthy of publication. When data were deemed publishable, the peer reviewed journal article became the primary vehicle through which scientists were able to realize data's value. Of course, the scholarly journal article occupies a central role in the communication and validation of research results. As Matt (IM-PI) put it to Renee (IM-GR), "If you don't publish it, it's like it never happened." But additionally, as a key measure of a scientist's productivity and impact on his or her field, peer-reviewed publication was the payoff for the significant time and money scientists invested in producing data. This had significant implications for the different ways scientists conceived of the value of unpublished and published (especially published-out) data, including who should reap the benefits of data's publication value and the time span over which data were valuable.

Assuming data were publish*able*—a trait that for the researchers was tied to data's ability to answer research questions—scientists regarded unpublished data as valuable resources that had yet to be exploited for the production of peer-reviewed, scholarly journal articles. Further, scientists described their unpublished data in ways that indicated they thought of data's use for creating publications as finite and theirs to exploit. Until data were fully exploited by the team, researchers were concerned with maintaining their data as a rare resource: one that was only accessible to the team and afforded the researchers a possible competitive advantage over scientists who did not have the data. For example, the following statements from researchers

from each of the teams represent just a sampling of what was brought up repeatedly in interviews with scientists:

> In general, people are hesitant to share [data] with the public or people who aren't part of the team until it's published, which is kind of standard. *(Amy, IM-PD)*

> We haven't published [the data] yet, so it's not available for anyone else. Because we need to get it published first. *(Jessica, NUS-PI)*

> We will have a lag because, again, especially Ethan and Gabe need to be getting papers out from [the data]. *(Kate, ID&M-PI)*

> Usually, [data sharing happens] after people publish it, just because people are protective over their data. *(Renee, IM-GR)*

Scientists' fixation on "protecting" unpublished data from other researchers demonstrates an assumption on the part of the scientists that the release of unpublished data outside of the scholarly publishing regime would threaten their own ability to realize an important goal of their hard work. To the scientists I studied, unpublished data were resources with unrealized value. If data were made common—for example, through deposit in the Station repository—teams would no longer have the exclusive right to exploit the data to their own advantage first. But perhaps more importantly, when scientists talked about their unpublished data, they described them *not* as resources that were endlessly exploitable, but as resources whose value for generating publications was depletable.

Mark (ID&M-PI) depicted publication as the necessary outcome of his team's investment in setting up their study and producing data and also indicated what was at stake were his team *not* the first to publish data from the model they developed.

> Even if [it] takes a few years, we have to publish a few publications first. We have a lot invested in it, and you [. . .] need the publications to sort of be the outcome from that investment. And I want people to know this is the model I developed. So if there ends up

> being ten different research groups that use this [model] down the road, I want people to know: this is the model [Mark Smith] developed. The first few publications have to be my publications. If I gave a copy of [the model] to someone and they said, "This is really great" and [they] did a whole bunch of model runs and published something with it, and that came out *before* my publication came out, that's not okay.

Matt (IM-PI), who was responsible—along with Amy (IM-PD)—for managing his team's data collection efforts, analyzing the team's data, and making sure the team was generating publications, also described the importance of turning unpublished data into publications.

> [. . .] the most important thing is to get it published. So [. . .] if we're stacking up publications and it's like the data from this individual project would fit into a publication, and then, once it's published then the data are out there and we can return to them if we need to, but it's mostly just like trying to keep processing this stuff and moving it along.

To scientists, data that remained unpublished were like money never spent: valuable means to an end that was never realized. Scientists' view of unpublished data as an untapped resource whose value had yet to be realized meant that the trajectory of data's value was defined by the event of publication. For some researchers, the time that it took to reap value from unpublished data in the form of publication was of minor concern. As Mark (ID&M-PI) said, "even if it takes a few years." For other researchers—most notably those whose tenure review or entry into the academic job market was looming in the near future—it was important to turn unpublished data into publications within a relatively short timeframe. For example, one of the postdoctoral researchers for the ID&M Team, Ethan (ID&M-PD), expressed frustration with the fact that his team had yet to generate enough data to produce publications. Planning to apply for tenure-track faculty positions in the fall, Ethan described the time and labor he had invested in the project and the relative lack of value he had yet to see from that investment.

> Ethan (ID&M-PI): [. . .] [this project is] the most work per unit of data so far because there *is* so much manual labor setting these things up, and there are a lot of costs going

> into setting these things up. I mean a hundred cattle tanks[12] is $20,000 right there. It is an *enormous* cost [. . .]
>
> Interviewer: [. . .] Is that something that you think about when you think about the data?
>
> Ethan: It's not something that I have done any formal thinking about like, "Okay, should we do this experiment? Here's the amount of work it's going to take, here's what we're going to get out of it," but I have thought about it a lot [. . .] less cerebrally and more emotionally. It's sort of irking me that after a year of work, we have very little data. It's becoming kind of a strain on me that I have to start looking for jobs now, and I don't have much to show for the last year. It's not that it's not useful work. It's been a useful year, and I think a lot of data will *eventually* come out of it. It's just that initial lag time is really unfortunate for my timing.

As Ethan explained, the work his team had engaged in thus far was not useless, but until the team generated publications from the data, he would have little to show potential employers for it.

Scientists viewed unpublished data as resources with the potential to yield peer-reviewed journal publications, and, as a result conceived of them as a valuable resource yet to be tapped. Conversely, researchers depicted published data as resources that had been exploited and that were then secondary to the peer-reviewed, published products that the data yielded. In talking about published data or the process of publishing data, scientists frequently used metaphors of extraction that indicated they regarded data as objects out of which publications could be "squeezed," "extracted," and "got out." As a result, scientists' view of published data's value depended on whether they thought their team had fully used up the data's potential for generating publications.

Some scientists argued that data's usefulness to the team diminished to virtually zero upon their publication, particularly for data—like the NUS Team's controlled experiment data—that were assumed to have no value beyond the questions that inspired their creation. Jessica (NUS-PI), for example, said the team's data would "be useful to us up until we publish them." As

[12] Ethan frequently referred to the mesocosms as "cattle tanks," which was what the metal tanks were designed and sold for.

both she and Elizabeth (NUS-PI) explained, after the data were published, the publication—and the results presented within it—would remain valuable for other scientists interested in nutrient uptake in streams. The data themselves, however, "would not be very meaningful" (Elizabeth). Mark (ID&M-PI) expressed a similar perspective on his team's modeling data.

> Interviewer: If you had to say how long the useful life of the data that you're generating with the model [. . .], how long do you think the useful life of the data is? [. . .] How long is the useful life of the data to your team?
>
> Mark: You mean the model itself or the data being produced?
>
> Interviewer: The data that are being generated through the model.
>
> Mark: Well the idea is to get peer reviewed papers published. So as long as it takes to get peer reviewed papers published, that's when the data are useful. Once the paper's published, that sort of becomes the story. And if I wanted to go back ten years later, to try to remember what that model was showing, I probably wouldn't go back to the model run [data] themselves. I'd go back to the paper. That sort of becomes the record.

These scientists not only realized the value of their investment through publication, but they also viewed the publication itself as the most valuable version of their work. As Mark (ID&M-PI) explained, even as the producer of the data, he was more likely to revisit the published paper rather than go back to the data. Importantly, Mark's team's modeling data and the NUS Team's data were drawn from controlled study systems, which, as I delineated in the previous section, led researchers to view their data as a resource with limited, short-term value. What I emphasize here, however, is scientists' use of publication status as an additional framework for conceptualizing data's value.

As my observations of and interviews with scientists made clear, researchers viewed publication as a process by which they extracted the most interesting findings from a set of data, developed a set of conclusions, and then presented them in a vetted form. Data, then, were valued for means for getting to the more value-added product. In talking about his perception of

the value of his team's modeling data to other scientists Mark's (ID&M-PI) characterization of published data exemplifies this perspective well.

> I don't see what [other scientists] would have to gain. If I do 10,000 model runs and we would write a paper from that . . . that would be extracting the most interesting conclusions from that set of model of runs. And we publish a paper. Other people that are interested in the research would read the paper. If they get the results of the 10,000 model runs and expect they're going to do something with it or learn something from it, I just don't really see that. I don't think other people would see that as being a valuable thing.

As Mark described, he expected his team to "extract" the most interesting things from the data and put them into the paper. Going further on this point, Mark emphasized in the same interview that the most useful information was to be found in the publication.

> The idea is to put what's useful in a peer-reviewed publication and have that in the open literature.

Similarly, Matt (IM-PI) explained that his team's mesocosm data would likely not be very useful to other scientists because his team was already planning to "squeeze" as much out of the data as they could and that his team was telling "the story" of the data through the publications.

> We have a lot of the soil through time, and I think that we are basically squeezing as much out of those data as we can. Maybe I'm wrong. I'm sure that there are some statistical wizards that could come in there and look at it in a different way and probably make more sense of it, but I think we're telling most of that story through the various publications that are coming out of it.

When characterizing their data's value according to their publication status and potential, scientists often referred to published data in ways that indicated they thought the data's value had been used up with a single publication. However, several scientists also described data that they still considered valuable after publication because they surmised their team could continue publishing from them. For example, Gabe (ID&M-PD) told me he thought his team's mesocosm data would be remain valuable for at least a decade during which they would be able to generate publications.

> [. . .] that dataset [. . .] I think will have a shelf life of, like, ten years, where you can really extract a lot of good stuff out of it and maybe get some more publications.

And Kate (ID&M-PI) told me about mesocosm data that she had collected nearly a decade ago that she was still going back to in order to write more publications. Likewise, Matt (IM-PI) explained that his team had several datasets from which they were still generating publications. Importantly, however, when scientists were making publication-based value distinctions they also frequently suggested that data they had tapped out of publication potential were tapped out for all scientists. Occasionally, researchers allowed that, "perhaps there are some statistical wizards that could come in there and look at it in a different way," (Matt, IM-PI), but more often they characterized data with low publication potential to the team as having similarly low value to others because—rightly or wrongly—they assumed the data's value for answering questions had been exhausted.

As I alluded to earlier, the issue of data's publishability was more important to some researchers' notions of data's value than others. While all researchers recognized peer-reviewed publications as the primary output of their team's research, those whose careers were not contingent on developing and maintaining a publication record (e.g. undergraduate researchers, research assistants, and others who did not plan on having an academic career) were far less concerned with data's publication status and potential than were the other researchers. Further, they expressed little worry that their data's value would be used up by others. Renee (IM-GR), for example, was ambivalent about publishing her data. As a graduate student, she was more focused on gathering and analyzing data for her master's thesis. In contrast to other scientists, such as Elizabeth (NUS-PI), Ethan (ID&M-PI), and Matt (IM-PI), Renee did not expect peer-reviewed publications to be the outcome of her investment, and hence she did not conceive of her data's value through the lens of their publication status or potential. As Renee explained, "I

don't really feel like I have an academic reputation to uphold or build. The more people that can use my data the better." Ultimately, in Renee's view, there was no reason to protect her data, because there was little for her to gain by doing so. This counterexample highlights the importance of looking closely at the social context and associated goals that researchers are working toward as they collect, analyze, and manage data. Renee was conducting her research right alongside the other IM Team researchers; and, yet, her understanding of what was to be ultimately gained through the data she was collecting differed dramatically from the outcomes that the other researchers were working toward with the larger project.

So far, I have described conceptions of value that were based on the type of study data represented and data's publication status and potential. Next, I detail one final basis on which scientists conceived of data's value: data's processing state.

5.4.3 Data's Processing State

> *But obviously the raw data isn't really what we need at the end of the day. We need to get . . . There are a lot of mathematical manipulations that you have to do in order for us to get the answers that we're looking for. So after we come up with our raw data, then we're able to go into the little Excel sheets [. . .] and plug in data and eventually get out little charts and graphs and things like that that are better able to tell us what's going on in the streams.* (Carolyn, NUS-UR)

As suggested by Hilgartner and Brandt-Rauf's (1994) data stream model, data's level of processing was an important basis for scientists' assessments of data's value. Carolyn (NUS-UR) explains in the interview excerpt above that raw data were not what her team really needed in order to answer their research questions (i.e. "tell us what's going on in the streams"). Raw data sat at some distance from the products scientists most cared about: the measures, charts, and graphs that would more directly help scientists answer their research questions. Carolyn's depiction of data's value according to their processing state was common across the three teams

and was also apparent in my observations of and interviews with scientists. Other scientists I interviewed, for example, emphasized differences between raw and more processed data by referring to the latter as the "data we really care about" and describing the former as just devices—albeit crucial devices—for getting to the more results-like data that they would later present in publications.

At the same time, as first-level inscriptions of wetland field plots, artificial stream channels, and mesocosm tanks, scientists' rawest data were the representations closest to the actual objects scientists were studying. With the exception of modeling data, their production relied on a significant amount of manual labor, and they could not be reproduced without reengaging with the objects they represented. This stood in stark contrast to scientists' even minimally derived data, which could be fairly easily reproduced almost anywhere, using little more than the raw data and a laptop. The costs associated with producing raw data were highly salient to scientists so that even while the researchers devalued their raw data in comparison to the more refined products, it was important to them to avoid having to reproduce raw, level-one data.

In making value statements about data based on their processing state, scientists talked about data in dichotomous terms: for example, "raw" data versus "derived," "processed," or "calculated" data; "raw" data verses "charts" or "graphs"; or "raw" data versus "results," "answers," or "what we're really after at the end of the day." However, when I asked scientists to clarify what they were referring to in using these terms, it became apparent that their data actually fell along a spectrum; from not at all processed on one end to highly processed on the other. In between the two extremes of the rawest form of data (e.g. macroinvertebrate lengths measured by Renee (IM-GR)) and the most heavily processed form of data (e.g. a graph showing

the relationship between macroinvertebrate biomass and the number of *Typha* stems in a plot) there were data that represented various levels of processing. Scientists called data that were subjected to no processing "raw," but they also frequently used "raw" to refer to data that resulted from some calculation and/or transformation and that were then used as inputs to get even more refined data.

For example, *Typha* stem heights—upon which the IM Team relied—were understandably regarded by scientists as raw because no processing whatsoever was involved in their production; the researchers simply recorded the *Typha* stem heights in meters, to the nearest centimeter. But other datasets that scientists referred to as "raw" relied on transforming data that were yet rawer. A conversation I had with Matt (IM-PI) exemplifies this point. Early in my study, Matt (IM-PI) briefly mentioned a "raw pollen dataset" that he and Evelyn (IM-PI) produced in a previous project and had recently handed off for deposit in the Station repository. Eager to learn more about it, I asked Matt to describe the dataset in greater detail. He subsequently explained that the pollen data were not *completely* raw; at least not in the same way that his team's *Typha* stem height measurements were.

> [. . .] the pollen dataset . . . that I guess is all . . . Even what I was referring to as "raw" there's some . . . it's derived. The raw pollen data are just counts, but you basically [. . .] The dataset that we have kind of corrects for [. . .] proportion. [. . .] Just all the count data aren't really important for analyses, so I guess there is some level of work that's been done to some of these datasets before they go into that—what I was referring to as "raw" [. . .]

Similarly, NUS Team scientists referred to their measurements of the phosphorus concentrations of water samples taken from their artificial stream channels as "raw," even though the scientists did not directly measure phosphorus concentrations. Instead, they used a spectrophotometer machine to get a reading of the color absorbance of each water sample. Then, by plotting the relationship between known standards' phosphorus concentrations and their

absorbances, the researchers converted the color absorbance readings of their water samples into phosphorus concentrations. They would later use the phosphorus concentrations—along with several other numbers—to calculate an uptake rate for the nutrients in the streams. NUS Team researchers characterized both the absorbance readings *and* the phosphorus concentration numbers as "raw," though, as I have described, these two types of measurements were the result of different levels of processing.

There was no clearly demarcated line between the data that scientists called raw and data they called derived. However, I found that when using processing state as a basis for differentiating data's value, researchers tended to reserve the terms "derived" or "calculated" for data that were subjected to more extensive forms of processing, such as tests of statistical significance, graphing, regression analysis, statistical correlations, rate calculations (such as the rate of ammonium uptake in an experimental stream channel), and analysis of variance (ANOVA). Processing state based conceptions of data's value concerned the distance of data from the results or anticipated products of the teams' work. The further downstream the data were on the chain of inscriptions scientists produced (i.e. heavily processed data that scientists would include in a publication), the more likely scientists were to characterize them as the goal of their work and, as a result, to ascribe high value to them. Conversely, the closer upstream data were (i.e. raw, level-one data), the more likely scientists were to devalue the data and cast them almost entirely in terms of what measures and products the data would allow them to produce.

For example, as Jessica (NUS-PI) described her team's process for producing nutrient uptake measurements, she explained that the uptake rate of each nutrient (a measure of how many micrograms of a nutrient per meter square per minute were taken up in the channel) was what her team "was after."

> Jessica (NUS-PI): We'll take twenty samples through our curve and five background samples and we'll run them for our nutrients and our chloride, and we'll end up with one uptake number. And that's what we're after; what we publish.
>
> Interviewer: The one you're interested in.
>
> Jessica: Yeah.

Other NUS Team researchers echoed this sentiment, asserting that they were working toward producing "a point on a graph" (Elizabeth, NUS-PI) and likening the output of their calculations on raw data to "answers to questions" (Janet, NUS-UR) and "answers that we're looking for" (Carolyn, NUS-UR).

Renee (IM-GR) explained to me on more than one occasion that all the raw data she was collecting over the summer (e.g. macroinvertebrate lengths and plant coverage numbers) were "essentially so I can get two numbers: biomass and diversity." She went on,

> Those are the ones I really care about. Biomass is my main . . . that's what I'm really interested in, because biomass is important for these other organisms.

Biomass, unlike macroinvertebrate lengths, represented an element central to the questions in her study. Stated another way, macroinvertebrate biomass was a key variable in Renee's research, and biomass measurements would allow her to compare differences between treatment sites.

Likewise, ID&M Team researchers characterized their native plant species biomass measurements and nitrogen mineralization data as expected outcomes of their work, referring to them—in contrast to raw height measurements and plant stem counts—as "what we're really after" and "what we're really measuring." Even though many of the derived data that scientists described as "what they were really after" were not necessarily synonymous with results (for example, plant species biomass was not, in itself, an answer to any of the ID&M Team's research questions), they were relatively close to results, especially compared to raw plant stem counts and heights.

Scientists frequently told me that that, in contrast to derived data, their raw data were not what they actually cared about or were after in their studies. Rather, they explained, they collected the raw data in order to produce data that more closely resembled what they were after (i.e. results or answers to their questions). As Carolyn (NUS-UR) described in the opening passage, her team was ultimately interested in understanding how leaf litter affected nutrient uptake in streams, and the raw data were not what they "really need[ed] at the end of the day." Jessica (NUS-PI) described the relationship between raw and processed data in more detail.

> What we are interested in is measuring nutrient uptake. That's sort of the ultimate . . . To answer our question of how nutrients are related and what influences whether or not there's more uptake of nitrogen versus uptake of phosphorous, the thing we're measuring is nutrient uptake. And so the first thing we have to do is we collect a whole bunch of water samples that have different levels of the nutrient that we're looking at [. . .]. We have to analyze those to determine what the concentration is. [. . .] That's sort of our first step.

That first step—collecting water samples with different nutrients in them and analyzing them to determine their concentrations—resulted in data that the scientists would use to produce the data they *were* interested in; in this case, a measurement of nutrient uptake in the channels. As the researchers characterized it, they only cared about the concentration of, say, ammonium in a sample taken from one of their stream channels as an input for calculating ammonium uptake in the channels as a function of the leaves in the channels.

Matt (IM-PI) similarly described his team's raw field data in means-ends terms that indicated his team collected raw data, such as vegetation data, only *so that* they could produce the measures that were closer to products that answered their research questions.

> We collect vegetation data, which [. . .] includes species and dominance coverage values for each species. From those data, we calculate diversity indices and some other measures [like the] coefficient of conservatism. There's a bunch of other different kinds of plant values that we can calculate from the different species-specific data.

Kate (ID&M-PI) went so far as to describe her team's raw data as not what they were "really measuring." Referring to the plant stem height data that I helped record for each of the ID&M Team's vegetated mesocosms, Kate asserted, "Well, measuring those heights? You're not measuring that. [. . .] We want to translate that to biomass." In other words, while in fact the researchers were measuring the heights of plant stems in the planted mesocosms, those numbers did *not* represent the variable that they were really interested in capturing. Rather, raw plant stem height measurements were used (along with regression formulas) for extrapolating the numbers they were "really after": measures of plant biomass.

At the same time that scientists devalued raw data in comparison to data that were closer to the results they sought, it was clear that raw data played a crucial role in enabling scientists to produce more results-like data. In many cases, the variables that were the focus of researchers' studies could not be captured directly. For example, scientists in both the IM and ID&M teams were concerned with measuring plant species biomass. However, as scientists from both teams explained to me, they could not measure biomass directly without removing the plants, drying them, and measuring their weights; and thereby destroying the field study and mesocosm experiment. Instead, the researchers needed a means of extrapolating biomass measurements from data they *could* collect directly without destroying their studies. The raw plant stem counts and height measurements were such means. Likewise, NUS Team researchers explained that, while they would never publish raw data like the phosphorus concentrations of each sample, raw data were "necessary to get to the number that you would publish" (Jessica, NUS-PI).

Further adding complexity to how scientists conceived of data's value according to their processing state, scientists' rawest data were also difficult, if not impossible, to replace. Scientists collected raw, level-one data by interacting directly with the objects they were studying. For

instance, the ID&M Team's plant stem heights were first-order inscriptions that required the researchers to go to their mesocosms with a ruler and measure each stem within a sample area. After such data were captured, they could not be recaptured without engaging once again in what was already a costly and time-consuming process.

As an example of the amount of time and effort involved in collecting raw, level-one data, consider that the production of raw water sample data from four stream channels took NUS Team researchers an entire day that began at 8:30 a.m. and ended around 6:00 p.m. The process entailed five different steps just to collect and prepare samples, and this five-step process was carried out for each of three nutrients on each of four channels (Figure 5.1).

Step 1: Measure and record the rate at which water is flowing into channel 1
Step 2: Pour a known amount of a nutrient ("a slug") (either nitrogen, phosphorus, or ammonium) into channel 1
Step 3: Wait for a conduction meter to show the nutrient has begun to reach the end of the channel
Step 4: Begin taking samples (15 in total) of the water from the end of the stream, approximately every 40 seconds, filtering each one of debris
Step 5: Go inside (there was an indoor facility at the stream lab) and measure and record each sample's conductivity
Step 6: Repeat Steps 1-5 for the other two nutrients in channel 1
Step 7: Repeat steps 1-6 for the remaining three channels (2, 3, and 4)
Step 8: Take all samples back to the main Station lab and prepare them for analysis by creating the standards to use as a basis of comparison
Step 9: Analyze the ammonium and phosphorus samples and standards (the researchers planned to analyze the nitrogen samples when they returned home)

Figure 5.1: The steps involved in collecting raw data for the NUS Team.

After collecting and preparing the samples, the team then had to analyze each of the samples for nutrient concentrations. At 15 samples for each of three nutrients, in each of four stream channels, the team generated 180 samples *on a daily basis* that they then had to analyze, thus transforming them into raw data.

For the other two teams, collecting raw, level-one data was no less labor intensive or time consuming than it was for NUS Team researchers. To collect their rawest data, IM Team scientists had to start by loading up the necessary field equipment (e.g. coolers for collecting samples, length measurement tools, waders, shovels, sample plot markers, a GPS device, a soil corer) after which they would drive anywhere from 20 minutes to two hours just to get to the wetland site they were working in on a particular day. On arrival, the researchers unloaded all of the equipment from the vehicle and hiked with it—often wearing waders to keep water and leeches away from their bodies—to the individual plots at the site. Then, data collection could begin: a process that took five people approximately one hour to complete for each plot.

To collect raw, level-one data from their mesocosms (plant counts and heights), ID&M Team researchers spent hours hunched over tanks on bended knees counting all of the stems (hundreds) of each species in a sample area of each mesocosm and measuring their heights.[13] They divided this task between two people, with an additional person (me, on the days I was observing their work) recording the data as they called them out (Figure 5.2). Each tank took approximately one hour to complete, and gathering the raw plant data for an entire set of 24 vegetated mesocosms took our three-person team three days to finish.

Scientists frequently mentioned the large amount of labor and discomfort involved in collecting their raw, level-one data. Kate (ID&M-PI), for instance, described the raw data her team was collecting from the mesocosms as "a pain in the butt to get" and the process "just plain physically hard." She went on:

> It's just really bad on your back and it's hot and it's miserable, and you've got your head stuck in there, and you're drooling. You're not drooling, you're sniffling. It's just plain not fun.

[13] ID&M Team researchers measured the heights of all of two of the four species in the tanks. For the other two (whose stems were far more numerous), they measured a subset.

Figure 5.2: ID&M Team researchers count stems in one of the mesocosms.

Likewise, Renee (IM-GR) explained, "It takes so much energy to create those [raw] data"; Matt (IM-PI) characterized raw data collection as "physically difficult"; and Elizabeth (NUS-PI) asserted that raw data collection was "difficult because of the time involved."

Given the extensive amount of work involved in collecting level-one data, from a practical standpoint, they could not be easily replaced (if they could be replaced at all). For example, when NUS Team researchers realized that the raw data they had collected in the first few days of their experiment were not what they needed (they added too little phosphorus to the channel to detect a change in concentration in the samples), the team had to engage with the entire, daylong process of nutrient addition, sample collection, and sample analysis once again. This came at a high cost to the team, not only in terms of the labor involved in redoing the entire process, but also in terms of opportunity costs. An important dimension of this team's questions concerned change over time; NUS Team researchers wanted to see how nutrient uptake changed as the leaves in the channels decayed over the course of several weeks. As Elizabeth (NUS-PI)

described it, each day's loss of raw data meant "less points in [their] final graph," lessening the impact of the phenomenon they wanted to show.

If level-one data were difficult for NUS Team researchers to replace, they were virtually impossible for IM and ID&M Team researchers to replace. First of all, the summer field season left neither team's researchers time to analyze data as they went about collecting them. By the time errors would have been spotted—in the fall or winter—the researchers would be a minimum of several hundred miles away from the Station and their study sites. Additionally, the passage of time was a key factor for both the IM and the ID&M Teams: once fall began and vegetation started to die back for the winter, the researchers—even if they were still at the Station—would not have been able to capture the raw data again (unless, of course, they wanted to restart their entire study; a prohibitively costly option for both teams). The high amount of labor involved in collecting raw, level-one data and their practical irreproducibility were important to scientists. Even while scientists used data's processing state to explain raw data were of relatively lower value than their more heavily processed counterparts (e.g. "the raw data aren't what we're really after at the end of the day"), researchers regularly indicated that the costs associated with producing their rawest data were such that the researchers wanted to avoid having to replicate their production.

Conceptions of value based on data's processing state, then, were not as straightforward as notions of value that were based on the type of study data represented or on data's publication status and potential. Scientists regarded data at the rawer end of the inscription spectrum as less valuable (for showing them what was going on/answering questions) than data at the more heavily processed end, yet also made it clear that level-one inscriptions—as practically

irreproducible representations of what they were studying—demanded somewhat more care in ensuring they only had to produce them once.

In closing this section, I would like to point out that when scientists used data's processing state to talk about data's value, they rarely brought up data's potential value to other scientists. Some scientists did occasionally use processing state to differentiate between data they would and would not publish, explaining that they would not publish their raw data but that derived metrics would go into a paper. However, when asked which, if any, of their data might be valuable to other scientists, researchers tended to focus on the previous two dimensions discussed (type of study system and data publication status and potential) rather than data's processing state.

5.5 Summary of Findings

The theories of value that frame my study emphasize that the value of objects—in this case, science data—is inseparable from objects' meanings to those who interact with them (Beckert & Aspers, 2010; Blumer, 1969; Najder, 1975). My examination of scientists' conceptions of data's value, then, is necessarily tied to questions of the meaning(s) of data to scientists. My findings reveal several meanings of data that were salient to scientists as they considered whether data were good or bad, how long data would be valuable for, who should or could reap value from data, and the purposes to which data could be put.

The most prominent meaning of data to scientists was as resources for addressing a gap in knowledge, and in this role scientists emphasized that data needed to fulfill three main criteria to be of value to the team. Data had to be comparable, trustworthy, and relevant to scientists' specific research questions. Data that did not fulfill these prerequisites were of little value to the

teams I studied, with researchers characterizing such data as "crappy," "not worth the time," "unusable," or a "clusterfuck."

Other meanings for data, however, were also apparent, most notably in scientists' characterization of how long their data would be of value and who could or should reap value from their data. Scientists depicted data not only as resources for addressing a specific gap in knowledge, but also as representations of particular kinds of study systems, resources for producing publications, and inscriptions of objects that allowed them to create yet more inscriptions or were closer to the desired endpoint. Employing these meanings of data, the researchers thought the value of data from controlled studies was more limited than the value of data from field studies. While researchers could conceive of their field data as valuable to other scientists and across a long span of time, they thought data from controlled studies would likely not be of much value "past the point of what the people who were doing that experiment really want the data for" (Gabe, ID&M-PD).

As resources for generating publications, researchers made it clear that data's publication status and potential were important to their notions of data's value. Scientists described unpublished data as untapped resources for generating peer-reviewed publications (the most rewarded product of their work), and therefore of very high value to the team. That value, however, could only be realized with data's publication. Once data were published—assuming the team did not plan to generate yet more publications from them—scientists primarily regarded the publication as the valuable product and record of their work; not the data themselves.

Lastly, scientists made differentiations of their data's value on the basis of their meaning as inscriptions of the things they were studying. As inscriptions, scientists were particularly attuned to whether data were inputs—in which case their value was described in means-ends

terms—or outputs that represented what the scientists were "really after" in their work. In this way scientists devalued raw data in comparison to their more heavily processed counterparts. At the same time, however, first-level raw data—as the most fundamental surrogates of scientific objects—required significant costs to produce; costs that were large enough to make their reproduction difficult if not impossible to carry out. As a result, scientists' rawest data stood out to scientists as unique resources that enabled them to manipulate, combine, and transform what they were studying into something more closely resembling results.

5.6 Discussion

The meanings of data that scientists emphasized as they talked about their data's value reveal the uses for data that were salient to scientists across the data stream (both as they worked and as they were asked to consider making their data available to others); the timespans involved in value considerations; and the specific factors that contributed to scientists' conceptions of data's value. Not surprisingly, scientists were primarily concerned with data's value for their team's own, relatively narrow uses: addressing a gap in knowledge and producing the outputs that would garner them credit and prestige for successfully filling the knowledge gap (e.g. peer reviewed publications).

However, while scientists mostly worried about their own limited uses for data, they did not necessarily exclude uses beyond their teams in their considerations of data's value. Previous research has suggested that scientists feel more compelled to share and archive data if they think doing so would advance science (Borgman, 2010), be of use to others (Cragin, Palmer, Carlson, et al., 2010), or have broader public benefit (Niu & Hedstrom, 2007). While all three of these ends share in common the concern that data be of some practical use to others if scientists are to

take the time to share them, they still leave unclear what *specific* uses or *benefits* scientists have in mind. In elucidating the specific secondary uses that were prominent in scientists' conceptions of data's value, my study adds considerable detail and nuance to this other work.

When asked about their data's potential value beyond their studies, scientists regularly cited metaanalysis, cross-site comparison, and time-based studies as worthy secondary uses for data and assessed data's value according to how well they thought the data could serve those ends. In calls to make data more open and to enable data's preservation, researchers and government officials have similarly highlighted data's capacity to serve as inputs in comparative studies, metaanalyses, and studies that look at change over time (Heidorn, 2008; Lauriault et al., 2008; National Academy of Sciences, 2009). And in ecology—where scientists have long been concerned with understanding long-term changes to ecosystems and changes as the result of different human perturbations—such uses for properly managed and preserved data have been heavily promoted (Michener & Brunt, 2000; M. Palmer et al., 2005; Whitlock et al., 2010). My study provides empirical evidence that value propositions based on the integration and comparison of data from more than one study is compelling to scientists conducting ecological research; particularly when they are asked to consider their data's secondary value.

Interestingly, data integration and comparison were the *only* uses researchers highlighted when asked to consider their data's value beyond their own studies. Aside from a single offhand comment from Matt (IM-PI) that "some statistical wizard" might possibly be able to look at his team's mesocosm data in a different way, none of the researchers cited data's potential value for verifying results, replicating a study, or reanalyzing a study's findings. Funders and other data sharing proponents have regularly lauded well-managed and shared data's value for validating and replicating science (National Academy of Sciences, 2009; National Research Council, 2003;

Tenopir et al., 2011; Uhlir & Schröder, 2007; Whitlock, 2011), but in my study such uses did not figure into scientists' consideration of their data's secondary value.

There are two possible explanations for this finding. First, validation and replication could conceivably threaten the scientists who produced the data by revealing errors in their work. While none of the scientists mentioned such a threat (after all, they did not even bring up reanalysis and validation as potential secondary uses), other studies have revealed a high concern among scientists that their data sharing and/or archiving efforts bring them benefit or, at the very least, not threaten their own interests (Campbell et al., 2002; Louis et al., 2002; McCain, 1991). Another possible reason scientists did not bring up the value of data for validation and reanalysis could stem from the absence of such activities in the day-to-day practice of science as they knew it. Contrasted with metaanalyses or time and site based comparisons, scientists did not discuss secondary validation and reanalysis as important aspects of their disciplinary community's work. Though the scientific ideal stresses the importance of validation (via replication of a study, for example) and reanalysis (Shapin & Schaffer, 1985), these appeared to be relatively rare practices for the scientists I studied.

In addition to identifying the uses that were salient to scientists as they considered their data's value, my study also sheds light on the timespans involved in scientists' value conceptions. Previous research has tended to employ absolute timespans (e.g. 1-5 years, 1-12 months) in trying to understand how long scientists think their data will be of value (Beagrie et al., 2009; Cragin, Palmer, Carlson, et al., 2010). My study, however, suggests that when conceptualizing the length of time their data might be valuable, scientists think primarily in terms of events, rather than in numbers of years. The most salient timespan began with the collection of raw, level-one data and ended with data's publication in a peer-reviewed scholarly journal. The length

of time between these two events was variable among the teams I studied: NUS Team researchers anticipated generating articles within several weeks of completing their work at the Station, but longer-term projects like those of the IM and ID&M Teams might not yield publications for several months to a year or more. Regardless of the actual span of time, several scientists designated publication as the event past which data had relatively low value to the team. However, data's value was not *necessarily* fully exhausted for the team after the first publication. In particular, IM Team researchers described several datasets that they were still planning to generate new publications from. This suggests another important event in scientists' conceptions of data's value: the point at which the data are "published-out" for the team because the researchers do not anticipate being able to mine any new publications from the data. Furthermore, this finding challenges many mandates that stipulate that scientists should deposit their data within some set time after collection or publication. For example, the Station's mandate that researchers deposit their data within one year of completing their collection might not be meaningful for scientists who anticipate a fairly long period of time before they feel the data are published out and, hence, no longer need to be protected from others.

Beyond their own uses for the data they produced, researchers suggested that if data were at all valuable to scientists outside the team, the data would be valuable for a virtually unlimited time period. Field data—which scientists described as of potential use to others—were depicted as potentially having value for "decades," "a really long time," or even "until the end of time." Data from more controlled experiments, on the other hand, were described by scientists as having a much shorter timespan of value: one that ended after the data were used to answer the study's questions and were published.

Finally, my study highlights several factors that were key to scientists' value conceptions: the depletability of data's value for generating publishable findings; the ephemerality and artificiality of the study systems used to produce data; and data's status as inscriptions. As scientists talked about their data's value, they repeatedly brought up data's publication status, differentiating between "published" and "unpublished" data. This binary categorization, however, belied a more complicated notion of data's value that was based primarily on the perceived potential of the data to yield publications for the team. Scientists regarded data that were likely to result in new publications for their team as having very high value, but included in this category unpublished data as well as data the team had already published but that had not yet been exhausted of its publication value for the team.

Furthermore, when scientists described data's value from the vantage point of data's role in generating publications, they were emphasizing not only the importance of publications as products of their team's work (a finding that has shown up in virtually every data sharing study, e.g. Baker & Millerand, 2010; Blumenthal et al., 1997; Borgman et al., 2006; Cragin, Palmer, Carlson, et al., 2010; Tucker, 2009), but also data's depletability as a resource for generating publications. This runs counter to many arguments in favor of data sharing, which insist data are "endless fuel for creativity" (Interagency Working Group on Digital Data to the National Science and Technology Council, 2009, p. 1); and that data's "value is enhanced, not exhausted, by the first publication of conclusions drawn from them" (Whitlock, 2011, p. 62). Scientists' conceptions of data's value included the assumption that some of their data were indeed *limited* fuel for creativity, and in such instances, scientists planned to exhaust that fuel themselves. The scientists only viewed their data's value in the more expansive way suggested by data sharing

proponents when the data were the product of studies of systems like wetland field sites and real streams.

Just as some data preservationists have suggested data's long-term value depends on the type of study system that yielded the data (W. Anderson, 2004; Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010; Simberloff et al., 2005; Steering Committee for the Study on the Long-term Retention of Selected Scientific and Technical Records of the Federal Government National Research Council, 1995b), the scientists in my study also based their notions of data's value on whether data were from mesocosms, models, artificial stream channels, real streams, or wetland field sites. However, unlike the data preservationists, scientists did not tell me their experimental data were of limited value because they could be easily reproduced by rerunning the experiment; nor did they indicate that field data had longer-term, more expansive value solely based on the data's uniqueness. Rather, when scientists used study system type to explain differences in their data's value, they focused primarily on study systems' degree of naturalness, their likelihood of persistence, and their comparability to other systems. Data that were the product of more natural, less constructed-for-a-particular purpose, and more likely-to-be-around-for-the-foreseeable-future systems (e.g. data from wetland field sites) were seen as data that could be added to and/or compared to other similar data. These virtually inexhaustible resources stood in stark contrast to data from more artificial, deliberately constructed-just-to-answer-this-set-of questions systems that were likely to be dismantled or neglected to the point of uselessness. These data (e.g. data from artificial stream channels or data from model runs) were seen as inherently limited. This suggests an additional property of data streams that is important to how scientists view the value of their data and hence the data's

suitability for sharing: the potential for data to be expanded upon in order to answer new questions.

Lastly, scientists revealed that data's processing state was important to their notions of data's value. In doing so, they foregrounded both data's power as inscriptions and their role as either intermediaries or end goals. Latour's notion of data as inscriptions of laboratory objects (where the laboratory is "the place where scientists *work*" (Latour, 1987, p. 64)) emphasizes inscriptions' ability to allow people to "act at a distance" (p. 223). Unlike the objects they represent, inscriptions are mobile, stable, and combinable. As Latour explains, "When you hold a piece of information you have the *form* of something without the thing itself": the list of water nutrient concentrations without the artificial stream channels (NUS Team); the plant vegetation data without the actual wetland (IM Team); the biomass numbers without the mesocosms (ID&M Team) (p. 243, emphasis in original).

Whether data were first-order inscriptions (inscriptions of the laboratory object itself) or 2^{nd}, 3^{rd}, and n^{th}-order inscriptions (inscriptions of inscriptions) (Hilgartner & Brandt-Rauf, 1994) played into to scientists' value conceptions, because first order inscriptions could not be reproduced without significant effort, if they could be reproduced at all. Scientist' rawest data replaced laboratory objects that were messy, tied to specific locations, and—in many cases—changing such that the timing of the production of the inscription mattered. Contrast that with scientists' second-level (and third, and fourth-level inscriptions), which—because they relied on, say, vegetation data instead of plants in a wetland—could be regenerated fairly easily, almost anytime and anywhere.

The focus on processing state to talk about data's value also reflected the importance to scientists of whether data were intended as intermediaries/inputs or goals/outputs. When data

were viewed as intermediaries they were cast in language that indicated their value only extended to their use as a means of getting to something else that was more important: e.g. "not what we're after at the end of the day, but what we need to get the results." More heavily processed data, on the other hand, were characterized as outputs, answers, and results, and scientists described them in ways that reflected the view that such data had more value to the scientists than the rawer counterparts (e.g. "the answers we're looking for," "what I'm really after").

This chapter has shed light on the first part of my research question: how scientists conceive of their data's value. However, it still largely leaves open the question of how value conceptions are reflected in scientists' data practices. How, for example, do scientists' conceptions of value based on scientists' narrow, study-specific concerns play into how scientists produce valuable data? Further, how do assessments that data are not good for addressing the scientists' study problem affect what scientists do with their data? This, and other enactments of value conceptions, is the subject of the next chapter.

CHAPTER 6: ENACTING CONCEPTIONS OF DATA'S VALUE

6.1 OVERVIEW

In the last chapter, I described how scientists conceived of the value of their data. Now, I turn my dissertation's focus toward delineating how scientists enacted conceptions of data's value in their data practices. Specifically, I describe how scientists worked to create data that they would consider valuable and what scientists did in response to their conceptions of data's value.

"Data practices" is an umbrella term employed by science and technology studies scholars to encompass the "research processes and activities related to scientists' work with data" (Cragin, Palmer, & Chao, 2010). In studying data practices, scholars have looked at a number of different activities surrounding scientists' work with data, including data collection, calibration, documentation, and sharing and archiving. In my examination of the intersection between value conceptions and data practices, I found that scientists enacted their notions of data's value all throughout their work, including when they decided which data to collect and how to collect them; when they engaged in quality control activities related to data collection and determined that data were not good for the team's purposes; when they managed data during the active portion of their life cycle for the teams; and when they made decisions related to public deposit of data.

As scientists collected and worked with data, their primary concern was with data's value to their team and their study's specific questions and/or hypotheses. Scientists spent considerable effort producing "good" data, where goodness was always measured by data's fit to the study at

hand. Furthermore, they regularly assessed data's value for their study, seeking to identify bad data and minimize its effects. However, in contrast to the considerable nuance exhibited in scientists' conceptions of data's value, scientists' collection and management of data varied according to little beyond good versus bad distinctions in data's value. For example, I did not find evidence that scientists managed field data (which, when asked, they described as having potentially broader and longer-term value) differently than data from more controlled studies.

As they collected data and conducted their studies, scientists did not think about data's value beyond whether or not they were good as resources for addressing a gap in knowledge as they had defined it. However, when asked to make their data more openly available—e.g. by funders or the Station—researchers indicated that their decision to share was based strongly on data's value for producing publications for the team. Data that teams were still working with and planned to publish were regarded as too valuable to the team to make widely available, and scientists were unwilling to deposit such data in a public repository. Conversely, when scientists thought data's publication value had been fully exploited for the team, they saw little threat in sharing. In addition to publication potential, scientists also suggested that study type influenced their decision to share data and told me that they felt less compelled to share data from controlled studies because they assumed such data had inherently limited value.

I have divided this chapter's findings into three main sections. The first section lays out the data mandates that scientists were beholden to for their projects and describes the baseline set of data practices I observed in the three teams. The mandates and baseline practices are intended to serve as the backdrop to understanding activities that were based on scientists' value conceptions. In the second section, I focus on how scientists worked to create valuable data and

the steps they took to identify data that were not of value to the project. The last main findings section describes the actions scientists undertook in response to assessments of data's value.

6.2 The Landscape For Data Practices at the Station

6.2.1 Station and Funder Data Management Mandates

Because they were using Station resources to conduct their research, all of the teams in my study were subject to the Station's "data management policy." Established in 2010 and promoted through the Station's website, the policy ostensibly required scientists and teams of scientists to deposit their data in the Station repository within a year of completing their study's data collection activities. At the time of my research, the policy was being implemented with considerable flexibility. For example, although the Station built and maintained its own data repository, scientists had the option to deposit their data with an alternate public repository of their choosing so long as they provided the Station with a description of their study's methodology and a link to the data. Additionally, despite the official "requirement" to deposit data, Station staff indicated to me that they understood and accepted that scientists would be unwilling to deposit some of their data and had, thus far, not exerted much pressure on faculty researchers to abide by the data sharing requirement.

In the summer of 2012, the information manager was beginning to meet with faculty researchers for the first time since the Station implemented the data management policy to find out what data they would be willing to deposit. As the information manager described these meetings, they were primarily focused on learning what datasets scientists were developing and which, if any, of the datasets they were "comfortable" archiving.

> I've reached out to some PIs individually to set up meetings. Those meetings are more or less data interviews where I'm [. . .] trying to identify all the datasets that are in play that

> the person's developing and then hopefully identifying one or two datasets [. . .] which the researcher's comfortable working with me on in terms of immediately archiving or at least identifying a timeline which they're comfortable with in order to eventually get those data into the system. *(Information Manager)*

Importantly, aside from server space, the Station did *not* provide resources for managing data as scientists were collecting and working with data; and, in fact, the information manager said that he deliberately stayed away from intervening in scientists' "project-level" data management activities.

> I don't get involved in how individual projects manage their data. [. . .] the closest I get to that is meeting with some of the summer fellows[14] prior to their collection of data to try to give them some tips. That's mostly just . . . that's not really even project . . . that's a relatively tight focus, and that's mostly designed because I need to get data from them a good month later. So, no, I don't get too involved. Actually I try *not* to get involved in project-level data management. *(Information Manager)*

Instead, the information manager—and other senior Station staff—was primarily concerned with the "management" of data after researchers completed their projects. The information manager explained that the Station's interest in making sure scientists' data were archived and made available to others was motivated by three main goals: the Station wanted to help researchers adhere to data management mandates (e.g. NSF and publisher mandates); to create a record that demonstrated the research output enabled by the Station; and to build a repository that future students and researchers could use.

> [. . .] one of the primary points is that it acts to support future research and education, which are the primary missions of the Station. Availability of data describing the Station and its property and the systems around here will only help [. . .] future students and researchers. So I think that's a big part of it. The nice thing about [. . .] data management mandates and other potential or current data mandates is that it puts us in a position of providing a service to researchers. That's definitely another part of it. And that was one of the specifications of [. . .] our information management system and the policies: they need to be built in such a way that they can help researchers meet mandates. Those are two primary items. Then the other value comes in terms of supporting arguments for the

[14] The Station hosted several NSF-IGERT and REU (Research Experience for Undergraduates) student researchers every summer.

> Station itself in terms of showing . . . we can show all our publications and show how many publications per year; the amount of work from the Station and grants that go through the Station. We have a better way of documenting what they've accomplished, and I think that's useful as well. *(Information Manager)*

At the time of my study, there were 60 datasets in the Station's repository. The majority of these were captured by Station staff, such as the resident biologists, and were historical datasets (e.g. the nutrient profile of a nearby lake from 1913-1950) or datasets that characterized some aspect of the Station's facilities (e.g. GIS data depicting ecosystem type boundaries at the Station). None of the teams I studied was yet at the point in the projects they were carrying out during the summer of 2012 where they were being asked by the Station to deposit those projects' data. One of the teams, however, held data from prior projects they had carried out at the Station. Another team initiated a meeting with the information manager with the goal of learning early in their study what would be required to archive data later. As a result, the information manager's interaction with researchers about archiving their data and/or preparing their data for later archiving varied across the teams.

The IM Team had been carrying out research at the Station for several years, under a series of different grant-funded projects. By the summer of 2012, the team had amassed several datasets, and the information manager was working with Matt (IM-PI) and Evelyn (IM-PI) to get some of the data into the Station's repository. The information manager's meetings with Matt and Evelyn did not deal with the data the team was generating for the study they were still carrying out: wetland response to *Typha* removal. Renee (IM-GR)—who was considered part of the team, but was working on her own master's thesis project—was in her last year of data collection and did not plan to return to the Station the following year. She and the information manager met twice during the summer of 2012 to discuss depositing her data when she finished her thesis in December.

At the time of my study, the ID&M Team was only just beginning to collect data. Ethan (ID&M-PD) was aware of the Station's data management policy and initiated a meeting with the information manager to apprise him of the data his team was collecting, learn more about which of the team's data the Station would be interested in archiving, and ascertain what his team would need to do to prepare the data for deposit. Ethan said his primary motivation for instigating the meeting—even though his team had thus far collected few data and planned to continue collecting data for at least another year—was that he saw himself as the person on the team in the best position to provide data documentation. Because 2012 would likely be his last summer on the ID&M Team's project (he planned to apply for faculty positions in the fall), Ethan wanted to better understand what was required of his team while he still had an opportunity to be of assistance.

Notably, the ID&M Team's computational data was not a topic of the discussion between Ethan (ID&M) and the information manager as they discussed future archiving of the team's data. Since the modeling portion of the study—unlike the mesocosm portion—was not carried out using Station facilities, the Station's data management policy technically did not apply to the data generated from the model. In fact, Mark (ID&M-PI) and Gabe (ID&M-PD) spent very little time at the Station (only visiting occasionally to help with the mesocosms) and conducted the modeling work at the their university.

NUS Team researchers and the information manager had never met to discuss archiving their data at the Station. The information manager paid a visit to the team early in the summer—as he did with all the research teams—to learn more about their project, but he did not initiate a conversation with the team about their data or the Station's data management policy. This was not out of the ordinary, given the fact that NUS Team researchers had never before conducted

research at the Station (i.e. unlike the IM Team, they had no prior data that was under the purview of the Station's data management policy) nor had they reached a point in their project where the data were, according to the Station's policy, supposed to be publically deposited. When I asked Elizabeth (NUS-PI) and Jessica (NUS-PI) what, if anything, they had been told regarding the Station's data policy, they both indicated little awareness of it. Jessica said that she remembered that something about data archiving was brought up at the Station's winter orientation meeting, but she was unsure of the specifics. Elizabeth, who had not attended the winter meeting, was unfamiliar with any Station policies that applied to her team's data.

> Interviewer: What about instruction from the Station regarding your data? Have you received any instruction from the folks here about data archiving or data management?
>
> Elizabeth (NUS-PI): No. I didn't read it, or I didn't receive it. Either one of those two. I don't think . . . I don't know. It's a great question. I know there's something about photos. Can be any photo that the Station takes are their property or something like that, but data . . .
>
> Interviewer: You're kind of doing it on your own?
>
> Elizabeth: I kind of feel like we're kind of under the radar here. I mean there very well may be something. I should probably ask.
>
> Interviewer: I'm just trying to understand the [. . .] context of what's going here with regard to data management. That's why I asked. Like if any resources were provided or instruction about what should be done with data? But it doesn't sound like it.
>
> Elizabeth: No. It isn't, no.

As for funder mandates that applied to teams' data, only one team—the ID&M Team—was obligated to include a "data sharing plan" in the project's grant application. The ID&M Team's work was part of a larger, NASA-funded multi-institutional project. In the project proposal, the investigators included a one-paragraph data sharing plan that specified "data collected and products produced" would be made available through a "publically accessible project website" and/or through the Oak Ridge National Laboratory's Distributed Active Archive

Center (ORNL-DAAC), "where NASA-developed field data, remote sensing products, and models are held" (project grant application). Interestingly, however, when I asked ID&M Team scientists whether their funder required them to archive or otherwise make their data available, some said that NASA did not require them to do anything in particular with data.

> NSF . . . you have to have a data management plan, and in the data management plan—well, this is not NSF. This is NASA, so they actually . . . they don't. *(Kate, ID&M-PI)*

Other researchers indicated some familiarity with NASA data mandates, yet made no reference to the sharing and archiving activities the team had committed to in the project's grant application.

> For field data, the granting agencies . . . they typically expect you to have a data management plan, so you write that into the proposal. And then part of that data management plan should be to archive your field data and make that available to other researchers in the community and to other scientists funded by the agency. This project is funded by NASA. NASA would say, "Give us a data management plan that says how you're going to put all the fieldwork on a server that's through a website that's going to be available to other NASA researchers." They would expect you to write that plan in the proposal. The follow through on those is very spotty if it exists at all. So I don't know if they would follow up on whether you did it or not. *(Mark, ID&M-PI)*

> But the funding agencies . . . this is NASA-funded, and I'm not sure if they have a policy or not, but I feel like they probably do. *(Ethan, ID&M-PD)*

While all of the teams were subject to the Station's "data management" policy, those mandates were not an immediate concern for any of the data that researchers were producing at the time of my study. Station mandates were focused on getting data into the Station's repository—or some other public repository—within a year after teams were finished collecting them, and, as a result, seemed to play little role in scientists' practices during their studies. Of the three teams, only the ID&M Team's project was accountable to funder mandates; however, as with the Station's mandates, the data sharing plan the researchers included in the funding application was not salient to the scientists as they collected and worked with data during the

field season. In fact, none of the researchers referenced the data sharing plans that they and their partners had stipulated in the grant application.

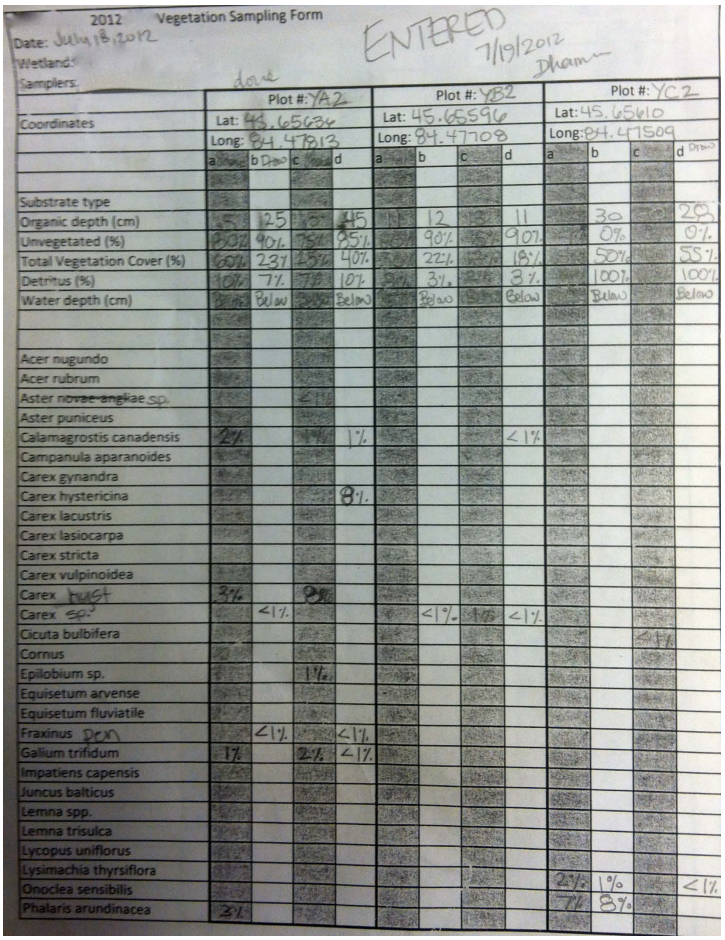
6.2.2 Baseline Data Management Tools and Practices

The basic artifacts and methods for data management were remarkably similar across the teams I studied and not appreciably unlike those described in a book published in 2000 on ecological data management (Brunt, 2000). Scientists recorded raw data onto paper templates; transferred the raw data to Excel spreadsheets, where they could sort, check, and analyze data; and stored their data in project-specific paper and digital files for indefinite within-team access. Before I delve into how scientists enacted value conceptions in their data practices, I describe the baseline data management practices I observed and that scientists articulated in my interviews with them.

Scientists in all three of the teams recorded the vast majority of their raw, level-one data by hand, using pencil and paper. With the exception of the ID&M Team's computational data—which were output digitally as they were generated—the measurement of variables and their inscription were two distinct activities that were often even carried out by different people on the team. For example, IM Team researchers counted *Typha* stems and measured their heights, calling the measurements out to another team member (usually Brooke (ID&M-UR), but also to me when I observed the team's work), who recorded the data on a paper sheet attached to a clipboard. Even when scientists used more sophisticated technical instruments—such as a dissolved oxygen (DO) meter or spectrophotometer machine—they recorded the data using pencil and paper. Such devices displayed data on a screen, but did not output data in any other manner. Recording data on paper templates was, as a result, the first step in capturing and

organizing data. I encountered two kinds of templates for recording raw data: "datasheets" and bound paper notebooks.

By far, datasheets were the most common artifact for collecting raw data. For example, the IM Team created a "vegetation sampling" datasheet for collecting data from their field plots (Figure 6.1), and a datasheet for raw data from their seedbank study.



2012 Vegetation Sampling Form

Date: July 18, 2012

Wetland:

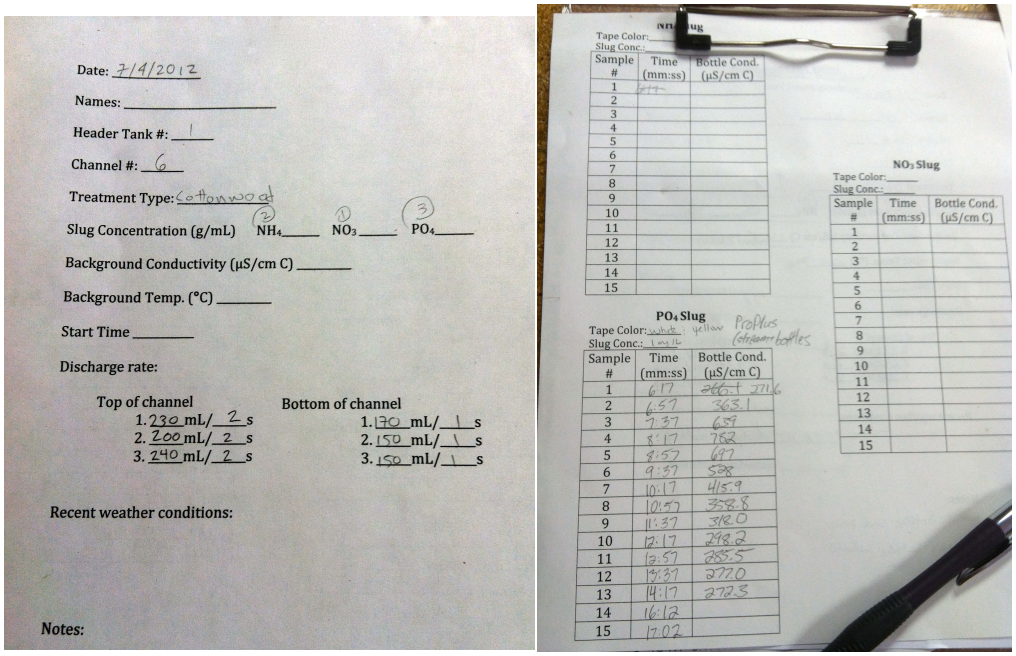
Samplers: dave

ENTERED 7/19/2012 Dhamm

Coordinates	Plot #: YA2 Lat: 45.65634 Long: 84.47813				Plot #: YB2 Lat: 45.65596 Long: 84.47708				Plot #: YC2 Lat: 45.65610 Long: 84.47509			
	a	b Dup	c	d	a	b	c	d	a	b	c	d Dup
Substrate type												
Organic depth (cm)	5	12.5	8	4.5	11	12	12	11		30		29
Unvegetated (%)	80%	90%	75%	95%		90%	85%	90%		0%		0%
Total Vegetation Cover (%)	60%	23%	25%	40%		22%	20%	18%		50%		55%
Detritus (%)	10%	7%	5%	10%	5%	3%	2%	8%		100%		100%
Water depth (cm)	2 cm	Below	3 cm	Below		Below		Below		Below		Below
Acer nugundo												
Acer rubrum												
Aster novae-angliae sp.			<1%									
Aster puniceus												
Calamagrostis canadensis	2%		1%	1%				<1%				
Campanula aparanoides												
Carex gynandra												
Carex hystericina				8%								
Carex lacustris												
Carex lasiocarpa												
Carex stricta												
Carex vulpinoidea												
Carex hyst	3%		8%									
Carex sp.		<1%				<1%	1%	<1%				
Cicuta bulbifera											4%	
Cornus												
Epilobium sp.			1%									
Equisetum arvense												
Equisetum fluviatile												
Fraxinus Dup		<1%		<1%								
Galium trifidum	1%		2%	<1%								
Impatiens capensis												
Juncus balticus												
Lemna spp.												
Lemna trisulca												
Lycopus uniflorus												
Lysimachia thyrsiflora												
Onoclea sensibilis									2%	1%		<1%
Phalaris arundinacea	3%								7%	8%		

Figure 6.1: The IM Team's vegetation sampling datasheet.

Renee (IM-GR) created her own datasheets for recording data from field plots and for raw macroinvertebrate data. The NUS Team made a datasheet to collect data as they were gathering their samples (Figure 6.2), and the ID&M Team made a datasheet for collecting raw plants data in their mesocosms (Figure 6.3).



Date: 7/4/2012

Names: _____

Header Tank #: 1

Channel #: 6

Treatment Type: Cottonwood

Slug Concentration (g/mL) NH4 (2) ____ NO3 (1) ____ PO4 (3) ____

Background Conductivity (µS/cm C) _____

Background Temp. (°C) _____

Start Time _____

Discharge rate:

Top of channel	Bottom of channel
1. 230 mL/ 2 s	1. 190 mL/ 1 s
2. 200 mL/ 2 s	2. 150 mL/ 1 s
3. 240 mL/ 2 s	3. 150 mL/ 1 s

Recent weather conditions:

Notes:

Tape Color: _____
Slug Conc.: _____

Sample #	Time (mm:ss)	Bottle Cond. (µS/cm C)
1	6:17	
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		

NO_3 Slug

Tape Color: _____
Slug Conc.: _____

Sample #	Time (mm:ss)	Bottle Cond. (µS/cm C)
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		

PO_4 Slug

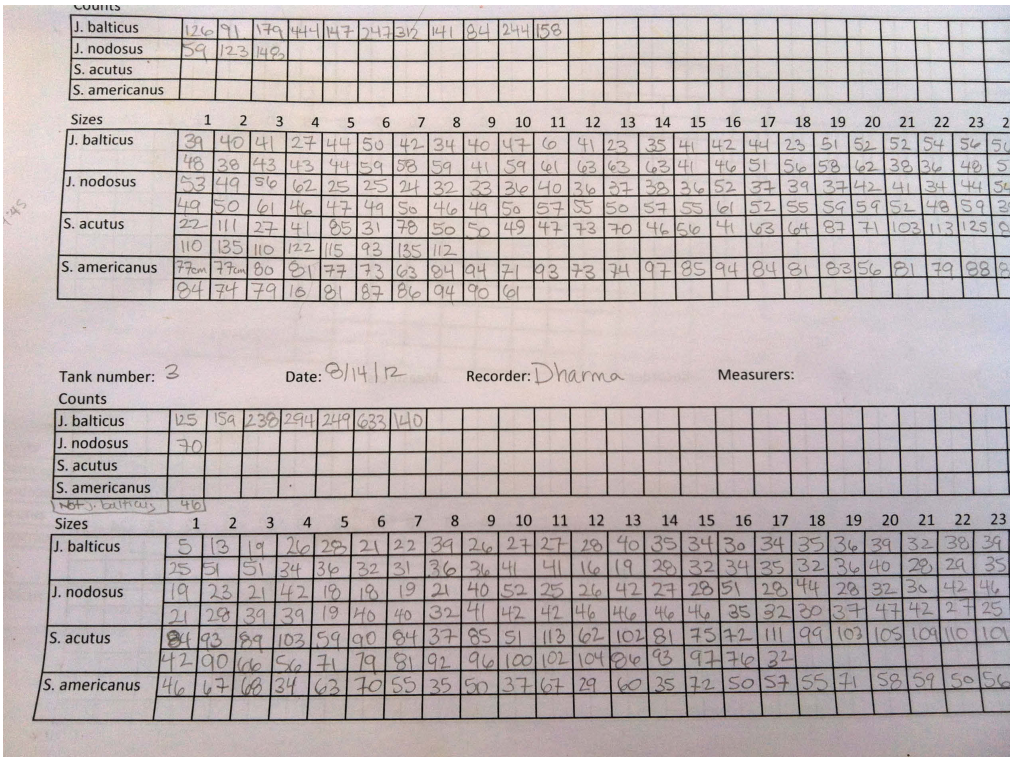
Tape Color: white yellow Proflux (stream bottles)
Slug Conc.: 1 mile

Sample #	Time (mm:ss)	Bottle Cond. (µS/cm C)
1	6:17	366.1 2716
2	6:57	363.1
3	7:37	659
4	8:17	782
5	8:57	697
6	9:37	508
7	10:17	415.9
8	10:57	358.8
9	11:37	318.0
10	12:17	298.2
11	12:57	285.5
12	13:37	277.0
13	14:17	272.3
14	16:12	
15	17:02	

Figure 6.2: Sides 1 & 2 of the NUS Team's water sampling datasheet.

The basic design of the datasheets was the same—no matter the team—and consisted of a printed-out piece of paper with labeled, empty slots for site-specific study information (e.g. plot and subplot numbers, channel number, or mesocosm tank number); the data collection date; and the variables scientists intended to measure (e.g. organic depth, bottle conductivity, or plant counts). Additionally, datasheets had a blank area for recording "notes," such as any problems that happened during data collection. As Elizabeth (NUS-PI) explained, "stuff always happens." Scientists printed several copies of the datasheets (usually on water-proof paper to ensure field conditions would not destroy the sheets), attached to them clipboards, and carried them into "the field"[15] when they were collecting data.

[15] Regardless of whether scientists were conducting their studies on mesocosms, artificial stream channels, or wetland plots, they often referred to data activities that took place away from an indoor laboratory as "in the field."



Counts

J. balticus	126	91	179	444	147	247	312	141	84	244	158													
J. nodosus	59	123	148																					
S. acutus																								
S. americanus																								

Sizes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	2
J. balticus	39	40	41	27	44	50	42	34	40	47	60	41	23	35	41	42	44	23	51	52	52	54	56	5
	48	38	43	43	44	59	58	59	41	59	61	63	63	63	41	46	51	56	58	62	38	36	48	5
J. nodosus	53	49	56	62	25	25	24	32	33	36	40	36	27	38	36	52	37	39	37	42	41	34	44	5
	49	50	61	46	47	49	50	46	49	50	57	55	50	57	55	61	52	55	59	59	52	48	59	3
S. acutus	22	111	27	41	85	31	78	50	50	49	47	73	70	46	56	41	63	64	87	71	103	113	125	9
	110	135	110	122	115	93	135	112																
S. americanus	77cm	77cm	80	81	77	73	63	84	94	71	93	73	74	97	85	94	84	81	83	56	81	79	88	8
	84	74	79	16	81	87	86	94	90	61														

Tank number: 3 Date: 8/14/12 Recorder: Dharma Measurers:

Counts

J. balticus	125	159	236	294	249	633	140
J. nodosus	70						
S. acutus							
S. americanus							
Not S. balticus	46						

Sizes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
J. balticus	5	13	19	26	28	21	22	39	26	27	27	28	40	35	34	30	34	35	36	39	32	38	39
	25	51	51	34	36	32	31	36	36	41	41	16	19	28	32	34	35	32	36	40	28	29	35
J. nodosus	19	23	21	42	19	18	19	21	40	52	25	26	42	27	28	51	28	44	28	32	30	42	46
	21	28	39	39	19	40	40	32	41	42	42	46	46	46	46	35	32	30	37	47	42	27	25
S. acutus	84	93	89	103	59	90	84	37	85	51	113	62	102	81	75	72	111	99	103	105	109	110	101
	42	90	106	56	71	79	81	92	96	100	102	104	86	93	97	76	32						
S. americanus	46	67	68	34	63	70	55	35	50	37	67	29	60	35	72	50	57	55	71	58	59	50	56

Figure 6.3: The ID&M Team's mesocosm plant datasheet.

In addition to datasheets, NUS Team researchers also relied on templates in each of two notebooks they used to record their samples analyses: one for ammonium, and the other for phosphorus. In the first page of each of the notebooks, Jessica (NUS-PI) wrote down the headings for all of the pieces of data that the undergraduate researchers should include as they analyzed the team's ammonium and phosphorus samples. After the first page of the notebook was filled in, the undergraduates used the previous day's page as a reference, copying the template to a new page and then filling it in as they analyzed the water samples.

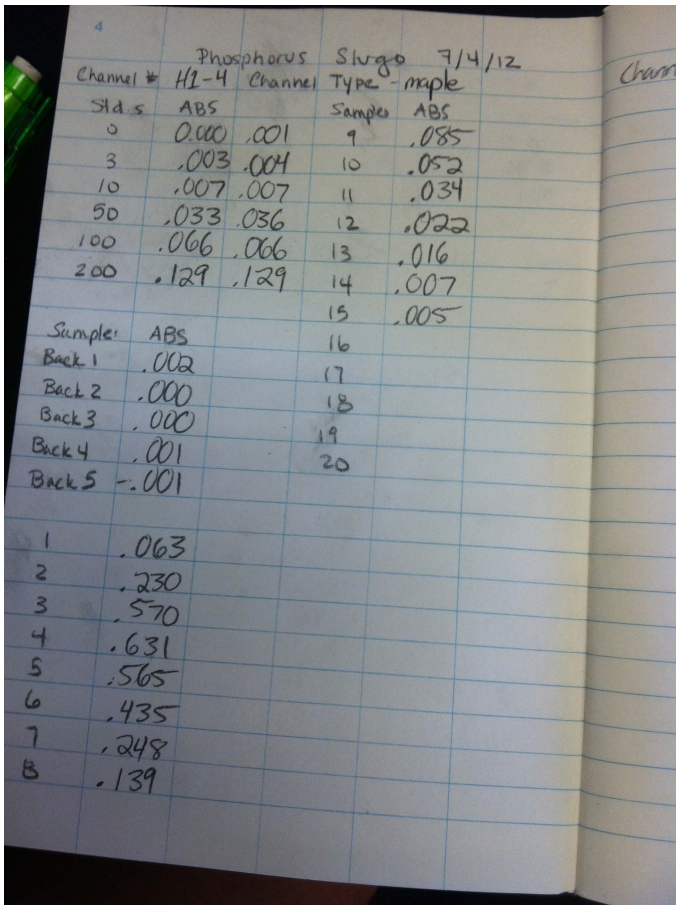


Figure 6.4: A page from the NUS Team's phosphorus notebook.

The paper templates—whether loose-leaf datasheets or pages in notebooks—served two main data management-oriented purposes for the scientists: they ensured that scientists recorded all of the pieces of data they needed for their study; and they organized the raw data to facilitate their understandability and later transfer to Excel files.

Several scientists told me that the datasheets helped their team make sure they captured all the data they needed to each time they collected data. Elizabeth (NUS-PI), for instance, noted that having the slots laid out served as a reminder to all of the team members of what they needed to record.

> Elizabeth (NUS-PI): [. . .] Laying out templates [. . .] so that people know how they need to be recording the data for the types of analyses that they're doing has also been really

> helpful, just in terms of consistency. A lot of that is just we have a big team and everybody does bits of everything, and so making sure that when you switch from one person doing something to another person doing something, it's obvious to them what they need to do based on just looking at a datasheet, even if they haven't explicitly been . . . you know, have been told how they need to set up their data.
>
> Interviewer: Because the slots are kind of there.
>
> Elizabeth: Yeah, the slots are kind of there. They can look in a notebook. They can look back at what somebody did last time, and they can see all the sort of metadata stuff that they need to write down at the top and they can see how to lay out their data as they start analyzing the samples.

Dylan (IM-RA) explained that datasheets made it possible for him and his colleagues to remember everything they needed to record.

> The projects are pretty involved, and it's impossible to really memorize all that stuff that we're taking, but as long as we have those datasheets as reference, we'll make sure that we don't leave without any data that we're looking for.

I witnessed the datasheets' memory-serving function when I assisted the IM Team as the "data monkey." The data monkey—a role often filled by Brooke (IM-UR)—was responsible for filling out the vegetation datasheet with the data that the other researchers on the team called out loud in the wetland study plots. Dividing the labor this way helped those responsible for measuring variables focus their attention on measurement; a task that almost always required the use of their hands (e.g. holding up a measuring stick to capture *Typha* stem height). When I recorded data, I noticed that I had to regularly remind the researchers that some piece of data was still needed when they thought they had completed data collection in a subplot. On one such occasion, Dylan (IM-RA) began to walk away from the subplot to help with plot maintenance activities. I quickly scanned the datasheet and noticed that a slot remained empty for "organic depth" in his subplot and reminded Dylan that he needed to measure it before we could move to the next plot. Overhearing this exchange, Matt (IM-PI) said, "That's what the data recorder is for;

to keep the caller honest." In a later interview, Brooke (IM-UR) explained that—as the team's data monkey—she was often responsible for making sure all the data were filled in for each plot.

> I'm always the one to be like, "Ok, now I need this, now I need this and this and this," and make sure that everything gets captured. Like yesterday when we were out, I was like, "Ok, Matt, you should give me the lat. long. coordinates, and Amy, I need substrate depth for plot B." So, yeah, I'd be the one to make sure everything was recorded at the site, because at the end of each plot they would say, "Ok, Brooke, do we have everything?" And I would go through the whole sheet to make sure that everything was filled out.

In addition to helping scientists remember what data to record, the datasheets and notebook templates also organized scientists' data to facilitate the data's later transfer to Excel spreadsheets. Without the datasheets, Jessica (NUS-PI) imagined her team would be "recording data on paper towels." The datasheets provided a structure for data, and in doing so made "it easier for data entry and data QA, quality control, types of things" (Elizabeth, NUS-PI).

Scientists collected raw, level-one data on paper, using templates that facilitated the consistent capture and organization of raw data. However, the paper form of data did not lend itself well to calculation and the generation of results. To create derived numbers and generate graphs and tables, scientists needed data to be digital. As a result, the second major data management activity for scientists in all the teams consisted of transferring the data from the paper datasheets and notebooks into Excel spreadsheets or workbooks (containing multiple spreadsheets). The more senior members of the teams (i.e. those who would be conducting data analysis later) created the spreadsheets, formatting them to facilitate data analysis. Because spreadsheets were meant to enable data analysis rather than raw data collection, spreadsheet templates were laid out differently from the paper datasheets.[16]

[16] Scientists also reported using statistical software to analyze their data, but none were involved in such activities at the time of my study.

> I mean, it's often the way it is . . . is that the most efficient way of taking data in the field is not going to be the most efficient way of putting it in for data analysis. *(Ethan, ID&M-PD)*

Often, the researchers formatted the cells in a spreadsheet to perform calculations that would yield derived numbers (e.g. uptake length or species richness) as the raw data were entered. Sometimes researchers waited until after the data were entered to run calculations.

In the NUS Team, the undergraduate researchers transferred data on a daily basis from the datasheets and notebooks into the team's preformatted Excel spreadsheets. All IM Team researchers involved in data collection transferred the data to spreadsheets as they had time. Matt (IM-PI) told me that while his team tried to get data entered into Excel "as soon as possible," a significant portion of the team's data entry did not occur until after the end of the field season. In the ID&M Team, Chad (ID&M-RA) was responsible for transferring his team's mesocosm data to Excel spreadsheets. He undertook this task as time allowed, but generally entered data into Excel within a week of collecting them.

Lastly, scientists described storing their team's data. Storage responsibilities typically fell with the PIs for the projects. PIs held onto paper datasheets indefinitely, keeping them in a file cabinet or file folders and saved digital data in a project-specific file folder on their computers. PIs also reported using some form of redundancy or backup protection, copying their digital files to thumb drives, external hard drives, departmental server space, and/or cloud-based data storage services such as Dropbox.

Now that I have described the mandates that were relevant to scientists' data practices and the basic tools and methods scientists used to collect and manage their data, I turn my attention to elucidating how scientists enacted their conceptions of data's value in their data practices.

6.3 Producing Good Data

6.3.1 Deciding What Data to Collect and How to Collect Them

Scientists' data collection activities were aimed at producing data that *they and their teams* could use to address *their* study questions. In scientists' view, there was no such thing as data that were good for every purpose or valuable in every context. As a result, data collection necessarily involved choices that shaped data into objects that were good for accomplishing some things and, conversely, likely not good for accomplishing others. Some of the earliest decisions related to data collection happened in the project planning stages, as scientists designed their studies and wrote their funding proposals. In the last chapter, I described how scientists' research questions shaped their high-level notions of what data ultimately needed to do in order to be valuable for the team, highlighting how they affected particular study design choices (e.g. controlled vs. more natural study systems; long- vs. short-term studies). Here, I turn my focus toward delineating how scientists enacted data's value as resources for addressing a gap in knowledge in the data they collected and the methods they employed for collecting data.

When they were available, scientists began with already established methods (within ecological science) for collecting data. Other times, scientists gathered data that had no established precedent, and they had to devise their own, unique approach to collecting data. In both cases, however, scientists paid close attention to the products they wanted to create, the phenomena they wanted to show, and the questions they ultimately wanted to answer, using methods aimed at producing the data that would be useful to their study. Considerations of data's value beyond the teams' studies had little impact on scientists' decisions about which data to collect and how to collect them. Even in instances where the researchers said their data could

have potential value beyond their studies, it was largely absent as a factor in scientists' data collection activities.

When tested data collection methods were available, scientists used them as a basis for gathering data in their own studies. Wetland vegetation sampling, nutrient measurement in streams, nitrogen mineralization studies, and plant abundance measurement were relatively common data collection efforts in ecological science, so researchers had some idea of how samples should be prepared for analysis, the specific variables that needed to be captured, and the tools or instruments that were appropriate for obtaining measurements. Even with recognized data collection protocols, however, the researchers considered the particular needs of their study as they adopted—and often adapted—methods for gathering data.

The IM Team's approach to vegetation sampling in wetland field plots, for instance, was based on a published method for vegetation sampling in the local region's wetlands (developed by fellow team member, Phil (IM-PI)). The published method indicated wetland vegetation community data collection should include percent coverage estimates of each plant species found in a sample quadrate; substrate type; organic depth; water depth; and water clarity. However, while this method served as a starting point for the IM Team's selection of variables to measure, the team made modifications to better address their study's questions, which revolved around wetland plant biodiversity response to different *Typha* removal techniques. Specifically, IM Team scientists discarded some variables—such as substrate type and water clarity—that they regarded as irrelevant to their study and added a few—such as detritus (i.e. *Typha* litter) coverage and litter canopy height—that they thought were important to their questions.

> We used his basic method essentially, with some modification, but his plot sampling methods, and we also adopted his datasheets. We modified slightly, because there are certain variables that we wanted to add, like litter depth, which is something he doesn't generally collect for standard vegetation monitoring that we thought was valuable for our

> study. We've kind of slightly modified the datasheet over time, but it's just more . . . maybe we threw out a couple of variables that he measured that we didn't need, and we added a couple of additional measures that we needed. *(Matt, IM-PI)*

Similarly, the NUS Team's basic method for collecting and analyzing water samples was "pretty standard" in stream ecology; however—as far as the PIs knew—no other researchers had looked at nutrient uptake in artificial stream channels, and there were peculiarities of the team's channels that led them to modify some of the methods. In fact, the researchers made several adjustments to their data collection methods and datasheets in the first couple weeks of their study as they better understood what specific things they needed to know in order to accurately calculate nutrient uptake in the channels. Carolyn (NUS-UR) described the changes her team made early in the project and the factors that guided those changes.

> Carolyn (NUS-UR): [. . .] in terms of knowing like what sorts of things we needed to collect, we definitely had to revise and change that. Like there has been a lot of talk along the way like, "Do we need to know just like the conductivity when we're collecting the actual samples, or should we know when the peak was occurring of the chloride?" I think a lot of it has to do with because we're not completely sure about how we're going to use the data once it's finished. [. . .] Like we went from just writing down when we were taking the samples to more like looking at the conductivity meter to help us guide how we were taking the samples and then even to start looking at things like, okay, we know we're collecting a sample like 6 minutes and 30 seconds and 7 minutes, but maybe we should be writing down the point at which we're reaching the highest conductivity. We have had to do a lot of revision based on not completely knowing what we've wanted, I guess.
>
> Interviewer: What's guided that revision? What kinds of things have happened that has made you guys go, "Oh, maybe we should be doing this?"
>
> Carolyn: I think some of that was as simple as we weren't getting the data we thought we were going to be getting so it was like, "Okay, we need to know more about what's going on, and we need to understand what's happening so maybe we need more information to do that." And I think some of it is also just like in order to maintain consistency we've realized that . . . like today for instance, we realized that throughout the course of the week we've been dumping in our slugs at different places. So like one day maybe it's 20 meters and one day maybe it's 19-1/2 meters. You know, a foot and a half doesn't seem like a big deal, but when it's in such a short stream it can make a difference.

These two examples highlight that even when established data collection methods existed within their field of study, scientists regarded them as subject to adjustment to better serve their project's specific needs. They added and discarded variables as appropriate and modified their approach to data collection to produce good data; data that were valuable for their studies. Many times, however, it was necessary for scientists to collect data on variables that were unique to their particular study and that had no established data collection protocols. I discovered early in my time at the Station that ecologists were a resourceful bunch; building their own study systems (e.g. stream channels made of linked-together plastic roof gutters); creating measurement devices (e.g. long plastic poles with meter markings); and constructing tools to facilitate sampling (e.g. quadrate markers made from PVC pipe). As Evelyn (IM-PI) said, "In ecology, there's a whole lot of . . . it's almost like creative . . . like you jerry-rig things all the time." Scientists also frequently improvised data collection methods. And—just as when they adapted established data collection methods to their own study—scientists paid careful attention to their purpose for collecting the data as they devised an original approach to capturing some variable.

Earlier, I described how IM Team researchers altered published wetland vegetation data collection methods to more appropriately meet the needs of their study. Two variables the scientists added were *Typha* litter height and percent coverage, because they felt *Typha* litter was an important factor in wetland native plant growth. Specifically, the researchers suspected that the litter negatively impacted native plant growth by creating too much shade (Figure 6.5). However, as far as they were aware, no other study collected data related to this factor; and they struggled to determine the best measure of what it was about the litter that mattered to native vegetation.

Figure 6.5: Sampling quadrate with Typha stems. The dry, brown stems were "litter."

Initially, the scientists measured the percentage of the quadrate that the *Typha* litter covered and the *Typha* litter's depth (the average height of the canopy of the standing litter). However, as the field season progressed, the researchers were less and less confident that these data captured the important aspects of the litter for their study. As Matt (IM-PI) explained, the *Typha* litter percent cover and height data were "not satisfying measurements," because they did not reliably get at the structure of the litter.

> Matt (IM-PI): I wish we had a better way of measuring this litter variable. [. . .] The litter height . . . the structure of this litter because we think it's important and we can see it and we can see how it affects the plant community underneath it, but we haven't been able to effectively measure it. That's a good example of something where it's like I wish we had a better way of [measuring].

Interviewer: You guys are doing percent cover right now?

Matt: Percent cover, and we're measuring depth.

[. . .]

Interviewer: So just measuring from the ground and seeing where the highest one—

Matt: Where the average . . . If there's a canopy, where's the average canopy. And it's just not a satisfying measurement because it's not taking into account the density and how much overlap there is. Because you can have five stems that sort of create a 60 centimeter high layer or you could have 50 stems that are creating a 60 centimeter high layer, and they're not the same; but a depth measurement is only going to get at the height.

Interviewer: Is it more about the shade it creates?

Matt: Yeah, that's why we're looking at light now.

The researchers began capturing another kind of data related to *Typha* litter: the light penetration underneath the litter, using a device that measured photosynthetically active radiation.

As IM Team researchers went about collecting their litter data, they were thinking forward to future data analysis activities. They were concerned that if they did not capture the right *Typha* litter data from their mesocosms, their analyses would yield very little in the way of meaningful associations with native plant growth. Projecting to a point later in the data stream, where scientists would produce derived data, was a common practice across the three teams. NUS Team PIs, for example, frequently talked about their data collection efforts in terms of the "ultimate graph" they would be able to generate with their data.

Interviewer: I notice you talk a lot [. . .] in terms of a "point on a graph." So are you saying that that's the thing that you're after, is the point on the graph?

Elizabeth (NUS-PI): I think so. I mean, when I visualize how a project is going to work, one of the things I like to do is draw what the graph might look just hypothetically. If I can see what the graph looks like, it means I know what's on my axis, and I know what data I need to collect if I can label an axis. [. . .] and there's obviously a hypothesis built into that graph too, so I feel like if I can draw a graph to begin with, then I can figure out what data I need to collect to make that kind of a graph. When I'm working through a project, [. . .] what I'm trying to do is figure out where these data that we're collecting

> right now are going to be on that graph. I want to see if there's a pattern and what we expect it to be or not.

Going into their study, NUS Team researchers had a strong sense of what their hypothetical, publication-worthy graph would look like: it ought to have "enough points to reasonably draw a line and say something about it" and to show a relationship between the N:P (nitrogen to phosphorus) uptake rate in a channel to the N:P of the leaf litter in the channel. This graph required not only several weeks' worth of uptake measurements ("enough points to say something"), but also data that showed leaf litter-driven nutrient uptake in the stream channels.

On that same theme, ID&M Team researchers thought ahead to the biomass data they needed to produce as they gathered raw vegetation data from their mesocosms. In ecology, biomass is often measured indirectly because direct measurement requires removing plants and drying and weighing them, thereby impacting the system scientists are studying. Instead, scientists take a small sample of plants of known size, dry and weigh them, and then create a regression formula that associates a size measurement with plant biomass. Plant stem heights are a common raw measurement used in this kind of extrapolation. However, the ID&M Team was faced with the challenge that two of the species in their mesocosms—*Juncus balticus* and *Juncus nodosus*—produce such a large number of stems (over 2,000 stems of just one species was common) that measuring the height of each stem would have taken a prohibitive amount of time. As a result, Ethan (ID&M-PD) sought a more efficient means of extrapolating biomass data for the two species, and identified a tight correlation—using data collected early in the field season—between the number of stems and the combined heights of the stems. In other words, Ethan devised a potential method of getting biomass data by counting stems and extrapolating the total height of the stems. One concern for the team, however, was that the correlation between stem counts and stem heights would not hold up as time went on and the plants were

exposed to the fertilization treatments that were part of the team's experiment. Fertilization could cause an increase in the stems' heights without increasing the stem numbers. If this were the case, the team's reliance on stem counts to calculate biomass would be faulty. To ensure Ethan's initial finding of a tight correlation between stem counts and stem heights held up, the researchers counted the stems of each of these two species and measured the heights of 50 stems in each mesocosm. At the time of my study, the researchers had not yet determined whether or not the relationship between stem counts and total stem height remained strongly correlated.

Scientists frequently balanced practical concerns related to time and money limitations with creating highly accurate data; and, again, the guideline for decisions was what researchers thought was actually necessary for their study. When a means of collecting data took a considerably greater amount of time than some other method, yet was not expected to have an appreciable effect on future data products or the conclusions drawn from them, researchers often chose less accurate methods. In Chapter 4, I described an example of such a situation in the IM Team's data valuation vignette. Renee (IM-GR) collected plant species name information for her macroinvertebrate study with less concern for accuracy than the rest of the team did for their study. For Renee's study, careful identification of plant species (i.e. clipping those plants she could not confidently identify in the field and examining them more meticulously at the lab) would have taken significantly more time, yet would have added very little to the quality of the analyses she planned to conduct.

Similarly, NUS Team scientists initially considered collecting 20 water samples from each channel, every time they added a nutrient slug to the channel. As Elizabeth (IM-PI) explained to me, using 20 samples could have led to more accurate derived data than the researchers would get from the 15 samples they actually collected. However, they concluded that

the derived numbers they were producing using 15 samples was sufficiently accurate for their study to justify forgoing the significantly more time-intensive option.

> Elizabeth (NUS-P): We could be more accurate in the derived data that we generate if we collected more samples. I guess we made a decision to collect fewer samples because it involves less time, recognizing that we could be more accurate if we collected more samples. But we're comfortable with the shapes of the curves that we're getting.
>
> Interviewer: You mean the 15 [samples] instead of 20?
>
> Elizabeth: Yeah, exactly. We're pretty comfortable with those numbers that we're getting out and the shape of those curves, so there isn't really a strong reason to collect more, and it would take a lot of time.

Scientists were intent on producing data that were good for their own studies, and—as I have just described—this often meant making choices that resulted in data that were valuable to the scientists' current study, with the associated implication that the data might not be useful to other types of inquiry. Interestingly, however, even for the kinds of data that scientists told me they thought could potentially be of value beyond their studies, longer-term considerations had relatively little bearing on data collection activities.

NUS Team researchers told me that they were not thinking of the potential broader value of their data as they collected them. As Tina (NUS-UR) said, "I guess I don't really think about the long-term. I just think about if it's good or bad [data]." This was not surprising, given that NUS Team scientists did not expect their data to have much value beyond their own narrow, controlled experiment. Elizabeth (NUS-PI) specifically attributed the absence of considerations of their data's broader value in the team's data collection activities to the controlled—rather than field—nature of the study.

> Interviewer: As you collect with, and work with, and manage your data [. . .] do you think about the long-term use of your data?
>
> Elizabeth (NUS-PI): [When we're doing a field study] I think about the larger framework for our data like in terms of the bigger picture questions and what's being worked on in

> this field, in general, and how our numbers and how what we're doing and our results would fit into that larger framework. I think—because we know we're using methods that are similar to what other people do—we know that our numbers are comparable to what other people are doing. We would always choose methods that would produce comparable results. But in this particular project, I am definitely less, you know . . . not thinking as much about [. . .] that sort of that relevance, as I might if we were out in the field.

Yet IM Team researchers, who were engaged in collecting field data that they said might have broader value, also indicated that the possibility of data's broader potential value had little influence on their data collection activities. For example, Amy (IM-PD) said some of her team's field data might be useful to other scientists studying wetland *Typha* invasion. However, she also said data's potential use to others was not especially salient while she gathered data.

> Interviewer: Are you thinking much about the long-term use of your data or potential long-term use of your data as you're out there collecting it or working with it?
>
> Amy (IM-PD): Not really, honestly. I can see how most of my experience has been shortsighted. It's just kind of thinking about publication and meeting those short-term goals [. . .]

In a separate interview Matt (IM-PI) agreed, asserting his team would consult standard methods texts as they planned and conducted data collection, but that facilitating broader or longer-term use of the data was not a consideration as the team collected data.

> I guess I would look—and there are some standard methods texts. As I would develop the methods for a project, I would look at the standard methods, and then—if they're available—and then record the data that they suggest. But as far as preservation goes, I think it's coming up with your project, and then a lot at the Station, but I guess I haven't thought really deeply about it.

Furthermore, as their modification of published methods demonstrates, IM Team researchers' primary criterion for selecting which data to collect and how to collect them was fit to the study at hand; just as it was for researchers on the other teams.

6.3.2 Making Sure Data Are Good

As I talked with scientists about the steps they took to create good data and observed them as they worked out and settled on data collection methods, it was clear that scientists did not take data's value for granted. In addition to employing data collection methods that they thought would meet their study's purpose, researchers also regularly checked the quality of the data they produced. In this section, I describe quality control measures that scientists implemented to ensure the data they collected—or were in the process of collecting—were what they would consider good.

In Chapter 5, I identified three qualities that were key to scientists' notions of what made data good. Data needed to be comparable to one another, trustworthy, and relevant to scientists' research questions. With these three prerequisites in mind, and an underlying assumption that data were, first and foremost, resources for addressing the gap in knowledge the researchers were at the Station to study, scientists periodically checked their data; both as they were collecting the data as well as after they had collected data and began compiling them for analyses.

In all the teams, researchers emphasized the importance of collecting data that were comparable across different researchers on the team as well as across data collection events. For example, all of the researchers on the IM Team should have generated close to the same plant coverage estimates for a given wetland subplot, and ID&M Team researchers should be able to duplicate their plant stem counts for any given mesocosm.

> [. . .] making sure that everybody is collecting the same data. That can be a challenge [. . .]. So trying to put in checks and balances to make that sure everyone's collecting the same kinds of data and it's repeatable . . . having clear standards and checking in with each other and making sure everybody's doing it right [. . .] *(Amy, IM-PD)*

> There are a lot of unsaid checkpoints in that process in which you're thinking, you're checking data, thinking about it. You're checking to make sure that the undergraduates[17] are counting right. Then you're checking the data to make sure [. . .] that you can replicate it. You're checking to make sure that all counts are right or within an area that you're expecting and then you can go out and recheck it, recount. *(Gabe, ID&M-PD)*

Often, it was enough for scientists to carefully follow the same procedure each time they prepared samples and collected data from them. The NUS Team's data were like this—as long as they were "really sure" that they did things the "same way every time," there was no need to check data for replicability. But I found that scientists viewed replicability as more difficult to achieve for some kinds of data, necessitating quality control checks during data collection.

The IM Team's plant cover estimates were one such kind of data. In contrast with the team's *Typha* stem height data, percent cover—as a visual estimate—was especially at risk of inconsistency or, as Evelyn (IM-PI) put it, "laden with variability." Matt (IM-PI) further explained, "Two people could have wildly different estimates of the same plot." To mitigate this variability, the IM Team regularly conducted what the scientists referred to as "calibration" of their plant percent cover estimates. Every 20th subplot (there were four subplots in every wetland field plot), the researchers gathered together around the quadrate. Each of the researchers examined the plants within the marker and mentally made note of the percent cover estimate they would assign to each plant. When everyone was done, the scientists shared their estimates with the rest of the team. If the estimates were close to one another, the researchers continued with data collection. When, however, the estimates differed significantly between the scientists, they spent several minutes discussing how they generated a particular estimate. The goal of this exchange was to "negotiate a good estimate everyone can agree with." Once an agreement was

[17] Occasionally, the ID&M Team hired one of the many undergraduates taking courses at the Station for an afternoon to help count and measure plant stems.

reached, the team proceeded with their regular data collection activities until the next 20th subplot.

ID&M Team researchers carried out similar replicability-oriented checks as they collected plant stem count data from their mesocosms, though not with the same formal regularity as the IM Team. In each mesocosm, ID&M Team researchers placed a sampling ring in the center of the tank and counted the stems of each of four native species (Figure 6.6). While certainly not a complicated data gathering activity—as one researcher said, "everyone can count"—in the densely vegetated tanks, it was easy to miss stems, lose track of which stems had already been counted, and forget what number one was on.

Figure 6.6: One of the ID&M Team's mesocosms with a sampling ring in the center of the tank.

To ensure stem count data were repeatable, the researchers occasionally recounted the stems within a tank. I observed Chad (ID&M-RA) carry out this recounting activity on one of the days I assisted his team with data collection. Upon completing his stem counts in the sixth tank of the day (all of which I recorded by pencil into a datasheet), Chad immediately counted the stems of one of the species again. After learning that his second count was within five stems of the first count (330 stems versus 325), his check was finished, and he felt secure that his stem counts were replicable.

In addition to the quality control activities that scientists carried out as they collected raw data, scientists engaged in quality control of their data *after* they collected and transferred them to formats more amenable to data analysis (e.g. Excel spreadsheets). Quality control of data after they were collected was focused on identifying "strange outliers" and numbers that "did not make sense" either in the context of the team's other data or in comparison to scientists' knowledge of ecosystems and the phenomena they were studying. Scientists considered post-data collection quality control an important step in making sure data were comparable, trustworthy, and applicable to the team's research questions.

While ID&M Team researchers were at least a couple of months from analyzing their first year's data, Ethan (ID&M-PD) had begun to carry out "exploratory analysis" on his team's raw data to make sure they were what he referred to as "reasonable."

> We have data from the summer already from a month ago that I've done some initial analyses on. And I like to do that. When you first get the data collected, get it entered, clean the data up, look for problems in the data, so that you don't come back four years later and wonder, "What the heck happened there?" I like to do that—within a month or two—at least take a cursory look at the data and make sure if things look like they're reasonable.

Large outliers or data that did not match scientists' own knowledge about how ecological processes worked were two key challenges to reasonableness; and potential indications that the data, or some portion of the data, were faulty.

> [. . .] just some basic things that I do is like, I'll just plot the data out and look for an outlier—a big outlier, like ten times bigger—then I know a decimal point got moved or something like that. In doing initial analyses if you do sort of a regression, you look at the errors and see, "Oh this data point has a really large error. It's not what you would expect." And then go back and look and make sure there are no notes written down about, "Oh, this tank was trampled in." Yeah, just sort of exploratory data analysis, looking for anything that looks odd. *(Ethan, ID&M-PD)*

As Ethan indicates in the excerpt, anomalies could be explained by something as simple as a misplaced decimal point that was the result of a transcription error or by a disturbance in one of the mesocosms. Kate (ID&M-PI) described the search for outliers or numbers that "did not make sense" as the search to make sure what the data showed was "for real" and not an "artifact" of data collection or processing.

> [. . .] that sort of awareness of data and making sure things that . . . being really skeptical. I guess, when I think about data, I'm a real skeptic. That means I'm constantly looking at things and saying, "Well, let's make sure that's showing us that for real, and it's not some artifact. [. . .] I guess what I'm coming around to is—for me—one of the biggest issues [. . .] is making sure that you're not seeing artifacts. You're not seeing anomalies that are a result of the way you treated data or a result of one or two bad data points that are driving a whole lot of differences or the manipulation of the data you did that just ended up with something really odd. And that happens amazingly often. It requires constant vigilance as you look at analyses along the way of making sure something makes sense.

As Kate indicated, anomalies in data were not a rare occurrence, nor were they necessarily a threat to the entire study. The real threat was to miss anomalies.

> [. . .] it's important to notice it. *That's* the most important part. It's not that big a deal when it happens. The big deal happens when it's not seen, when it just goes through and you never know it. *(Gabe, ID&M-PD)*

Just as with the ID&M Team, IM Team scientists talked about the importance of examining data for anomalies. Evelyn (IM-PI) said her team regularly checked for outliers as

they compiled and began to analyze data. While strange anomalies could reflect a real natural variability in the ecosystems the team was studying—and therefore possibly indicate that the study did not include enough replicates—they could also reflect data collection problems or a faulty mesocosm. For example, Evelyn talked about "one funky" mesocosm that leaked and was overgrown with terrestrial ferns; this mesocosm often gave the team erratic numbers. Knowing that data were collected under problematic conditions—and noted as such on the paper datasheets—helped scientists to feel assured that the data were not showing something real about what they were studying.

> What we do is when we're out in the field and we're taking measurements, if something is hard to measure or we're not confident in the way that we sampled, we'll write a note in the side. We'll say, "In plot 4D, there was like, whatever, some big problem there and we couldn't get the sample. Or there wasn't enough water, or it was flooded." Or whatever . . . whatever the issue is. Because then when you plot all your data, and that one becomes a huge outlier, you remember, "Oh yeah, that was the one that we had all that trouble in." *(Evelyn, IM-PI)*

Outliers and anomalies were data that did not conform to the rest of a dataset, and, as a result, scientists regarded them as suspect. But researchers were also concerned with data that fell outside the bounds of what they would expect based on other published research or their own previous work. As Gabe (ID&M-PD) explained,

> [. . .] we check it with the literature, and we make sure that this makes sense and it's within the bounds of logic that we'd been expecting; so many milligrams per kilogram or something.

While two of the three teams I studied were not yet far enough along in their research projects to conduct anything but the most preliminary of data analyses, the NUS Team offered the opportunity to observe how scientists applied "fit to expectations" to assess the quality of their data. The NUS Team was unique among the teams I studied for the researchers' frequent analysis of data as they carried out their stream experiments. As I noted earlier, the field season

generally marks an intensive period of data collection for ecologists (at least those I met at the Station), with relatively little time available for data analysis. NUS Team scientists, however, analyzed their data regularly throughout the field season.

NUS Team researchers had carried out similar studies in real streams, but had never run an experimental study in artificial stream channels. Furthermore, they were aware of no other studies that had used artificial stream channels to look at nutrient uptake driven by leaf litter. As a result, they were unsure whether their study would work and did not want to wait until "December when we're trying to write this up for a NABS [North American Benthological Society] abstract" to find out "that it's failed" (Jessica, NUS-PI).

> We're analyzing the data sort of as we go. That's why we have Wednesdays and Saturdays to . . . in part, because we needed to make sure it worked. *(Jessica, NUS-PI)*

> I think it's better to [analyze the samples] as you go through the process. Because also, we could go . . . everything could completely not work, and we wouldn't find out until December. I just couldn't handle that. *(Elizabeth, NUS-PI)*

In checking to "see if [the] systems are working or not," the scientists were looking primarily for a match between the results they expected—based on their previous work in real streams—and their actual results. Specifically, the researchers expected their data to show nutrient uptake in the channels that had leaf litter and to show no nutrient uptake in the channels without leaves (the control channels). Furthermore, they expected nutrient uptake to differ based on the type of leaf litter that was in a channel (i.e. cottonwood, maple, and a mix of cottonwood and maple). To check that their study was working, NUS Team researchers transferred the nutrient concentration data that they recorded on paper to Excel worksheets that were preformatted with formulas for calculating nutrient uptake. The PIs then compared their study's nutrient uptake results with an approximation of what they should see, given the parameters of

the experiment. When those numbers differed significantly, it indicated to the scientists that something was faulty in their experimental setup and, therefore, needed to be addressed. The data they were gathering were not good, and their justification for conducting the project threatened.

Thus far, I have described scientists' attentiveness to the specific needs of their study as they collected and examined data and tried to identify bad data. In the next section, I describe what scientists did in response to their assessments of data's value.

6.4 Responding to Assessments of Data's Value

6.4.1 Dealing with Bad Data

In spite of the activities scientists engaged in specifically to produce data that would be valuable to their teams, sometimes scientists were faced with bad data: data that—in their estimation—were of little use for their study and the questions that underlay their study. Sometimes, bad data consisted of one or two points among a set of hundreds of pieces of data. Scientists could identify their cause as a problem with data collection confined to one plot or one mesocosm, and the data's exclusion from analyses was both justified and fairly inconsequential. Other times, bad data were a more substantive threat to scientists' study, requiring adjustments to a study system's setup and/or scientists' data collection activities. In these cases, scientists were unable to use a significant portion of their data, potentially threatening their ability to address the questions they set out to answer. NUS Team scientists, for example, found themselves unable to generate good data, despite altering their experimental setup. They ultimately decided it was more worthwhile to end their study than to continue producing data that could not be used to test the hypotheses that framed their work.

Scientists told me that when they confronted a few bad data points among many and they could identify the source of the problem (e.g. the data were from a mesocosm that consistently yielded weird data), they "threw out" the problematic data points. Referring to the erratic data her team often found themselves with from one peculiar mesocosm, Evelyn (IM-PI) explained the justification for excluding some data from analyses.

> Sometimes we'll just throw the data out if it's really whacked. Because we have reason to, you know. It's not like . . . You can't just throw data out if it's an outlier because that could be the difference between here and here, and that's a natural variability. But when you know that [. . .] what you're sampling isn't real because of some interference or some bad measurement, that's when you can justify, "Let's take it out, because I think that was the influence of all those ferns" or whatever.

Importantly, in the instance Evelyn describes, the researchers surmised—because the data came from a disturbed mesocosm—that the outliers were not "real." Furthermore, the outliers were not so numerous that they threatened the IM Team's ability to use the data to generate results. As a result, it was unnecessary for the researchers to adjust their study or data collection methods to improve data going forward. It was enough to throw those data points out, where "throwing out" meant excluding the data from the team's analyses.

However, I encountered two instances in which bad data were enough of a threat to scientists' studies that they felt it necessary to take steps aimed at improving the data they were collecting: Renee's production of incomparable data in her first year of data collection (in 2011); and the NUS Team's production of data that indicated a faulty experimental setup. In both cases, the scientists altered their studies, either changing their methods of sampling and data collection or the study systems themselves. Additionally, the researchers excluded data as resources that could be used for their studies, quarantining them to separate files where they would not threaten their good data.

As I described earlier in the dissertation, Renee (IM-GR) was challenged in the first year of her master's study to produce data that were comparable to each other. The main problem was that many of the plots Renee surveyed had such low water levels that—in the drier plots—she was unable to collect data on several important variables and to employ her standard sample collection protocol. For example, in the dry wetland plots, Renee could not collect dissolved oxygen data, because the DO meter needed at least two inches of standing water to take a reading. Additionally, Renee could not use her pumping method to obtain macroinvertebrate samples from the drier plots, because there was not enough water to pump. Instead, she had to take a soil core, which meant that she could not be sure if results from her data indicated differences in her sampling method or real differences in the macroinvertebrates in the plots.

At the end of that first summer, Renee (IM-GR) was left with data that both she and Evelyn (IM-PI) referred to as "apples and oranges," a phrase they used to emphasize the data's lack of comparability and, hence, uselessness for Renee's study. Because the data were "unusable," the first major action Renee took was to exclude them from her thesis analyses.

> Renee (IM-GR): [. . .] So we ended up not using a bunch of data just because we couldn't compare them.
>
> Interviewer: [. . .] so what's going to happen with that?
>
> Renee: We're just pretty much, for the most part, using the samples that were collected with the stovepipe.
>
> Interviewer: Like you said, monitoring what was going on there instead of doing a comparison?
>
> Renee: [. . .] the soil core was mostly used in these wet meadow areas so basically when I wrote—I've written this chapter, the last chapter of my thesis—and basically the way I did it was I said, "I collected all this data in the wet meadows, but there wasn't enough water so I had to use soil cores." Basically I'm just not going to use that data. I *just* used the data from the emergent marsh. So I said, "In an emergent marsh, how does *Typha* affect the invertebrate community?" is all I did. It was just a big hassle. And you know it

> was my first year, and I was new so . . . And it was all stressful because I'm like, "I can't use a quarter of my data."

In other words, Renee modified her research questions (i.e. by adding "in an emergent marsh [. . .]") to fit the data she gathered in the wet-enough plots. Importantly, Renee did not dispose of the bad data, even though she regarded them as unusable: "[. . .] last summer is not worth just tossing them. You don't ever want to just toss data." She held on to the datasheets that contained the raw data and kept the Excel files that contained a copy of the raw data. For Renee, omitting the data by not including them in her analyses was sufficient for negating the impact of bad data.

In addition to not using the data from the drier plots she confronted in that first summer, Renee (IM-GR) also took steps to make sure that she would not face the same issues in her second (and last) summer. Specifically, she changed her study design to look at only plots that were on the outer edge of her team's wetland sites. "Edge plots" were very likely to have high enough water levels to allow Renee to gather all the necessary data and use the same sampling methods across plots.

> Renee (IM-GR): The whole idea of setting up *this* summer was to say, "Okay I need to . . . " I knew what environmental variables I needed to relate these biomass data back to. Kind of choosing the sites and choosing the plots that I was going to use, all went into I need to be able measure DO and I need to be able to measure water depth. That's why I'm only sampling the edge plots. That's also why I'm not sampling at [Wetland Name].
>
> Interviewer: So dry.
>
> Renee: Exactly. It's kind of a sticky subject too. Because you can't just be like, "Well, I'm just not going to sample here because it doesn't really fit into the way I want to sample." So it's kind of tricky.
>
> Interviewer: But, like you said, it's important to get consistent data that you can actually compare that to.
>
> Renee: Right. That was the goal for this summer. After the data kerfuffle from over the winter, I was like, "I can only sample in these plots that have water in them." That kind of limited me to these two sites and the 12 plots instead of 24.

Like Renee's first year thesis work, the NUS Team also had trouble generating what they would have considered good data. The main challenge for NUS Team researchers, however, was not the collection of comparable data. Rather, they had trouble producing data that were relevant to the team's research questions. As described in the last section, NUS Team researchers analyzed their data regularly as they were gathering them to make sure that their experimental setup was working appropriately. In the first couple weeks of their study, the scientists were frustrated to find that nutrient uptake had taken place in all of their channels, including the channels without leaves (the controls). Additionally, the team's results showed a much higher level of nutrient uptake in the channels populated with leaves than the scientists' back-of-the-envelope estimations indicated they should have. To the scientists, these two findings indicated a basic fault in the study setup, which was designed to isolate nutrient uptake as a function of leaf litter. The finding that nutrient uptake occurred in the control channels and at a much higher rate than expected in the other channels meant that something besides (or in addition to) the leaves was taking up the nutrients. The PIs repeatedly emphasized that they were not interested in studying nutrient uptake as a function of any other factor aside from leaf litter.

> [. . .] that's not our question. We don't really . . . That's not what we're interested in. That wasn't the whole point of why we built all these experimental channels; to grow algae and fine particulate organic matter of unknown C to P to N ratios. *(Jessica, NUS-PI)*

As a result, if the researchers were to continue their study, it was critical they be able to exclude whatever factor, or set of factors, was confounding results by taking up nutrients.

It occurred to the scientists that the gutter materials themselves affected nutrient uptake; or that the stream water intake system brought in a significant amount of organic material (such as algae and other fine particulate matter) that took up nutrients. There was little the researchers could do about the gutter materials once their study was underway, because constructing the

channels had taken them approximately two weeks and used up a significant portion of their project's funding. As for the possibility that other organic materials were in the channels, the team had already placed filters made of nylon hosiery on the water output valves that emptied into the channels; and the researchers spent time every evening manually brushing off algae from the channels. The only reasonable option left to the team was to change the water source from stream- to groundwater; and this is what they did approximately halfway through their study.

To the disappointment of NUS Team researchers, the switch to groundwater did not improve the team's results. They still found that nutrients were taken up in their control channels and that the amount of nutrient uptake was much larger than it should have been. As the PIs saw it, the data were unusable because they did not show nutrient uptake as a function only of leaf litter. Further demonstrating the importance of the relevance of data to scientists' particular research questions, the researchers concluded that the data were of such little value that they did not justify the team's completion of the study. With three weeks left, NUS Team researchers disassembled their channels and quit the study to pursue what they characterized as more worthwhile endeavors.

> We didn't do anything wrong. It's not like we messed up. And we're figuring this out early enough that we have the option to salvage the rest of our summer to do something else rather than realize in December when we're trying to write this up for a NABS abstract that it's failed. So in some ways everything is good. We did it in a way that allowed us to make sure we were answering the question we wanted to ask, not "Did it work?" And we're not interested in how much nutrients that fine particulate organic matter takes up. *(Jessica, NUS-PI)*

Like Renee (IM-GR), NUS Team scientists saw little value in data that could not be used to address their study's questions. Early in the study, when the researchers discovered something was wrong with their data, but thought they might still salvage the study, the PIs told me that

they would move the data to a different file so that the bad data would not "mess up" the team's good data.

> I think that once we learn that an experiment hasn't worked and we know why—we figured out why it didn't work—we pretty much discard that data. We might keep those files electronically, but we won't go back and revisit them, and, inevitably, my guess is they might get moved to a separate—all the stuff that didn't work that we've actually entered electronically—might get moved to a separate file so it doesn't mess up our pretty datasheets.[18] We won't get rid of the hard copies of any of that information, but we won't go back and look at it. *(Elizabeth, NUS-PI)*

"Discarding," as Elizabeth described it, did not mean permanently deleting data; rather what her team did with bad data would more appropriately be called "delete by omission." The team would still have access to the data—both the paper form and the digital—should they want to access them, but they would be sequestered from other, more valuable data. Furthermore, Elizabeth doubted they would be looked at again.

When NUS Team researchers decided later that their experiment was a failed one, the PIs sometimes facetiously said they would trash the data. But in actuality, they told me, they did not anticipate destroying the data: not the paper datasheets nor the Excel spreadsheets they had transferred their raw data to. If nothing else, the data might be useful for future study designs.

6.4.2 Enacting Notions of Data's Value Beyond Good Versus Bad

Among the teams in my study, scientists did not, for the most part, treat data differently according value conceptions that extended beyond whether data were good or bad. In fact, aside from when scientists contemplated making their data available through a publically accessible repository, I encountered just two examples of data practices that varied as the result of more nuanced value conceptions; both were in the ID&M Team.

[18] Elizabeth was referring to her team's Excel spreadsheets when she used the word "datasheets" here.

ID&M Team researchers engaged in particular data documentation and/or saving practices depending on whether their data fit into the category of "things we're really interested in" or "data we're just collecting to set up our study and/or account for potential anomalies later" (i.e. "peripheral," "ancillary," or "testing" data). For the mesocosm portion of the team's study, the former category included the team's native plant biomass and soil mineralization data. Sometime after the data were entered into an Excel workbook, the researchers created a "metadata sheet" to go along with the data. Take, for example, the team's "Soils Mineralization" workbook, which contained the team's soil mineralization data. In addition to separate sheets within the workbook for soil weights, raw data extracts of the nitrogen content of all the soil samples, and the calculated nitrification rates, the workbook included a spreadsheet titled "metadata." The metadata worksheet contained a one-sentence description of the dataset (e.g. "net nitrogen mineralization rates"); a list of the field methods used to collect the data; and a "file description" that enumerated and explained each spreadsheet in the workbook and all the variables contained in each sheet.

In sharp contrast, the scientists did not create metadata sheets for data that fell into the category of what both Ethan (ID&M-PD) and Kate (ID&M-PI) termed "peripheral" or "ancillary" data, such as the chemistry of the water going into the team's mesocosms at each of the two sites; and a set of data for looking at the flow patterns of water through the mesocosm tanks. When I noticed and asked Ethan about this difference in his team's documentation practices, he explained that he and the other researchers on the team were "not really interested in the [peripheral] data." Rather, they were just collecting these data to set up their experiment.

Not only did the researchers not document their ancillary data, but—as I detailed in the ID&M Team's data valuation vignette—they did not anticipate archiving such data at the

completion of the project. Until the information manager suggested otherwise to Ethan (ID&M-PD), the researchers did not see how their ancillary data would be useful to any purpose beyond setting up their study and accounting for anomalies they might find later. In fact, Kate (ID&M-PI) asserted that the data would be useless to anyone else.

> That's something that we might use to illustrate or to demonstrate that our treatments work in a way we'd like them to work, but they're not data that anybody else could ever use. *(Kate, ID&M-PI)*

Mark (ID&M-PI) and Gabe (ID&M-PD) also described data storage practices that differed according to whether their computational data were testing data or data from "real model runs." Model development—which was the stage of modeling work the researchers were engaged in at the time of my study—for the ID&M Team involved an iterative cycle of testing and modification. During testing and modification of the Wetland Ecosystem Model, Mark and Gabe ran the model, generating what they called "testing data." They examined the data to make sure the model was "behaving;" if it was not, they sought to understand why and made modifications accordingly.

> [. . .] what we're looking for is the model "behaving badly." And so things like that would be locking up, would be the population declining for no real reason . . . for the model acting in a way that is not only unexpected but even illogical and usually has a basis in something we're aware of, and sometimes it's not. *(Gabe, ID&M-PD)*

According to Mark (ID&M-PI) and Gabe (ID&M-PD), data from real modeling runs—runs they had not yet conducted, but would conduct later to answer research questions—were redundantly saved on a desktop computer and an external hard drive, or what Gabe referred to as the "long-term data files." However, as Gabe explained, he and Mark would not take the same steps with testing data. Gabe might save the test run data on his computer, but—unlike data from real model runs—he would not store data on any other device.

> Gabe (ID&M-PD): What Mark does is that, all this beta testing stuff he does . . . it's mine, which means it's not going for any long-term stuff.
>
> Interviewer: What do you mean? Sorry I'm not—
>
> Gabe: He's not . . . he does not really care. [. . .] it's not like he doesn't care about the data. We go through the data, [. . .] but he doesn't catalog it. If Mark's not going to put it in some long-term database, it's not going to get saved. I save it; it's on my computer.

From the point of view of ID&M Team scientists, what the ancillary data and testing data shared in common was that they were not collected to answer the team's research questions. Rather, such data were produced to set up the experiment, help the team interpret future data, or test to see that the model was functioning properly. In some ways, treating data differently according to whether they were for answering questions or were more peripheral was related to the precondition that data be relevant to research questions in order to be considered good by scientists. However, the scientists did not consider the ancillary or testing data to be bad or useless, so much as they thought of them as tangential to the study's research questions and the interests of the scientists and, therefore, less in need of documentation and data backup or archiving measures.

Despite the relative lack of influence that notions of data's value—aside from whether data were good or bad—apparently had on most areas of scientists' data practices, when researchers were asked to make data more widely available, they relied heavily on conceptions of value rooted in data's publication potential for the team. When researchers surmised that data had high publication value for the team, they felt that making them available outside of the team would squander the data's value. As a result, they were unwilling to deposit such data in public repositories such as the Station's. When, however, scientists thought that data's publication value had ended for the team, they saw little threat in making data widely accessible and were much more willing to publically deposit data.

Of the teams I studied, only IM Team researchers had deposited any of their data with the Station. Early in the 2012 field season, Matt (IM-PI) handed off of two datasets to the information manager: one was a GIS dataset that mapped *Typha* stand age in a nearby wetland and the other a pollen dataset from the same wetland that showed the relative dominance of various plants (including *Typha*) through time. IM Team researchers used both of the datasets together in a published journal article that showed an association between pollen dominance and the spatial dominance of *Typha*. Matt said the decision to "hand the data over" to the Station was motivated primarily by the fact that the data were published and his team did not anticipate making any more "real use of them."

> [The information manager] was just trying to identify some datasets we felt comfortable sharing, essentially. Then [. . .] we decided on one particular dataset that we'd actually already published.
>
> [. . .]
>
> It was published, and we didn't feel like there was any other real use for it. I mean, there's probably some more information you could get out of it, but, for the most part, we did what we thought we could do with the data. *(Matt, IM-PI)*

As I pointed out in the last chapter, scientists often implied that the important publication-related distinction regarding data's value was between published and unpublished data. Further observations and statements, however, indicated that what was really at issue in scientists' conceptions of data's value was data's publication potential for the team. For example, IM Team researchers told me that they held other published datasets that they were unwilling to deposit at the Station at this time because they anticipated generating more publications with them.

> Some of our *other* datasets that we *didn't* want to share . . . maybe we're still working on them or they're not clean or there's more . . . we might be able to use them for some other purpose. *(Matt, IM-PI)*

One such dataset was comprised of approximately ten years' worth of data from the team's mesocosm study. Unlike the GIS and pollen datasets that the IM team passed along to the information manager for deposit, the researchers expected to write more papers using the mesocosm data. Matt explained, "We're still writing papers with the data, and it's our data."

While the other teams had not yet been asked to make their data available through the Station's repository, the scientists similarly emphasized that they would not share data until after they were done authoring publications with them.

> We haven't published it yet so it's not available for anyone else. Because we need to get it published first. *(Jessica, NUS-PI)*

> We will have a lag because, again, especially Ethan and Gabe need to be getting papers out from it. *(Kate, ID&M-PI)*

> The first few publications have to be my publications. If I gave a copy of this to someone and they said, "This is really great," and did a whole bunch of model runs and published something with it, and that came out *before* my publications came out, that's not okay. *(Mark, ID&M-PI)*

One exception that provided an interesting contrast to the focus on publication potential as an important basis of data sharing decisions was Renee's (IM-GR) willingness to share her thesis data almost as soon as she finished collecting and organizing them. Renee met with the information manager twice during the summer to discuss depositing her data at the Station and explained to me that she was willing to share her data more quickly than the other researchers because she was not concerned with publishing the data.

> Renee (IM-GR): [The information manager] will eventually have all my data.
>
> Interviewer: He will?
>
> Renee: Yeah, because pretty much any research that's conducted out of the Station, he gets the data for. Usually, it's after people publish it, just because people are protective over their data. I really don't care that much [. . .]

Unlike the other scientists I talked to, the potential publication value of data was of minor concern to Renee. As a result, it did not factor into her decision to share the data.

In addition to data's publication value to the team as a basis for sharing decisions, scientists also indicated that study system type factored into their sharing behavior. For example, NUS Team researchers told me that they were not inclined not initiate public deposit of their current study's data (with the Station or any other data repository) primarily because the data were from a tightly controlled experiment and were therefore expected only to be useful for answering the study's questions. The scientists emphasized repeatedly that they could not imagine that such data would be of much use to others.

> Interviewer: Based on what you've said so far in this interview, would it be accurate to say that you don't plan to share or archive these data in any way aside from writing a paper?
>
> Elizabeth (NUS-PI): I think I would never delete all these files. I would never . . . but in terms of really long-term—like post my death type of archiving—I currently don't plan to do anything.
>
> Interviewer: Like deposit them in some kind of disciplinary or repository or—
>
> Elizabeth: No.
>
> Interviewer: Okay. And you don't plan to deposit them with the Station?
>
> Elizabeth: We haven't been asked to do anything like that. If they asked, I would say, "Sure, you're welcome to take this," [. . .]. I probably would not reach out to them about this kind of data, again, because it's an experiment in these channels as opposed to observations of the natural system, which I might be more inclined then to say, "Would you like some component of this data?" because it would contribute to baseline information or something.

Scientists on the other two teams expressed similar uncertainty about depositing data from controlled studies. Matt (IM-PI), for instance, said that he was not sure that his team would deposit their mesocosm data with the Station, even after the data's publication value had been

exhausted for the team. Just as with NUS Team researchers, Matt was doubtful that data from a more controlled system would be useful to others.

> Matt (IM-PI): I guess, eventually, we will likely share those data. But since it's a controlled study. I don't know. It doesn't seem . . . I don't know. I don't know how useful it will be for them to have those data, but I think eventually, once we've kind of worked with the mesocosm data a bit more, then we might give them the system too.
>
> Interviewer: When you say you don't know how useful it would be, are you saying that you don't—
>
> Matt: Well, I guess I don't really understand what the goals of the data project are. It seems to make a lot of sense that data that are collected about the ecosystems on the Station . . . these long-term changes in ecological conditions on the site; that seems like that could be really valuable for the Station. But these mesocosms are this sort of ephemeral thing. We put them in, they're changing really rapidly . . . most people take them out. It's not like it's looking at the changes in the forest over time.

Kate (ID&M-PI) told me her team would "probably" deposit their mesocosm data with the Station after they were finished generating publications from them, and Ethan's (ID&M-PD) meetings with the information manager suggested the same. However, in explaining why she had not yet made mesocosm data from a study she conducted approximately a decade ago (not at the Station), Kate said, "I think there's less interest in mesocosm greenhouse experiment data being out there than field data."

Unlike data's publication potential, study system type did not determine scientists' willingness to share data. While scientists expressed unwillingness to share data that still had high publication value for the team, they were not necessarily averse to share data from more controlled studies. In fact, scientists from all three teams said if they were required to share data and the data were published out for the team, they would share them, no matter if the data were from field or controlled studies. At the same time, however, researchers also made it clear that public deposit of data was somewhat less compelling when—as with controlled experiment data—data were regarded as having little potential value to others owing to the perception that

they could not be extended or expanded upon with additional variables or comparison with other datasets.

6.5 Summary of Findings

In this chapter, I have described how scientists enacted conceptions of data's value in their data practices. I found that as scientists collected and managed data during their projects, they were primarily concerned with creating good data and identifying and reducing—or at least minimizing the effects of—bad data. Scientists' work involved a number of choices—particularly about what to collect and how to collect it—that made data into objects that the scientists thought of as good for some things and not good for others. As scientists worked to create good data and checked to ensure the data they were producing would be valuable to their own studies, they regularly considered the products they wanted to create, the phenomena they wanted to show, and the questions they wanted to answer. In cases where scientists judged data to be useless for the particular purposes of the team, they worked to remove the data from their analyses and improve their data going forward. When, as with the NUS Team, researchers determined their study would not yield data the team could use, they discontinued their projects to work on "more worthwhile" activities.

There was little evidence that scientists thought much beyond good and bad distinctions in the value of their data until they were asked to make data publically available. Then, scientists were primarily concerned with whether data had publication value to the team. When scientists anticipated generating publications from data (even if they had already published the data once), they were unwilling to potentially squander that value by making them widely available. Conversely, scientists thought of data that they and their teams were done "squeezing" for

publication as having low publication value to their teams; and they therefore saw little threat in sharing them. Several scientists also suggested that the type of study data were from influenced their decision to deposit data in a public repository. Scientists conceived of data from more controlled studies as having little value beyond the team or the questions that motivated their production, and, therefore, while they were not unwilling to share such data, they were not especially compelled to share them either.

6.6 Discussion

My research helps to clarify how scientists manage data in a local context as they produce and use the data in their own work. Echoing the finding—presented in Chapter 5—that scientists were mostly focused on data's value for their team's own fairly narrow uses, scientists' data practices were heavily geared toward producing data that could be used to answer their study's research questions and/or test their study's hypotheses. Scientists continually enacted data's value for addressing a particular gap in knowledge by selecting and altering data collection methods, conducting quality control on data, quarantining data, and adjusting their studies. Furthermore, as scientists decided on the variables they would measure and their methods of data collection, they made it clear that the production of data that was good for their studies necessarily circumscribed the potential uses of those data. When, for example, Renee (IM-GR) decided that her plant data did not have to be highly accurate, she understood those data as good for her study, but not good for the team's larger project. In other words, despite the normative claim that data are inherently valuable, scientists' data practices are consciously aimed at producing valuable data given the context of the study at hand.

In keeping with results from other data practices research (Birnholtz & Bietz, 2003; Borgman et al., 2006; Borgman, Wallis, & Enyedy, 2007; Louis et al., 2002; Tucker, 2009), my study provides more evidence of the important role of publication in scientists' decision to make data widely available. However, by focusing on scientists' enactment of value conceptions in their data practices, my research also shows that publication *status* is not necessarily the important determinate of scientists' willingness to share. The researchers I studied were principally concerned with whether data had publication value to the team, regardless of whether the data were published yet or not. When researchers thought their data still held the promise of helping the team to generate new publications, they protected their team's ability to exhaust that value (by withholding the data) before making the data available to others.

Publication value was paramount to scientists' willingness to share data. However, researchers also indicated their perception of data's usefulness to others influenced their data sharing behavior. Like the social scientists Hedstrom and Niu surveyed (2008), the scientists I studied were less compelled to share data that they regarded as probably not useful to others. Extending the work of Cragin et al. (2010) and Hedstrom and Niu, my study reveals specific data qualities that informed scientists' judgment of data's secondary value. Specifically, scientists assumed that the usefulness of their data to others was predicated on the ability to compare, extend, or add on to the data. As a result, scientists across all three teams thought there was less reason to share experimental data—like those from mesocosm studies and artificial stream experiments—than there was to share field data, because experimental data had finite value that the scientists planned to extract themselves. Researchers similarly sought to exhaust field data's publication value to the team before making them more widely available, but they could imagine additional uses for field data that went beyond their team's interests. These findings help explain

the seemingly conflicting findings that scientists are less likely to share data of high value (Tucker, 2009), yet more likely to share data that they anticipated would see high use (Hedstrom & Niu, 2008). The scientists I studied did not want to engage in activities—like data sharing—that threatened their own realization of data's value. And at the same time, they did not see the rationale for sharing data that were likely to have their value tapped out by the time the scientists were willing to share them.

Notably, scientists demonstrated very little nuance in data practices among these different types of data during the early part of their work; a particularly striking finding given the distinctions in value scientists described when I prompted them to talk about data's value. In other words, while scientists exhibited subtlety in how they thought about data's value, such subtleties were not important to them as they produced and worked with data (at least during the field season). There are a couple of possible reasons for this disconnect. First, even while scientists viewed some kinds of data as having potential value that extended beyond the project occupying their attention, they may not have had the knowledge or expertise to translate these value conceptions into particular practices. Researchers did what they needed to do to produce and use data for the purposes at hand, and they devised systems that worked well for their near-term needs.

A second explanation for the valuation-data practices disconnect—somewhat related to the first—is that scientists were quite simply not compelled to consider their data's broader value as they collected and worked with data. The Station's information manager deliberately stayed out of scientists' "project-level" data management activities, waiting until projects were nearing completion (or after) to discuss the deposit of project data into the repository. Furthermore, mandates—whether Station-level mandates presented at orientation meetings and on websites; or

funder mandates that required researchers to state sharing plans in their funding applications—did not seem to make much of an impression on activities early in data's life cycle. Decisions made early in data's life-cycle have significant impact on data's viability beyond individual research projects (Cedars Project Team, 2001; National Academy of Sciences, 2009). My results indicate that, while compelling secondary value propositions (Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010) existed for scientists, the value propositions had not been leveraged as well as they might have been to have the greatest impact on future access.

CHAPTER 7: CONCLUSION

In this concluding chapter, I summarize the findings and contributions of my study, outline some of the implications of my findings, and suggest areas of future work.

7.1 SUMMARY OF THE STUDY

This dissertation examined data practices through the lens of scientists' conceptions of data's value. My study was based on a multi-case examination of three small teams of scientists who carried out ecological research at a U.S.-based university field station. Using interviews and participant observations, I answered the following research question: How do scientists conceive of the value of their data, and how do scientists enact conceptions of value in their data practices? The theoretical basis for my research was constructivist, drawing on the data stream model as articulated by Hilgartner and Brandt-Rauf (1994) for studying data access issues as well as on theories of value. As I described in Chapter 1, the data stream model suggested factors that might enter into scientists' assessments of data's value. Specifically, Hilgartner and Brandt-Rauf highlighted what they called "several important properties" of data streams: data streams are comprised of a heterogeneous collection of elements (e.g. equipment, samples, software, techniques, inputs, and outputs); the elements differ not only in their form, but also in their rarity and level of reliability or factual status; and data streams are composed of chains of products or inscriptions that vary according to processing state. Hilgartner and Brandt-Rauf also argued that the heterogeneous nature of data streams demands that researchers interested in data access

issues take a broader perspective of data, paying attention to the stream rather than some predetermined output of research. While emphasizing some of the bases on which scientists might assess data's value as well as how purposes for data vary across the stream, the data stream model does not specifically deal with the nature of value or valuation. As a result, I relied on philosophical theories of value that describe valuation as not only about an object's traits, but also about the benefits or purposes at issue (Beckert & Aspers, 2010; Rescher, 1969). Benefits or purposes are bound up in the socially situated meanings that objects take on for people (Blumer, 1969; Wenger, 1998). Data's meanings to scientists, therefore, were key to my examination of scientists' conceptions of data's value and their data practices.

Before presenting the findings from my study, I described in detail each of the three teams I studied (Chapter 4). I also presented three vignettes—one for each team—that demonstrated the role of data valuation throughout scientists' research processes and the influence of value conceptions on how scientists created, managed, and worked with their data. The vignettes showed how assumptions about data's value could influence the data scientists collected, the alteration of a study's design as well as the discontinuation of a study altogether, and scientists' data archiving and sharing plans.

Chapters 5 and 6 comprised the findings from my study, with each chapter focusing on a different dimension of my dissertation's research question. In Chapter 5, I answered the first part of my question: how do scientists conceive of their data's value? I found that the most central component of scientists' conceptions of data's value was data's usefulness for addressing the specific gap in knowledge their teams had defined. Valuable (i.e. good or useful) data helped the scientists answer their research questions and/or test their hypotheses, and scientists emphasized three main characteristics that made their data good. Only one of these properties was cited in the

data stream model: data's reliability (or what my scientists characterized as trustworthiness). Two other characteristics emerged as of key importance to scientists' determination that data were valuable: comparability and relevance to their specific research questions.

Scientists demonstrated nuance in their notions of data's value that went beyond whether data were good or bad, however. More nuanced value conceptions were particularly apparent when scientists were asked to consider how long their data might be of value and who and for what purpose others might find their data valuable. Using the designations "field data," "experimental data," "unpublished data," "publishable data," "raw data," and "derived data," scientists asserted that data's value varied according to their type. When describing data according to the study used to produce them (i.e. field versus experimental or modeling), scientists emphasized that data's long-term value and potential value to others depended on whether data could be combined with other similar data (i.e. metaanalysis) or be expanded upon with further study on the same system (e.g. studies that looked at time-based changes). Researchers described data from controlled studies as having much more limited value outside the team and over time than did data from field studies.

Scientists also differentiated data according to their publication status and potential. In doing so they highlighted the perceived potential of the data to yield publications for the team and the threat that sharing data would have on their own team's realization of data's value. Researchers regarded data that were likely to result in new publications for their team as having very high value that needed to be protected and exploited by the team. A team's data were—according the data stream model—rare data stream elements as long as they remained private resources. Scientists were reluctant to make them common by sharing them widely until they felt they had fully exploited data for the team's own uses. Interestingly, in using publication potential

as an important marker of data's value, scientists made it clear that—in contrast to the assertions of data sharing proponents—they regarded at least some of their data as depletable resources with finite value.

In the data stream model, inscription level was described as a key characteristic of the products scientists generate, and this was a characteristic that bore out in scientists' conceptions of data's value. When making value-based distinctions between their raw and more heavily processed data, scientists focused on whether data were means for achieving something else or were closer to products of their work. The closer data were on the processing spectrum to results (i.e. heavily processed data), the more likely scientists were to characterize them as an end goal of their work and to ascribe greater value to them. Raw data—particularly level-one data—were, on the other hand, thought of as resources for achieving something else, with a heavy emphasis placed on their instrumental value. At the same time, researchers' raw, level-one data could not be easily reproduced. Even while they were less valued as what the researchers "were really after," scientists viewed such inscriptions as practically irreplaceable resources.

In Chapter 6, I turned my attention toward answering the second part of my study's question: how do scientists enact conceptions of data's value in their practices? I found that as researchers carried out data collection and early data management activities, their data practices were predominantly aimed at producing data that were good for their own studies and identifying and rooting out data that were not. I showed how this overriding focus on the value of data for their own studies shaped the data scientists collected, the methods they used to collect data, and their evaluation of the data they produced. Beyond enacting conceptions of data's value based on whether they were good or bad, scientists' data practices exhibited negligible concern with other value conceptions, such as those based on the type of study scientists were conducting. This

changed as scientists were asked to make their data more widely available to others. As stakeholders, such as the Station, approached the scientists about depositing data in a public repository, data's value for producing publications was an especially prevalent factor in scientists' decision to share. When researchers perceived data as having high publication potential—no matter if the team had already published the data—scientists were decidedly unwilling to make the data available to people outside the team. On the other hand, when scientists felt they had wrung publication value (for the team) out of a given set of data, they were willing to share them more widely.

Secondarily, researchers indicated that data's potential use to others was a guiding factor in their data sharing decisions. Unlike publication potential, however, this did not affect scientists' willingness to share data. Rather, when researchers viewed data as having low potential reuse value—as they did with experimental data—they were not as compelled to engage in making those data publically available.

7.2 Implications

This study was designed to address a gap in our understanding of data practices, particularly the practices of researchers engaged in small science. I noted that studies of data sharing and withholding have thus far focused extensively on incentives and motivations. I argued that while incentives and motivations are important factors in scientists' sharing and withholding behavior, data practices research had thus far largely left out other potentially important factors in the decision to share data or create data that endure past the life of a project. Based on findings from previous research, I asserted that value conceptions were likely to play an important role in scientists' data practices. My dissertation has revealed several factors that

were important to how scientists conceived of their data's value, and, further, how scientists enacted value conceptions in practices that extended from collecting data to depositing those data in public repositories. The results I have presented suggest implications in two main areas: data practices research (both practical and theoretical); and efforts aimed at increasing data sharing behavior. I discuss each in more detail.

Scientific data practices research has focused heavily in recent years on understanding the various factors at play in the apparent lack of data sharing across many disciplines. Low rates of data sharing are frequently attributed—by data sharing proponents and scholars alike—to a mismatch in incentives and motivations in the scientific research context (e.g. Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010; National Academy of Sciences, 2009; Tucker, 2009). My research confirms that scientists are concerned with bolstering and protecting their ability to generate the products for which they are rewarded (i.e. peer reviewed publications); often at the expense of creating and managing data for reuse. However, my findings also suggest that there is more at play in what scientists do with data—including making data widely available—than incentives and disincentives.

Specifically, the findings from my study complicate what is frequently labeled as "withholding" behavior when scientists do not share data. An important part of scientists' decision to share data does concern data's publication value to them and their teams. When the scientists I studied viewed data as having high publication value, they made it clear they were unwilling to share the data, and their decision not to publically deposit data could then be accurately called "data withholding." However, scientists also considered some kinds of data inherently limited in value and unlikely to be of much use to others or for a much longer period of time than their study. Scientists were not unwilling to share such data, yet they may not have

made the data available because they could not imagine any benefit to others in having access to the data.

While the effect is the same in either case—data are not made available—the misattribution of the rationale behind scientists' lack of sharing has important implications for our understanding of scientists' data practices. The kind of "not sharing" behavior that occurs as the result of a perception that data would not have much value to others has very little to do with scientists' willingness to share and a great deal to do with a lack of any compelling purpose conceived in doing so. This suggests caution when interpreting the results of studies that focus on repository submission rates without also attending to the views of scientists.

In terms of implications for data practices research, my study also expands upon and adds nuance to the data stream model, which was put forth by Hilgartner and Brandt-Rauf as a framework for studying data access issues. My study lends credence to the model's emphasis on reliability, rarity, processing state as important characteristics of data stream elements. The scientists in my study were clearly concerned with data's reliability, relying on it as an indicator of their data's value (i.e. whether the data were "good") for the team. Additionally, in their use of publication status and potential to differentiate data's value, scientists demonstrated that the rarity of data influenced their data access practices. When data were perceived as still having publication value to the team, researcher wanted to maintain those data's status as rare resources that were inaccessible to anyone outside the team. As those data's value was exhausted for the team, the researchers no longer sought to maintain the data's status as rare objects.

My findings also suggest important additional characteristics of data elements that would enhance the data stream model's usefulness for understanding data access issues. In particular, my study indicated that comparability and relevance to research questions—in addition to

reliability—are important data traits in scientists' assessments of data's value. Like reliability, these two characteristics were assessed and reassessed throughout scientists' work. Furthermore, when scientists judged data to be incomparable to one another or irrelevant to their research questions, they took steps aimed at either improving data going forward or at excluding them from analyses. While neither directly concerned access decisions, such activities had implications for what was made publically available. For example, when Renee (IM-GR) determined that some of her data were incomparable, she excluded them not only from her own analyses, but also from the set of data she planned to hand over to the Station. My findings also indicate an additional key data characteristic that more directly concerns data access issues: data's expansibility. Particularly when they were asked to consider their data's value to others outside the team, scientists described the potential to add to, compare, or otherwise expand on data to develop new insights that went beyond the team's interest. Researchers studying data access practices should pay close attention to this characteristic of data elements as they examine sharing and other data access activities.

The second area where my results have important implications is in the promotion of data sharing and archiving. Recent years have seen a number of efforts aimed at compelling scientists to share the data from their research, particularly when that research is federally funded (e.g. National Science Foundation, 2011a; National Science Foundation, 2011b; United States. Executive Office of the President. Office of Science and Technology Policy, 2013). Yet, to date, compliance with mandates remains low in many disciplines (Piwowar & Chapman, 2008, 2009; Tenopir et al., 2011). My study raises two important points for those who seek to increase the production of archive-ready data as well to compel scientists to share and archive their data. First, funders, repository managers, and publishers should get scientists thinking early in their

projects about the potential value of their data to others. It is not enough to have mandates in place at the beginning of a project. My study shows that when scientists are put in a position to consider data's longer-term value, they view different kinds of their data as having different potential reuse value. Yet, there was little evidence that these distinctions were salient to scientists as they collected and managed data early in their projects. In other words, while scientists held nuanced conceptions of their data's value that might have been leveraged to influence activities like data documentation practices, they were hardly ever asked—beyond my constant questions—to think about data's potential broader value. Someone like the Station's information manager could make data's broader value more salient to scientists by interacting with them about their data earlier than at the end of the projects. For example, we saw how the information manager got Ethan (ID&M-PD) to consider that the data his team assumed to be useless to others because of their "peripheral" nature might have actually been very useful to other researchers at the Station. It is worth reemphasizing that Ethan instigated the meeting with the information manager—what kind of impact could the information manager have had on data practices had he been deliberately involved with teams earlier?

Second, my study highlights the contrast between data sharing proponents' persistent characterization of data as valuable by virtue of simply being data; and scientists' conceptions of data's value as contingent and frequently limited and depletable. I do not wish to claim that one perspective is right and the other wrong; rather I argue that vague, non-specific claims about data's value might reduce the impact of statements intended to promote data sharing and archiving. The scientists in my study were emphatic in the view that all data were not equally valuable when it came to data's potential use by others. Furthermore, there was widespread consensus across the teams about the kinds of secondary uses that were particularly meaningful.

Data's value for validation and replication were never brought up as reasons for sharing and/or archiving data. Instead, scientists were compelled by uses that involved extending, adding onto, or comparing data. Data sharing proponents should emphasize these kinds of uses—and recognize that scientists do not view all data as equally suitable for these secondary uses—as they seek to increase data sharing and archiving.

7.3 Future Research Directions

My study points to several areas that deserve further attention as researchers, funders, publishers, and repository managers continue to develop an understanding of scientists' work with data and, in particular, how scientists' practices impinge or facilitate the use of data by others. The first two research directions I suggest are intended to address the limitations of the current study and which I highlighted in Chapter 3, but reiterate here; the third represents a possible expansion on my study's contributions.

First, my dissertation focused almost exclusively on the views and activities of scientists during the field season: one segment of an entire process involved in carrying out a study. As a result, my observations were heavily tilted toward data collection activities and the valuations of data that scientists made during a fairly early stage of their projects with little observation of activities—like data analysis and data deposit—that generally occurred at a later point in scientists' studies. My study showed that conceptions of value that concerned how long data might be of value and whether others would likely find them valuable did not have much salience to scientists as they collected and organized data for their own use. However, scientists indicated that such value conceptions play a more significant role as they reach a point in their projects where others are asking them to make their data widely available. A logical extension of

my work, therefore, would focus on the later stages of scientists' research projects. Specifically, much could be added to our understanding of the role of value conceptions in data practices by looking more closely at scientists' work with data during analysis, publication, and data deposit.

A second avenue of research would test this study's findings in other settings. My dissertation focused on the views and practices of small teams of scientists conducting ecological research. Data practices studies often attend to one particular discipline (e.g. ecology, space science, genomics), which can challenge efforts to apply the resultant findings to other contexts and domains. Other researchers studying the data practices of scientists emphasize the importance of uncovering attitudes and practices that carry across scientific disciplines and identifying key dimensions of difference (Cragin, Palmer, Carlson, et al., 2010; Cragin, Palmer, & Chao, 2010).

In the current study, I deliberately selected a site that allowed me to examine the views and practices of scientists engaged in different kinds of research. However, even while the teams I studied represented various subdisciplines of ecology, they shared many similarities in their data collection and management activities that are likely unique to ecological science. In an effort to generate findings that might hold up in other settings, I highlighted the factors underlying value conceptions and their enactment in data practices. For example, while the distinction between field data and experimental data might be unique to ecological science, the use of data's capacity to be added to, compared to other data, or extended as a measure of data's potential value to others is a finding that I anticipate being replicated in many other settings and disciplines.

An expansion of my study's research in this direction might examine another small-science discipline to determine to what extent scientists' view of data as a resource with potential

secondary value depends on the ability to combine or add to data. If such a similarity is found in a different scientific domain, how does it get translated into data types? For example, one might imagine that in a field with highly technical instruments that advance quickly, scientists make value distinctions between data produced with instrument type A and instrument type B.

Finally, my findings suggest a study that examines the effect of specific interventions on scientists' views of data value and/or their data practices. In my study, scientists showed little concern with data's more expansive potential value as they carried out their studies. At my research site, the information manager deliberately waited to approach scientists about archiving their data after they completed their studies. How might earlier and more direct guidance about data management and data's potential use to others affect scientists' views and behaviors? Were I to carry out such an expansion of this study, I could imagine going back to the Station and recruiting the information manager to intervene earlier in scientists' studies and compel them to think about data's potential data's value to others as they began collecting and working with data. This could extend my findings significantly by revealing what, if any, effect such a value-focused intervention has on data practices.

APPENDIX 1: INTERVIEW GUIDE FOR SCIENTISTS

Understanding the science

1. Can you describe your work to me (not just what's going on here)? What are the kinds of research questions you're interested in? How do you go about answering them? What are the main outputs of your research? What do you consider to be your disciplinary community?

2. What journals do you publish in? What conferences do you attend? What funding agencies typically fund your kind of research?

3. Where are you at in your scientific career? What do you see as your research path going in the future?

Station Project

1. Tell me about what you are working on here this summer. What are the basic steps involved in doing this research? What kinds of data will you (or are you) be collecting? What variables are you interested in?

2. How does your work here fit in with your other research work?

3. What work was done on this project before getting to the Station and what will be done after you leave?

4. What do you hope is the main outcome or hope to accomplish by the end of your time here?

Managing and working with data

1. What are the steps involved in collecting data for this project? What is required to move from the things you're looking at to data that you can use to answer your questions and publish findings? What kinds of data do you collect and what forms are they in?

2. How difficult are these data to generate? What was/is required to get from concept to "good" data? What are good data? What are crappy data? What would it take for someone else to generate them? Are the tools or techniques you're using unique?

3. How long of a timespan is there typically between collecting the data and analyzing or using them? How will you use the data?

4. What are you doing with data to ensure that you can access them and understand them when you need to? How do you find and understand data that you've collected later? Will data be shared with team members that aren't involved in data collection? How so? How will you insure that they understand them?

5. Are all of your data treated the same with respect to ensuring access and understanding in the future? If not, what is the difference?

6. How do you know as you collect and work with data that some data are worth extra steps?

7. Are there rules within your team about how to manage data or which data should be managed? What do you tell the other team members about managing data as they work with them?

8. How did you learn what to do with data so that you could understand and use them later?

9. How long is the useful life of your data? (How long do you anticipate these data will be useful to you and your team)? What other things might you use them for? What about for others?

10. What are data for?

Managing for other uses

1. Do you think your data have (or will have) value beyond the research questions (purposes) that motivated their generation? If so, for what and for whom?

2. Do you think about long-term use of your data as you collect with, work with, and manage them? How so? Does it influence what you do with data? Do you think any of your data are worth long-term preservation? For whom? For what? Why?

3. Do you plan to share or archive your data? Which ones? Why? Tell me about data you've collected (or are collecting) that you plan to (or have already) deposit at the Station. Do you deposit your data in any other repositories? Have you made other data available by archiving or sharing it? Tell me about it.

4. Were you to make your data available to others, do you think they'd be used? By whom? For what? For how long would they be useful to others?

5. Have others let you know that your data might be of value beyond your project?

6. Have you generated other data like these in the past? Have they been useful to yourself or others over time? Have others requested to use them?

7. Do you make use of (have you ever) of data created by other scientists?

Policy and Cultural Environment

1. What, if any requirements or expectations exist in your field regarding data preservation and sharing? For example, does your funder require it? Journals that you publish in?

2. Have you received any instruction from the Station regarding what you should do with data such as how you should manage or share them? What do you think? What's your interpretation of this?

3. Are you aware of any standards for data in your field that are relevant to what you are collecting?

BIBLIOGRAPHY

Akmon, D., Zimmerman, A., Daniels, M., & Hedstrom, M. (2011). The application of archival concepts to a data-intensive environment: Working with scientists to understand data management and preservation needs. *Archival Science, 11*(3-4), 329-348.

Amann, K., & Knorr Cetina, K. (1988). The fixation of (visual) evidence. *Human Studies, 11*, 133-169.

Anderson, C. (2004, December 10). The Long Tail. *Wired*.

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

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L. et al. (2004). Promoting Access to Public Research Data for Scientific, Economic, and Social Development. *Data Science Journal, 3*, 135-152.

Association of Research Libraries. (2006). *To stand the test of time: Long-term stewardship of digital data sets in science and engineering*. Arlington, VA: National Science Foundation.

Atkins, D., Droegemeier, K., Feldman, S., Garcia-Molina, H., Klein, M., Messerschmitt, D. et al. (2003). *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*: National Science Foundation.

Attfield, R. (1987). *A Theory of value and obligation*. London: Croom Helm.

Baker, K., & Millerand, F. (2010). Infrastructuring ecology: Challenges in achieving data sharing. In J. Parker, N. Vermeulen & B. Penders (Eds.), *Collaboration in the New Life Sciences*: Ashgate.

Beagrie, N., Beagrie, R., & Rowlands, I. (2009). Research data preservation and access: The views of researchers. *Ariadne*(60).

Beckert, J., & Aspers, P. (Eds.). (2010). *The Worth of Goods: Valuation & Pricing in the Economy*. Oxford: Oxford University Press.

Bietz, M., & Lee, C. (2009). *Collaboration in metagenomics: Sequence databases and the organization of scientific work*. Paper presented at the ECSCW '09.

Birnholtz, J., & Bietz, M. (2003). *Data at work: supporting sharing in science and engineering*. Paper presented at the GROUP '03.

Bjorndal, K., Bowen, B., Chaloupka, M., Crowder, L., Heppell, S., Jones, C. et al. (2011). Better science needed for restoration in the Gulf of Mexico. *Science, 331*, 537-538.

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2008). *Sustaining the digital investment: Issues and challenges of economically sustainable digital preservation*.

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2010). *Sustainable economics for a digital planet: Ensuring long-term access to digital information*.

Blumenthal, D., Campbell, E., Anderson, M., Causino, N., & Louis, K. (1997). Withholding research results in academic life science: evidence from a national survey of faculty. *JAMA, 277*(15), 1224-1228.

Blumer, H. (1969). *Symbolic Interactionism: Perspective and Method*. New Jersey: Prentice-Hall, Inc.

Borgman, C. (2007). *Scholarship in the digital age*. Cambridge, MA: MIT Press.

Borgman, C. (2010). *Research data: Who will share what, with whom, when and why?* Paper presented at the China-North American Library Conference.

Borgman, C., Wallis, J., & Enyedy, N. (2006). Building Digital Libraries for Scientific Data: An Exploratory Study of Data Practices in Habitat Ecology.

Borgman, C., Wallis, J., & Enyedy, N. (2007). Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal of Digital Libraries, 7*(1-2), 17-30.

Borgman, C., Wallis, J., Mayernik, M., & Pepe, A. (2007). *Drowning in Data: Digital Library Architecture to Support Scientific Use of Embedded Sensor Networks*. Paper presented at the JCDL 2007.

Bowker, G. (2000). Biodiversity Datadiversity. *Social Studies of Science, 30*(5).

Bowker, G. (2006). *Memory practices in the sciences*: MIT Press.

Brunt, J. (2000). Data management principles, implementation and administration. In W. Michener & J. Brunt (Eds.), *Ecological Data: Design, Management and Processing*. Oxford: Blackwell Science.

Campbell, E., Clarridge, B., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. et al. (2002). Data withholding in academic genetics: evidence from a national survey. *Jama, 287*(4), 473-480.

Campbell, E., Weissman, J., Causino, N., & Blumenthal, D. (2000). Data withholding in academic medicine: characteristics of faculty denied access to research results and biomaterials. *Research Policy, 29*(2), 303-312.

Cedars Project Team. (2001). *The Cedars Project Report*.

Collins, H., & Pinch, T. (1998). *The golem: What you should know about science*. Cambridge, England: Cambridge University Press.

Committee on Issues in the Transborder Flow of Scientific Data, & National Research Council. (1997). *Bits of power: Issues in global access to scientific data*. Washington, D.C.

Costello, M. (2009). Motivating online publication of data. *BioScience, 59*(5), 418-427.

Cragin, M., Palmer, C., Carlson, J., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A, 368*, 4023-4038.

Cragin, M., Palmer, C., & Chao, T. (2010). *Relating data practice, types, and curation functions: An empirically derived framework*. Paper presented at the ASIST2010.

Creswell, J. W. (1994). *Research design: Qualitative & quantitative approaches*. Thousand Oaks, CA: Sage Publications.

Digital Curation Centre. (n.d.). DCC Curation Lifecycle Model. Retrieved August 23, 2011, 2011, from http://www.dcc.ac.uk/resources/curation-lifecycle-model.

Edwards, P. N., Jackson, S. J., Bowker, G., & Knobel, C. (2007). *Understanding infrastructure: Dynamics, tensions, and design*.

Erlandson, D., Harris, E., Skipper, B., & Allen, S. (1993). *Doing naturalistic inquiry: A Guide to methods*. Newbury Park: Sage Publications.

Erpanet. (2003). *The selection, appraisal and retention of digital scientific data*. Lisbon: Erpanet.

Fienberg, S., Martin, M., & Straf, M. (Eds.). (1985). *Sharing research data*. Washington, DC: National Academy Press.

Gardner, T., Cote, I., Gill, J., Grant, A., & Watkinson, A. (2003). Long-term region-wide declines in Caribbean Corals. *Science, 301*(5635), 958-960.

Gitelman, L. (Ed.). (2013). *"Raw data" is an oxymoron*. Cambridge, MA: MIT Press.

Gurevitch, J., Curtis, P., & Jones, M. (2001). Meta-analysis in ecology. *Advances in Ecological Research, 32*, 199-247.

Gutman, M., Schürer, K., Donakowski, D., & Beedham, H. (2004). The selection, appraisal, and retention of digital social science data. *Data Science Journal, 3*.

Hagstrom, W. (1965). *The scientific community*. New York: Basic Books, Inc.

Hedstrom, M., & Niu, J. (2008). *Incentives for data producers to create "archive-ready" data: Implications for archives and records management*. Paper presented at the Society of American Archivists Research Forum.

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends, 57*(2), 280-299.

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The fourth paradigm: Data-intensive scientific discovery*. Redmond, Washington: Microsoft Research.

Hilgartner, S. (1997). Access to data and intellectual property: Scientific exchange in genome research *Intellectual property rights and the dissemination of research tools in molecular biology: Summary of a workshop held at the National Academy of Sciences, February 15-16, 1996*. Washington, DC: National Academy Press.

Hilgartner, S., & Brandt-Rauf, S. (1994). Data access, ownership, and control: Toward empirical studies of access practices. *Science Communication, 15*(4), 355-372.

Hine, C. (Ed.). (2006). *New infrastructures for knowledge production*. Hershey, PA: Idea Group.

Humphrey, C. (2006). e-Science and the Life Cycle of Research.

Interagency Working Group on Digital Data to the National Science and Technology Council. (2009). *Harnessing the power of digital data for science and society*.

Kahn, S. D. (2011). On the future of genomic data. *Science, 331*(6018), 728.

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 (CSCW), 15*(4), 321-358.

Kirschenmann, P. (2001). "Intrinsically" or just "instrumentally" valuable? On Structural types of values of scientific knowledge. *Journal for General Philosophy of Science, 32*, 237–256.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Cambridge, MA: Harvard University Press.

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts*. Princeton, New Jersey: Princeton University Press.

Lauriault, T., Craig, B., Taylor, D., & Pulsifer, P. (2008). Today's Data are Part of Tomorrow's Research: Archival Issues in the Sciences. *Archivaria, 64*(0).

Lemos, R. (1995). *The Nature of value*. Gainesville: University Press of Florida.

Lincoln, Y., & Guba, E. (1985). *Naturalistic inquiry*. Beverly Hills: Sage Publications.

Louis, K., Jones, L., & Campbell, E. (2002). Sharing in science. *American Scientist, 90*(4), 304-307.

Maxwell, J. (1996). *Qualitative research design: An interactive approach* (Vol. 41). Thousand Oaks, CA: Sage Publications, Inc.

McCain, K. (1991). Communication, competition, and secrecy: the production and dissemination of research-related information in genetics. *Science, Technology & Human Values, 16*(4), 491-516.

McSherry, C. (2001). *Who owns academic work? Battling for control of intellectual property*: Harvard Univ Press.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*: University of Chicago Press.

Michener, W., & Brunt, J. (Eds.). (2000). *Ecological data: Design, management and processing*. Oxford: Blackwell Science Ltd.

Miles, M., & Huberman, A. (1994). *Qualitative Data Analysis* (Second ed.). Thousand Oaks: Sage Publications.

Najder, Z. (1975). *Values and evaluations*. Oxford: Clarendon Press.

National Academy of Sciences. (2009). *Ensuring the integrity, accessibility, and stewardship of research data in the digital age*. Washington, D.C.: National Academies Press.

National Aeronautics and Space Administration. (2009). *Guidebook for proposers responding to a NASA research announcement*. Washington, DC: NASA.

National Research Council. (2003). *Sharing publication-related data and materials: responsibilities of producership in the life sciences.* .

National Science Foundation. (2011a). Dissemination and sharing of research results. Retrieved October 12, 2011, from http://www.nsf.gov/bfa/dias/policy/dmp.jsp.

National Science Foundation. (2011b). NSF Data Management Plan Requirements. Retrieved March 12, 2012, from http://www.nsf.gov/eng/general/dmp.jsp.

National Science Foundation Cyberinfrastructure Council. (2007). *Cyberinfrastructure Vision for 21st Century Discovery*: National Science Foundation.

Nelson, B. (2009). Data sharing: Empty archives. *Nature, 461*(7261), 160-163.

Niu, J. (2009). *Perceived Documentation Quality of Social Science Data.* University of Michigan, Ann Arbor.

Niu, J., & Hedstrom, M. (2007). *Incentives and barriers in data sharing----a survey report*: Working paper. Not published.

Noor, M., Zimmerman, K., & Teeter, K. (2006). Data sharing: how much doesn't get submitted to GenBank? *PLoS Biology, 4*(7), 1113-1114.

O'Hanlon, L. (2010, November 22, 2010). 'David and Golaith' black hole clash simulated. *msnbc.com*. Retrieved from

http://www.msnbc.msn.com/id/40325031/ns/technology_and_science-science/t/david-goliath-black-hole-clash-simulated/#.T4bm979STWY

Overpeck, J. T., Meehl, G. A., Bony, S., & Easterling, D. R. (2011). Climate data challenges in the 21st century. *Science, 331*(6018), 700-702.

Palmer, C., Cragin, M., Heidorn, P., & Smith, L. (2007). *Data curation for the long tail of science: The case of environmental sciences*. Paper presented at the Digital Curation Conference Retrieved from https://apps.lis.illinois.edu/wiki/download/.../Palmer_DCC2007.pdf

Palmer, M., Bernhardt, E., Chornesky, E., Collins, S., Dobson, A., Duke, C. et al. (2005). Ecological science and sustainability for the 21st century. *Frontiers in Ecology and the Environment, 3*(1), 4-11.

Pearce-Moses, R. (2005). Glossary of Archival and Records Terminology. Retrieved October 12, 2011, from http://www.archivists.org/glossary/substring.asp?SearchString=appraisal&SearchSearch=Search.

Piwowar, H., & Chapman, W. (2008). A review of journal policies for sharing research data. *Nature Precedings*.

Piwowar, H., & Chapman, W. (2009). Public sharing of research datasets: a pilot study of associations. *Journal of Informetrics, 4*(2), 148-156

Reddy, F. (2012). Simulations uncover 'flashy' secrets of merging black holes. Retrieved January 20, 2013, from http://www.nasa.gov/topics/universe/features/black-hole-secrets.html.

Reichman, J., & Uhlir, P. (2001). *Promoting public good uses of scientific data: A contractually reconstructed commons for science and innovation*. Paper presented at the Conference on the Public Domain. Retrieved from http://www.law.duke.edu/pd/papers/ReichmanandUhlir.pdf

Reichman, O., Jones, M., & Schildhauer, M. (2011). Challenges and Opportunities of Open Data in Ecology. *Science, 331*, 703-705.

Rescher, N. (1969). *Introduction to value theory*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Research Information Network. (2008). *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs*: Research Information Network.

Saracevic, T., & Kantor, P. (1997). Studying the Value of Library and Information Services. Part I. Establishing a Theoretical Framework. *Journal of the American Society for Information Science, 48*(6), 527-542.

Shapin, S., & Schaffer, S. (1985). *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life*. Princeton: Princeton University Press.

Simberloff, D., Barish, B., Droegemeier, K., Etter, D., Fedoroff, N., Ford, K. et al. (2005). Long-lived digital data collections enabling research and education in the 21st century: National Science Foundation.

Stake, R. (2006). *Multiple Case Study Analysis*. New York, NY: Guilford Press.

Steering Committee for the Study on the Long-Term Retention of Selected Scientific and Technical Records of the Federal Government National Research Council. (1995a). *Preserving scientific data on our physical universe*. Washington, D.C.: National Academy Press.

Steering Committee for the Study on the Long-term Retention of Selected Scientific and Technical Records of the Federal Government National Research Council. (1995b). *Study on the Long-term Retention of Selected Scientific and Technical Records of the Federal Government: Working Papers*. Washington, D.C.: National Academy Press.

Talja, S. (2002). Information sharing in academic communities: types and levels of collaboration in information seeking and use. *New Review of Information Behavior Research, 3*, 143-160.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A., Wu, L., Read, E. et al. (2011). Data sharing by scientists: Practices and perceptions. *PLoS One, 6*(6), 1-21.

Tucker, J. (2009). *Motivating subjects: Data sharing in cancer research* Unpublished Dissertation, Virginia Polytechnic Institute, Falls Church, VA.

Uhlir, P., & Schröder, P. (2007). Open data for global science. *Data Science Journal, 6*.

United States. Executive Office of the President. Office of Science and Technology Policy. (2013). Increasing Access to the Results of Federally Funded Scientific Research. Retrieved March 28, 2013, from http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

Van House, N. (2004). Science and technology studies and information studies. *Annual Review of Information Science and Technology, 38*, 3-85.

Vertesi, J., & Dourish, P. (2011). *The Value of Data: Considering the Context of Production in Data Economies*. Paper presented at the CSCW 2011.

Wallis, J., Borgman, C., Mayernik, M., & 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).

Wallis, J., Borgman, C., Mayernik, M., Pepe, A., Ramanathan, N., & Hansen, M. (2007). Know thy sensor: Trust, data quality, and data integrity in scientific digital libraries. *Research and Advanced Technology for Digital Libraries*, 380-391.

Weiss, R. (1994). *Learning from strangers*. New York: The Free Press.

Wenger, E. (1998). *Communities of practice*. Cambridge: Cambridge University Press.

Whitlock, M. (2011). Data archiving in ecology and evolution: Best practices. *Trends in Ecology and Evolution, 26*(2), 61-65.

Whitlock, M., McPeek, M., Rausher, M., Rieseberg, L., & Moore, A. (2010). Data archiving. *The American Naturalist, 175*(2), 145-146.

Yin, R. K. (1994). *Case study research: Design and methods* (Second ed.). Thousand Oaks, CA: Sage Publications.

Yin, R. K. (2009). *Case study research: Design and methods* (Fourth ed.). Thousand Oaks, CA: Sage.

Zimmerman, A. (2003). *Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists.* Unpublished Dissertation, University of Michigan, Ann Arbor, MI.