

Pre-publication version

Final version appeared in *Science, Technology, & Human Values*, 2008, vol. 33, no. 5, pp. 631-652.

<http://sth.sagepub.com/content/33/5/631>

New Knowledge from Old Data:
The Role of Standards in the Sharing and Reuse of Ecological Data

Ann Zimmerman
School of Information
University of Michigan
105 S. State Street
3438 North Quad
Ann Arbor, MI 48109-1285
USA

Work phone: +1-734-764-1865

email: asz@umich.edu

In this paper, I analyze the experiences of ecologists who used data they did not collect themselves. Specifically, I examine the processes by which ecologists understand and assess the quality of the data they reuse, and I investigate the role that standard methods of data collection play in these processes. Standardization is one means by which scientific knowledge is transported from local to public spheres. While standards can be helpful, my results show that knowledge of the local context is critical to ecologists' reuse of data. Yet, this information is often left behind as data move from the private to the public world. The knowledge that ecologists acquire through fieldwork enables them to recover the local details that are so critical to their comprehension of data collected by others. Social processes also play a role in ecologists' efforts to judge the quality of data they reuse.

Keywords: data sharing; data reuse; ecology; objectivity; standardization

Ecologists are on the cusp of potentially significant changes to their social and scientific practices. These transformations are being driven in part by the collection of new types of data, by the online availability of large volumes of data of interest to ecologists, by technologies that make it easier to manage and integrate disparate data, and by slowly changing views about the value of secondary analysis to address important ecological questions. The increased availability of data along with demands for researchers in ecology and other fields to share and reuse data could transform the culture, practice, and communication of science (e.g., Brown 2003; Hilgartner 1995). Yet, this vision is not a foregone conclusion (Hine 2006), and a host of problems make the benefits of data sharing and reuse remarkably difficult to realize. These challenges include issues of data ownership, a lack of incentives for scientists to share, technical hurdles related to incompatible hardware, software, and data structures, and costs to document, transfer, and store data. While scholars have yet to fully grapple with these problems, especially in small sciences such as ecology, other authors have pointed to a lack of standards as one of the major impediments to the sharing and reuse of scientific data. They assert that successful systems of data sharing depend on various kinds of formal standards, including those related to data collection, description, storage, and quality control (e.g., National Research Council [NRC] 1995, 1997).

Standards are a set of instructions that specifies how something will be done, and they play an important role in the transfer of knowledge from one location to another (Berg 1997; Edwards 2004; Porter 1995). As Berg stated, standards "are vehicles through which order can be brought to all those practices where variation reigns" (1997, 1082). Standards can be developed and applied to a wide range of areas such as the specification of data storage in a computer, categorization of animals, people, or things, and diagnosis of medical ailments. In this paper, I focus on the use of standard approaches to collect ecological data and their relation to data reuse. These standards, when they exist, consist of methodological protocols, instruments to measure physical or biological properties, and guidebooks for the identification of plant and animal species. The emphasis on the function that standards play in the transport of knowledge from one location to another naturally alters the extent of reliance on informal knowledge, which is also an important part of information reuse. In order to address gaps in our understanding about the role that the presence or absence of standards has on knowledge transfer, I studied the experiences of ecologists who used shared data. I find that although standardized methods of

data collection can be helpful to the secondary use of data, informal judgment is an important part of the process of creating new knowledge from existing data. The abstractions and generalizations that are essential to the conversion of scientific findings into public knowledge often obscure the local, contextualized information that makes the findings reusable.

In the remainder of this introduction, I review previous literature and define how *data* and *reuse* were used in this investigation. In the section that follows, I describe further the concept of standardization and the theoretical framework that provided the lens under which ecologists' experiences were analyzed. In the remaining sections, I provide an overview of the field of ecology and the nature of ecological data, and I discuss the methods and results from my investigation. I conclude with a discussion of the study's implications and limitations.

Data and Data Reuse

Data are central to many studies of science, including those that have analyzed scientific controversies, investigated the work practices of scientists, and posed hypotheses about technology's affects on the production and communication of knowledge (e.g., Beaulieu 2001; Brown 2003; Collins and Pinch 1998; Hilgartner 1995). Within this context, most authors have defined data in broad terms. For example, Hilgartner and Brandt-Rauf included "biological materials, reagents, novel instrumentation, and other scarce or limited resources" in their definition of data (1994, 356). Similarly, McCain (1991) described the research-related information exchanged by geneticists to encompass raw data, computer programs, extensive tables and text, craft knowledge, and physical research products. Hilgartner and Brandt-Rauf also argued that "data should be conceptualized not as the end-products of research, but as part of an evolving data stream" (1994, 359). In this article, data refer to inscriptions that appear in the form of measurements and observations of the natural world. In addition, my definition includes information relevant to the data that is independent of the data themselves but without which the data would be incomprehensible. Examples of pertinent supporting information include a description of the methods used to obtain an observation or to conduct an experiment; the location of an observation or experiment; and attributes associated with an observed species, such as taxonomic information, physical characteristics, or natural history information. Much prior research has been concerned with access to data, especially the conditions under which

scientists grant or deny access (e.g., Campbell et al. 2002; Hilgartner and Brandt-Rauf 1994; McCain 1991). Little attention has been paid to how data are reused once they are obtained, which is the main interest of the study reported here.

Sociologists of science have studied how knowledge is constructed when researchers conduct laboratory experiments that generate new data (Latour and Woolgar 1979). In related work, they have investigated the social interaction necessary for scientists to replicate experiments and duplicate instruments (Collins [1985] 1992). The reuse of data to generate new knowledge, which might be expected to be both similar to and different from replication and construction, has seldom been a topic of study. To address this gap, I studied ecologists who gathered data from published and unpublished documents, computerized databases, and institutional records, and from researchers both known and unknown to them and who used these data to study a new research question. Thus, I define *reuse* as the use of data collected for one purpose to study a new problem.¹ My specific aim was to examine the processes by which ecologists understand and assess the quality of data they did not collect themselves and to analyze the role that standard methods of data collection have on reuse. Although the ecologists I studied sometimes knew personally the individuals who originally collected the data they reused, these relationships did not fundamentally alter the decisions that ecologists made about whether or not to use particular data. In all cases, an ecologist's ability to understand data was the key to their reuse, and it was the knowledge that ecologists acquired through the collection of their own data in the field or laboratory that enabled them to comprehend data collected by others.

Conceptual Framework: Standards as Distance Spanners

The use of data outside their original context implies distance. In order for data to be reused, they must be able to travel beyond the location in which they were produced. Sociologists and historians of science have identified standards as one system by which scientific knowledge moves from the local sphere into the wider world. A lack of standards, particularly for data collection, storage, and management has been continually identified as a major impediment to the sharing and reuse of scientific data. For these reasons, I employed a

conceptual framework that focuses on standardization as a vehicle to overcome distance. In particular, I drew from Theodore Porter's (1995) theory of measurement as a social technology and similar ideas embodied in Bruno Latour's (1999) concept of circulating reference.

Porter explored the history of quantification. His ideas, particularly the notions of *objectivity* as a technology of distance and *standardization* as a substitute for lack of trust, are useful for casting light on the study of secondary data use. One of Porter's key points is that quantification is a technology of distance that is well-suited for communication that goes beyond the boundaries of locality and community. Other authors have also described how standards in measurements, money, and writing contribute to the successful management of distance (Alder 2002; King and Frost 2002). Standard approaches to gather, describe, and store data make it easier to aggregate and compare data gathered in multiple studies, and thus they are one means to overcome the distances that scientists can encounter when trying to use data they did not collect themselves.

Standards are important in transforming local knowledge into public knowledge. In order to perform this function, however, standard measurements involve a loss of information, or what Latour (1999) referred to as *reduction*. Reduction allows for “much greater compatibility, standardization, text, calculation, circulation, and relative universality...” (Latour 1999, 70). By reducing the world to inscriptions, local knowledge becomes public knowledge, and in this way it becomes amplified. Latour defined the dual ability of science to bring the world closer yet also to push it away as *circulating reference*.

How does one move from the first image to the second—from ignorance to certainty, from weakness to strength, from inferiority in the face of the world to the domination of the world by the human eye? ... The sciences do not speak of the world but, rather, construct representations that seems always to push it away, but also to bring it closer (Latour 1999, 30).

Although activities in a laboratory are most often the locus of study for sociologists of science, Latour made the point above in his written account of field scientists working together in the Amazon forest. Latour observed, “For the world to become knowable it must become a laboratory” (1999, 43).

Despite their power to negotiate distance, standards can be difficult to achieve for cultural, political, scientific, and technical reasons. All these factors apply to ecology, as I will describe in the next section. Standardization requires agreement, coordination, and stability (King and Frost 2002; Mallard 1998). These are not small considerations, and they can make standardization arduous and expensive to achieve (Berg 1997; Mallard 1998).

Ecological Data and the Practice of Ecology

Ecology is the study of the interrelationships between the earth's organisms and their environment. As members of a community of practice, ecologists share an interest in a domain of knowledge and a set of approaches to help them deal with this domain successfully. The community of practice concept encompasses the formal and the informal. "It includes what is said and what is left unsaid; what is represented and what is assumed" (Wenger 1998, 47). Ecologists conduct experiments in field and laboratory settings and observe phenomena in the natural world using a formidable array of tools and approaches that they learn as part of their enculturation to the discipline (Kohler 2002; Roth and Bowen 2001a, 2001b).

Wherever they work, ecologists are confronted with variance among individuals within a population, with the unpredictable character of the natural environment, and with immensely complex systems with large numbers of variables and many subtle interactions. For example, even with a field guide and specimens in hand it can be difficult to distinguish one tree from another because a guidebook cannot show all the possible variations in leaves, bark, or structure that occur over time. The result is that ecological data are unusually complex, presenting daunting problems of interpretation and analysis that pose particular problems for secondary use (cf., Bowker 2000; Bowser 1986; NRC 1995, 1997).

The culture of ecology also hampers the sharing and reuse of data. A relatively young science, ecology is currently characterized by modest single investigator studies conducted at limited spatial and temporal scales (NRC 1995). Ecologists do not have an established infrastructure for sharing data, although efforts are underway to change this (Michener, Brunt, Helly, Kirchner, and Stafford 1997; Jones, Berkley, Bojilova, and Schildhauer 2001). In general, sharing takes place between close associates and relies heavily on social interaction (Committee on the Future of Long-term Ecological Data 1995). The reasons for this are complex and include

few incentives for ecologists to share and a culture that values creative and independent research above secondary use of data.

The complexities of environmental problems, funding opportunities and pressures, and the online availability of large datasets are prompting ecologists to ask new questions and to look for ways to expand the spatial and temporal scales of their science. Ecology is self-consciously attempting to transform itself in order to better identify, understand, and manage environmental challenges such as vanishing biodiversity, changing climate, and altered biogeochemistry (e.g., NRC 2003; Senkowsky 2005). The most ambitious effort to transform ecology to big science is the National Ecological Observatory Network (NEON), still in the planning stages. If funded, NEON will consist of a vast network of sensors and associated technological and facilities infrastructure to support the study of continental-scale ecological processes (NRC 2003). This broader agenda challenges habituated ways of seeing the natural world and makes formerly unobservable processes candidates for measurement (Senkowsky 2005). It also creates the need for new epistemological tools to deal with large volumes of data.

Methods

The results presented in this paper were part of a larger investigation designed to address the following question: What are the experiences of ecologists who use shared data? My primary method of data collection was semi-structured in-depth interviews with thirteen ecologists who reused data. I identified potential subjects to interview by searching the methods and acknowledgments sections of 1999-2001 issues of two prominent Ecological Society of America journals, Ecology and Ecological Applications, to locate articles whose results were based wholly or partially on secondary data use. I chose the three-year time period in an attempt to get a large pool of subjects while also trying to locate articles that were recent enough to help ensure that the data reuse experience could be remembered. In addition, I selected papers to obtain diversity in the types of data reused and in data sources accessed.

The interview questions that are most pertinent to the results reported here focused on ecologists' experiences in understanding and judging the data they reused. For example, I asked ecologists to describe their greatest challenge in using the data; the role, if any, standards played in their use of the data; how they judged data quality; how the experience of using data collected

by others differed from using their own data; and what information would have made it easier for them to use the data. Interviews were conducted face-to-face and over the telephone and took place between June 2001 and February 2002. On average, interviews lasted 90 minutes.

Eight of the papers that reused data were published in Ecology, and five were published in Ecological Applications. Five papers each were published in 1999 and 2000, and three papers were printed in 2001. All papers had a minimum of two authors. The maximum number of authors affiliated with a paper was five and the average numbers of authors was three. The purpose of the research reported in the papers that ecologists published in Ecology and Ecological Applications varied, but they can be categorized as testing existing models, theories, or methods or as examining ecological processes at larger spatial or temporal scales. Some papers combined the authors' own data with that collected by others, but in all cases, data collected by others were a key part of each paper's findings.

I conducted one interview with each respondent, and I interviewed one author associated with each journal paper. In all cases except one, the ecologists I interviewed were also the first authors of the published papers.² At the end of each interview, I asked each ecologist how many years he/she had been an ecologist because I wanted to study the relationship between experience level and the process of reuse. I left it to each ecologist to choose the point of reference from which to calculate his/her years of experience. Based on this, I divided ecologists into two categories after the interviews were complete: experienced and less experienced. Experienced ecologists are defined as those with 15 or more years of experience, and less-experienced ecologists are defined as those with 14 or less years of experience. Under this classification, at the time of the interviews, five ecologists qualified as experienced (average = 26.8 years) and eight were defined as less experienced (average = 9.125 years). I planned initially to use the number of years since a Ph.D. degree was obtained to determine years of experience. Many studies have used this metric (e.g., Oromaner 1977), but I found that, at least for ecologists, this metric has deficiencies. As I illustrate in the sections that follow, ecologists' identities are tied strongly to their own experiences in the field or laboratory, and therefore, educational level or years since degree is not the most useful indicator of experience. For example, Alan and David had completed their doctoral degrees in the last five and seven years, respectively, but each had 20 or more years of professional work experience.³

Locating and Acquiring Data

In this section, I briefly report the methods that ecologists used to locate and acquire data, the sources from which they obtained data, and the types of data they reused. Further details about these findings have been reported elsewhere (Zimmerman 2007).

The ecologists I interviewed used a wide variety of data (e.g., species observations, plant, soil, and water chemistry measures, and experimental lab and field data) from a diverse array of sources. Four ecologists drew most of the data they reused from an existing electronic database that was available on the Internet, on CD-ROM, or from an institution. These databases contained information on instances of biological control, water chemistry, forest characteristics, and tree rings. In most cases, however, ecologists gathered single data points from multiple sources such as journal articles, reports, computerized databases, institutional records, and people and aggregated them to form a new data set. For example, one ecologist described how she extracted data from tables, graphs, and text that appeared in multiple journal articles. Another collected data on bird observations from natural history museums and also wrote letters to individuals both known and unknown to him that he thought might have data he needed.⁴ Three ecologists conducted a meta-analysis, which by definition, implies the use of multiple data sources.⁵ Sometimes ecologists knew personally, or by reputation, those who had originally collected the data, but at other times they did not. As I show in later sections, personal knowledge of the data collector, or whether or not the work involved meta-analysis, did not fundamentally change the approach that ecologists used. The ability to understand data was the most important factor for reuse, and it was this requirement that binds ecologists' experiences together.

Results

Gathering one's own data helps with reuse of data collected by others. The ability to comprehend data is the key to their reuse, and ecologists rely heavily on knowledge from their own experiences in the field or laboratory in order to understand data and to judge their quality—two closely linked processes.⁶ Although much of the knowledge they rely upon is informal, ecologists are able to articulate many aspects of the field-based knowledge they employ to reuse

data. In the sections that follow, I show that there is an element of informal judgment even in the use of ostensibly standardized data.

The Relationship of Ecological Identity to Reuse

Whether ecologists collect their own data or reuse data gathered by others, there are certain qualities that define one as an ecologist. Ecological researchers have the specialized knowledge to attempt to distinguish true patterns or variations in nature from artifacts of data that are the result of inaccurate observations or experimental error. The ability to do this requires ecologists to have particular knowledge in order to make decisions about what data to acquire, to understand the data in order to use them appropriately, and to make informed interpretations about ecological processes. Formal disciplinary training along with insights gained in the field lead to familiarity with particular types of data. As one ecologist explained in talking about the data he reused, "I chose it to be the kind of information that is readily available and that I am familiar with." Ecologists who use data collected by others strive to comprehend them as well as data they collect in the field or laboratory themselves. They do this in order to insure the quality of the results they produce through the use of those data and because their findings, as with any scientific knowledge they create, will be subject to future peer scrutiny. As Ellen, one of the interviewees, said in regard to her reuse of data, "If I honestly could not figure out what they had done, then I would not use that data point." Other ecologists reiterated this sentiment. Ecologists work hard to understand data they reuse, and they will not use data they do not comprehend. Of course, the act of understanding is itself a complex and contextualized process.

The special importance of field-based insights is reflected in the similar responses I received when I asked ecologists if they felt it was necessary to have collected their own data in order to understand data collected by others. This question encouraged ecologists to consider what it is that distinguishes them from scientists in other fields, and it provided a view into the aspects of their knowledge that they drew from to reuse data. Ecologists often took their specialized knowledge for granted. Upon reflection, though, ecologists like Bill were able to articulate some aspects of their knowledge:

Well, I don't know if it is necessary, but I think it is important for an ecologist to find some way to root their ideas in reality. An armchair ecologist often has ideas that have no basis in reality. I do think it's important that you do a mixture of synthesis as well as actual data collection in the course of your career.

Ellen, who had combined her own data with forestry data collected by others in her earlier Masters work, provided further insight into Bill's statement about the need to "root ideas in reality." Ellen explained her advisor's influence on shaping her thinking in this regard.

One of his points was that in order to be a good modeler you have to understand what the data mean because you could plug some sort of regression into a model, but if you don't know how the natural system actually is responding... I mean what the variations about that linear line might be or whether maybe it is not linear. Maybe it is an exponential function or logarithmic or maybe it is, you know, who knows? You are not... All you are is a computer scientist. You are not an ecologist.

An ecologist possesses the knowledge to "understand what the data mean." In other words, ecologists are able to make informed judgments about whether data accurately represent the natural world, and they are able to separate spurious data from accurate representations of ecological processes. Of course, this does not mean that ecologists will always accept each other's determinations—communities of practice thrive on diversity (within limits) as well as harmony—but they share the view that familiarity with data is integral to reuse.

Based on their specialized knowledge, ecologists attempt to "see for themselves" the original data collection. As Charles, noted, even seemingly "simple" data, such as a measure of elevation, require particular insight to understand.

I think doing fieldwork is a big help in understanding the data. ... Even something as clear as "elevation"—without the field experience you don't know how that variable might have been measured and the common errors involved with its measurement.

Ellen used data from a large forestry database compiled by the federal government. She spent time in the field with the crews who collected the data in order to learn more about their measurement procedures. She described how the ability to visualize the measurement of tree height helped her once she was back in the office analyzing the data.

How much can we rely on this deviation? How hard is it to measure deviation in the field? Or, do you really want to use height data in your model because, you know, it is really wicked hard to measure height, especially the way they do it. You can't get far enough away from the tree and maybe in a really thick stand it is going to be harder. Those kinds of issues and really getting in the ground and being out in the field really help you to understand how to analyze the data.

Ellen's experience in the field also provided her with personal knowledge of the skills and dedication of the data collectors that gave her added confidence in the data. The remarks from Charles and Ellen are representative of statements made by other ecologists I interviewed. As Michael summarized it, time in the field "gives you an appreciation for how the data are actually collected."

Several ecologists noted that insights gained from the field are used to *reconstruct* data. As Ellen told me about the data she reused: "You'd have to go back and try to reconstruct." Katherine said:

Every type of data collection could be optimized in different ways. You know, I feel like I keep modifying and learning this my whole life. But in fact, one of the best things that I really learned was as a graduate student. As a grad student, I learned what data are necessary for someone to repeat your work. And so, you know, even something as simple as an ANOVA table — the most common error that I came across in the literature are people reporting F values and a p value. And there is no way from an F value and a p value that you can reconstruct what that person did. What their model was. There's nothing.

To the social science observer, *reconstruction* might seem more like a process of *deconstruction*, but since the former is the word that ecologists used, it is interesting to consider why. Deconstruction refers to the action of undoing; one starts with the end product and moves backward. Whereas, the ecologists I interviewed mentally transported themselves to the sites where the data they were considering for reuse were originally collected, and from this mental vantage point they visualized the data gathering process — literally from the ground up.

An ecologist's depth of informal knowledge is influenced by years of experience, and thus, the use of this knowledge differs in subtle ways from one ecologist to another. Less experienced ecologists sometimes talked about the role of veteran scientists, who they relied on to help them reuse data until they possessed specialized, local knowledge themselves. This was most evident among ecologists who referred to work they did as Master's students. Ellen's advisor, for example, alerted her to the need to weigh all data, her own and that collected by others, against her knowledge of ecological processes. Susan noted that her inexperience as an aquatic ecologist was the largest hindrance to her understanding of the data she acquired. Thus, she relied on others to help her.

Then again, I had a lot of people I could ask. So, I think that was the biggest challenge because I was a new grad student and didn't know a lot about lakes. I was learning about aquatic ecology. ... I think the best metadata was directly from the people. I was lucky in that sense that there were a lot of people around that knew... that were familiar with the water chemistry data and familiar with how you define watersheds, and I had a lot of direct help.

Experienced ecologists also convey to those they mentor that not all data are to be trusted equally because skills vary among data collectors. For example, scientists familiar with water chemistry measurements pointed Susan to data they deemed trustworthy and steered her away from sources they viewed as unreliable.

Ecologists discussed several specific aspects of knowledge that they gained through the collection of their own data and that they relied on to use data they did not collect themselves. Their experiences in the field or laboratory, in combination with formal disciplinary knowledge, provided ecologists with the expertise to understand the critical link between research purpose,

methods, and data; to recognize the limitations of particular types of data; and to visualize potential points of data collection error. Ecologists also related the important "sense" of data, a tacit form of knowledge, which they gained by gathering their own data.

Recognizing the Importance of Purpose

Ecologists discussed the importance of knowing that the purpose for which data were gathered guides appropriate reuse of them. It was in response to my question about the role of standard methods to secondary data use that prompted ecologists to note the critical link between research questions and data. The purpose for which data were gathered is connected to their reuse in several ways. First of all, research purpose dictates methodological choices, which in turn affects the data that are generated. As Andrea noted, "so much in the results depends on how you did the study," and ecologists recognize that "people use different things for different reasons." Numerous factors affect the selection of research methods by ecologists, including the scientific question to be addressed, the environment in which a study is conducted, the taxa to be studied, and practical considerations, such as time, money, and skill. This may seem like common sense, but it is a point worth highlighting to avoid situations in which overzealous advocates may attempt to place burdensome standards on research communities. Andrea related how the methods selected to measure phosphorus in plants can vary depending on the goal of the research.

Well, I guess in a lot of cases, the methods that you use depend on the questions that you are asking. So, for example, I will go back to the example of soil phosphorus, where I was talking about the sequential extraction method. Well, so there are these six different forms, or whatever—six or eight or seven or something like that—different forms of phosphorous that are found in soils. And if you really want to know what the total is you have to do the sequential method where you extract one form after the next. But plants can only use some of those forms. There are certain forms of phosphorous that are completely unavailable for plant uptake, and so if you are interested just in the total amount of phosphorous in the soil, you do need to do the sequential methods and then add

them all up. But, if what you are interested in is how much of that phosphorous plant roots can actually take up and plants can use then you need to only look at two forms of that phosphorous. And if you are interested in phosphorous pollution leading to algal blooms, then there the forms of phosphorous are very specific as well because there has to be phosphorous that can become dissolved in the water and available for algae. So, that is another thing altogether.

Separately, Andrea and Katherine noted that the preferred method of measuring plant tissue nitrogen requires an expensive machine, and that scientists who cannot afford the instrument may rely on an earlier method. Depending on the research purpose, the use of an older method may not negatively influence the study results. Cal summarized it by stating, "As with any kind of data collection, there are always a lot of factors that go into how you choose your methods." There are legitimate reasons for the use of different research methods, which helps explain why the ecologists I interviewed did not place an overriding emphasis on methodological standardization. What is important is that a secondary data user is able to discern the methods used to generate data from different data sets, so that, as Susan said, "If they are different at least you know why."

Research purpose dictates the methods that are used to collect data, and this, in turn, limits secondary use of those data. All data have limitations, and these limitations are pitfalls to reuse if they are not understood. Ecologists employ their knowledge about the relationship between purpose, methods, and data limitations to make sophisticated decisions about appropriate reuse of data.

Dealing with Uncertainty

Data comprehension is closely related to many assessments ecologists make about the quality of data they consider for reuse. The ability to visualize data collection and to understand where errors can occur is a key aspect of ecologists' abilities to judge data quality. Knowledge about what can go wrong comes primarily from ecologists' own data collecting experiences, especially their assessments of the degree of skill required to gather particular data, and from their perceptions of or personal knowledge about the competence and commitment of specific

data collectors. As I show in the text that follows, trust in data sometimes stems primarily from *what* is being observed or measured while at other times it is tied closely to *who* is doing the work. In either case, the results I present in this section demonstrate that standard data collection approaches are inadequate substitutes for indicators of data quality and that there is a social element present in all aspects of ecologists' data reuse experiences. However, the uncertainties that ecologists confront and deal with rarely appear in their published works (Star 1985).

Dealing with Uncertainty: Focusing on the Object of Study

The variability of specific parameters in nature affects how “easy” data are to collect and ecologists weighed this factor when they made data reuse decisions. Fricker described this as trust based on empirical knowledge “of whether the topic is one about which people are generally trustworthy” (2002, 382). In this case, the object that is being observed or measured is the focus of attention. For example, Michael chose to exclude a certain Phylum from the zooplankton data he acquired because he knew that they are hard to identify.

I think sort of the main difference is that people are different in how good they are at identifying species. Actually, I didn't include rotifers, which is another group of zooplankton in the study because the taxonomy on them isn't as good. Some people would go to a lake and find four species of rotifers where I was sure that if somebody better went they would find plenty. So, those just seem more suspicious. People who are bad taxonomists are going to find two species. Good taxonomists are going to find lots of species. This is a function of people—how good people are in identifying them.

Michael reduced his concerns about data quality by eliminating rotifers from the data he collected. Michael chose not to include this group of hard to identify organisms because his uncertainty about the data was outside his personal level of comfort. He did, however, choose to include crustacean data even though there are some species that are difficult to distinguish from one another. Michael noted, however, that with crustaceans, “There are a few like that as opposed to a lot like that,” and so he was willing to accept the data despite some uncertainty.

When using data collected by others, it is not always possible to see what one would like to see, a situation that leads to uncertainty. Each ecologist I interviewed had a different tolerance for uncertainty. Susan's concerns about the quality of the data she reused were lessened by the fact that she chose a water chemistry variable, dissolved organic carbon (DOC), which does not vary greatly from season-to-season.

Yea, I would say with each data point comes some uncertainty—potentially different methods and the different data sets. There can be some year-to-year variability. But I would say, part of why I used DOC is that it doesn't vary as much as a lot of other things from year to year. And I would say the methods for measuring it on... I would say... it's a lot more standard than a lot of other things. For example, looking at soil chemistry can be more complicated.

Susan's comments indicate that low variability in nature in combination with standard methods of measurement can facilitate ecologists' understanding of data as well as help them to make quicker assessments about data quality. Ecologists scrutinized each data point before they reused it, but they were able to assess some data more quickly than others. While it is difficult to know if it would be possible for ecologists to reach agreement on what is "simple," their responses provide clues to situations in which standardization is most likely to help them to understand data and to speed up their ability to assess data quality.

Dealing with Uncertainty: Focusing on the Data Collector

When data are difficult to collect or tolerance for data uncertainty is low, ecologists call upon other, more personal, insights from the field to assess the trustworthiness of data sources. In these cases, the first focus of their scrutiny is on *who* originally collected the data. Since judgments about an unknown data collector's skill are often difficult to assess based on information that accompanies data, it is not surprising that firsthand knowledge of the skills or values of other scientists, whether positive or negative, enters into reuse decisions. For example, Stephen spoke about "commitment to the organisms" when he explained to me why he used

species lists generated by scientific programs that he was familiar with and that had been around for a long time.

If you wanted to do this today and use modern species lists I suspect a lot of it would be species lists generated by technicians. Whereas, the lists that I am using are generated by graduate students and professors — people who are really spending a lot of time becoming specialists in identifying these organisms. Not to say the technicians don't, but I do see a difference. You know, just sort of a different level of commitment to the organisms.

Stephen made judgments about the skill of data collectors before he began to gather data to reuse, and in this way, he sought to avoid the need to make quality judgments once he had the data in hand. Stephen and others defined commitment in terms of consistency and qualifications of personnel, including years of experience and dedication to the work at hand. Alan, who collected data on a bird species that is hard to identify, eventually decided to eliminate multiple data points from one data source because he had firsthand knowledge of some of the data collectors.

The reason I didn't use it was that some of these people running the routes don't know what they're looking at and counting, and I found that out because I was working with two people running some routes in Maine because one of my study sites was in Maine. They helped me, and then I went over to help them on their survey. They were seeing little brown shorebirds out in wetlands and misidentifying large numbers of them. So, I got to thinking that maybe that's the case in other segments of this database, so I did not use it.

Alan attributed these misidentifications to the fact that observers change frequently. The existence of standard field guides that assist people to identify phytoplankton and shorebirds did not assure Stephen or Alan of data quality. Their reasons for choosing not to use the data from particular databases demonstrate the complex arrangements of knowledge that make up many assessments of data. It is important to note, though, that while ecologists use personal

knowledge about data collectors when it is available to them, these insights play a secondary role in ecologists' choices about whether or not to reuse particular data. Knowing people and recognizing who does good work does not lead to automatic acceptance of data, but it does help to lessen concerns about data quality. Bill, an experienced ecologist, admitted that, "Because you know something about a person's work—they're widely published and have done a number of things—and by reputation, you will tend to trust the information." Tanya trusted data she used because it was collected by "professional dendrochronologists." As an inexperienced ecologist, she had little first-hand knowledge of these scientists, and so without information to the contrary, she perceived them as trustworthy.

Ecologists go to great lengths to ensure their understanding of data. As Bill said, "There was an iterative process of going back to the paper maybe multiple times to extract more information from it. If the information wasn't there, we would either find another source or contact the authors, if we could." Ecologists have to work hard to comprehend data they reuse. Personal insights and connections can provide access to information that aids comprehension of data or speeds up the process of determining quality, but they are not a substitute for disciplinary knowledge and fieldwork experience.

Ecologists also described their experiences in the field or laboratory as giving them a "sense" for data, and they drew from this insight to reuse them. As Nancy said, "When you're in the field, most of what you learn is not the data points you're collecting—it's just that sense." Or, as Alan phrased it, "Once you have done similar work, you kind of get a feel for—I think I do, anyway—how people operate in the field." Summarizing the findings of sociologists of science, Porter said, "There is an element of unarticulated expertise built into every attempt to solve problems according to explicit rules..." (1995, 214). My findings confirm Porter's observation. Field and laboratory experiences acquaint ecologists with the vagaries of particular data, which help them to understand and judge the data they reuse.

Discussion

In this paper, I analyzed the experiences of ecologists who used data they did not collect themselves. The results of my investigation show that while standard approaches to data collection can make it easier for ecologists to understand and judge data they did not collect

themselves, there are many instances in which methodological standardization is difficult, if not impossible, to achieve. Alternately, even when they do exist, standards do not function as substitute measures for data quality because they do not tell a secondary user if the measurements or observations were made skillfully. Ecologists use formal disciplinary knowledge along with insights gained through fieldwork to understand and assess the quality of data they reuse in both the presence and absence of standards. The fact that ecologists are able to articulate many aspects of their informal knowledge means it is possible to identify and incorporate some of this knowledge into formal data reuse systems and to make it available to scientists in other disciplines. Berg stated, "Through explicating that which was implicit, through making public what was private, patterns of practice become open for scrutiny and contestation" (1997, 1086). My results also show that trust alone is an insufficient vehicle to transfer knowledge about data across distance. Certainly, if trust does not exist, data will be rejected out of hand. However, it is the ability to comprehend data collected by others that is the key to their reuse, and trust does not play a significant role in understanding.

This study looked at data sharing within one community of practice, and it analyzed reuse of ecological data among ecologists with fieldwork experience. In their study on the enculturation of field ecologists, Roth and Bowen (2001b) observed that field ecologists might differ from theoretical ecologists or from management-oriented ecologists. Thus, the reuse of data by other types of ecologists might be different. However, science studies scholars have shown that tacit knowledge and social exchange are important components of many fields. So, while the data reuse experiences of field, experimental, and theoretical scientists may differ, evidence also strongly suggests that informal knowledge plays an important role in many, if not all, domains (e.g., Collins [1985] 1992; Knorr Cetina 1999). In addition, my study focused on "successful" data reuse experiences, and therefore, it may exaggerate ecologists' abilities to overcome the challenges described in the literature. In spite of these limitations, my findings help to explain the success of formal data sharing systems, which has implications for ecologists' efforts to increase the sharing and reuse of data. My results also suggest changes for the education of future ecologists.

Formal data sharing systems emphasize standardization, peer review, and quality control, but their real strengths may come from the forum they provide for discussion; from the opportunity they offer the individual to capitalize on collective wisdom; and from the presence of

intermediaries who bear much of the labor of cleaning, describing, storing, packaging, disseminating, and preserving data (Markus 2001). These benefits can reduce the mental and physical energy that scientists must expend to reconstruct, integrate, and judge data. One of the ecologists I interviewed, who had a previous career as an economist, provided insight into how this works with data such as those made available by the U.S. Census Bureau.

The economics data is often much more organized and processed. In economics, typically people are working with a shared data set. There are hundreds of people that work with the current population survey, for example, and you can go and find out, “Well, what are the problems with this data set?” Everyone can tell you, “Oh yeah, ’79 was a really bad year, and there’s a glitch, and you are going to have to reprocess this field if you want to use it.

Successful systems for sharing data succeed not only because of standardization, but also on account of the information they make public about data. Berg (1997) noted that such discussions have a spillover effect because they can help to enhance understandings *among* disciplines.

Ecologists have been working to establish metadata protocols to standardize written information about data sets to make possible data sharing and reuse (Michener, Brunt, Helly, Kirchner, and Stafford 1997; Jones, Berkley, Bojilova, and Schildhauer 2001). Michener and his colleagues proposed a metadata scheme to capture such instructions and documentation. At the same time, they acknowledged that there is no end to metadata: “There is no unique, minimal, and sufficient set of metadata for any given data set, since sufficiency depends on the use(s) to which the data are put” (1997, 335). Birnholtz and Bietz (2003) found this to be the case in their study of data from three scientific disciplines. Their interviews with scientists in one project revealed that even a metadata model with 297 fields for each data point was not sufficient to fully understand what took place in an experiment. Birnholtz and Bietz observed that “...data are not simple carriers of meaning. [C]onverting raw data into scientific or social meaning is an active, context-dependent process” (2003, 341; See also Bowker and Star 1999.) My findings show that the knowledge ecologists gain through the collection of particular types of data is a key part of this active, context-dependent process. In the face of growing pressure to work with available data, ecologists must think carefully about how to sustain career experiences in the

field. Secondary analysis of data may become an ecological specialty driven by access to databases, by new theories, and by efforts to encourage the use of existing data. Yet, the reuse of data without a feeling for the ecosystem could be futile and misguided.

Instead of emphasizing *replication*, it might be more valuable for educators to train scientists to describe their methods so that others who might use their data can *visualize* the data collection process. This is a subtle, but important distinction. When using data collected by others, ecologists depend on information about data that enables them to put their field-based knowledge into play. My results, like those of Cambrosio and Keating (1988) show that scientists are able to articulate many aspects of their informal knowledge, and some of this knowledge is amenable to explication. Training new scientists to include field-based insights in their documentation of data could go a long way to improve the longevity of their data.

At this time, ecologists demand close scrutiny of each data point that they reuse. One transformation for ecology might be said to have occurred when ecologists no longer feel the need to analyze data at such a fine scale. As Allchin (1999) noted, a strategy of “direct checking is not very economical.” For one, this will no longer be possible as datasets increase in size due to the implementation of new technologies, such as sensor networks, that generate enormous amounts of data. Armed with a better understanding of how ecologists currently use data they did not collect themselves, we can be attuned to new approaches they might develop to understand and ensure the integrity of data they reuse as they increase the scales at which they work.

References

- Alder, K. 2002. *The measure of all things: The seven-year odyssey and hidden error that transformed the world*. New York: The Free Press.
- Allchin, D. 1999. Do we see through a social microscope?: Credibility as a vicarious selector. *Philosophy of Science* 66 (Supplement): S287-S289.
- Beaulieu, A. 2001. Voxels in the brain: Neuroscience, informatics and changing notions of objectivity. *Social Studies of Science* 31 (5): 635-680.
- Berg, M. 1997. Problems and promises of the protocol. *Social Science & Medicine* 44 (8): 1081-1088.
- Birnholtz, J., and M. Bietz. 2003. Data at work: Supporting sharing in science and engineering. *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work*: 339-343.
- Bowker, G. C. 2000. Biodiversity datadiversity. *Social Studies of Science* 30 (5): 643-683.
- Bowker, G. C., and S. L. Star. 1999. *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press.
- Bowser, C. J. 1986. Historic data sets: Lessons from the past, lessons for the future. In *Research data management in the ecological sciences*, edited by W. K. Michener, 155-179. Columbia, SC: University of South Carolina Press.
- Brown, C. 2003. The changing face of scientific discourse: Analysis of genomic and proteomic database usage and acceptance. *Journal of the American Society for Information Science and Technology* 54 (10): 926-938.
- Cambrosio, A., and P. Keating. 1988. "Going monoclonal": Art, science, and magic in the day-to-day use of hybridoma technology. *Social Problems* 35 (3): 244-260.
- Campbell, E. G., B. R. Clarridge, M. Gokhale, L. Birenbaum, S. Hilgartner, N. A. Holtzman, and D. Blumenthal. 2002. Data withholding in academic genetics: Evidence from a national survey. *JAMA* 287 (4): 473-480.
- Collins, H. M. [1985] 1992. *Changing order: Replication and induction in scientific practice*. Chicago, IL: University of Chicago Press.

- Collins, H. M., and T. Pinch. 1998. The sex life of the whiptail lizard. In *The golem: What you should know about science, 2nd ed.*, 109-119. Cambridge, UK: Cambridge University Press.
- Committee on the Future of Long-term Ecological Data. 1995. *Final report of the Ecological Society of America Committee on the Future of Long-Term Ecological Data (FLED) (Vols. 1-2)*. Washington, DC: Ecological Society of America.
- Edwards, P. N. 2004. "A vast machine": Standards as a social technology. *Science* 304 (5672): 827-828.
- Fricke, E. 2002. Trusting others in the sciences: *A priori* or empirical warrant? *Studies in History and Philosophy of Science* 33 (2): 373-383.
- Hilgartner, S. 1995. Biomolecular databases: New communication regimes for biology? *Science Communication* 17 (2): 240-263.
- Hilgartner, S., and S. I. Brandt-Rauf. 1994. Data access, ownership, and control: Toward empirical studies of access practices. *Knowledge: Creation, Diffusion, Utilization* 15 (4): 355-372.
- Hine, C. 2006. Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science* 36 (2): 269-298.
- Jones, M., C. Berkley, J. Bojilova, and M. Schildhauer. 2001. Managing scientific metadata. *IEEE Internet Computing* 5 (5): 59-68.
- King, J. L., and R. L. Frost. 2002. Managing distance over time: The evolution of technologies of dis/ambiguation. In *Distributed work*, edited by P. Hinds and S. Kiesler, 3-26. Cambridge, MA: MIT Press.
- Knorr Cetina, K. 1999. *Epistemic cultures: How the sciences make knowledge*. Cambridge, MA: Harvard University Press.
- Kohler, R. E. 2002. *Landscape and labscapes: Exploring the lab-field border in biology*. Chicago, IL: University of Chicago Press.
- Latour, B. 1999. Circulating reference: Sampling the soil in the Amazon forest. In *Pandora's hope: Essays on the reality of science studies*, 24-79. Cambridge, MA: Harvard University Press.
- Latour, B., and S. Woolgar. 1979. *Laboratory life: The social construction of scientific facts*. Beverly Hills, CA: Sage Publications.

- Mallard, A. 1998. Compare, standardize and settle agreement: On some usual metrological problems. *Social Studies of Science* 28 (4): 571-601.
- Markus, M. L. 2001. Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success. *Journal of Management Information Systems* 18 (1): 57-93.
- McCain, K. W. 1991. Communication, competition, and secrecy: The production and dissemination of research-related information in genetics. *Science, Technology, & Human Values* 16 (4): 491-516.
- Michener, W. K. 2000. Transforming data into information and knowledge. In *Ecological data: Design, management and processing*, edited by W. K. Michener and J. W. Brunt, 142-161. Oxford: Blackwell Science.
- Michener, W., J. Brunt, J. Helly, T. Kirchner, and S. Stafford. 1997. Nongeospatial metadata for the ecological sciences. *Ecological Applications* 7 (1): 330-342.
- National Research Council 2003. *NEON: Addressing the nation's environmental challenges*. Washington, DC: National Academy Press.
- National Research Council. 1997. *Bits of power: Issues in global access to scientific data*. Washington, DC: National Academy Press.
- National Research Council. 1995. *Finding the forest in the trees: The challenge of combining diverse environmental data: Selected case studies*. Washington, DC: National Academy Press.
- Oromaner, M. 1977. Professional age and the reception of sociological publications: A test of the Zuckerman-Merton hypothesis. *Social Studies of Science* 7 (3): 381-388.
- Porter, T. M. 1995. *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, NJ: Princeton University Press.
- Roth, W-M., and G. M. Bowen. 2001a. 'Creative solutions' and 'fibbing results': Enculturation in field ecology. *Social Studies of Science* 31 (4): 533-556.
- Roth, W-M., and G. M. Bowen. 2001b. Of disciplined minds and disciplined bodies: On becoming an ecologist. *Qualitative Sociology* 24 (4): 459-481.
- Senkowsky, S. 2005. Planning of NEON moves ahead. *BioScience* 55 (2): 106-12.
- Star, S. L. 1985. Scientific work and uncertainty. *Social Studies of Science* 15 (3): 391-426.

Weltzin J. F., R. T. Belote, L. T. Williams, J. K. Keller, and E. C. Engel. 2006. Authorship in ecology: Attribution, accountability, and responsibility. *Frontiers in Ecology and the Environment* 4 (8): 435-441.

Wenger, E. 1998. *Communities of practice: Learning, meaning, and identity*. Cambridge, UK: Cambridge University Press

Zimmerman, A. 2007. Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries* 7 (1-2), 5-16.

AUTHOR'S NOTE: I gratefully acknowledge the ecologists who so willingly shared their experiences with me. For their comments on the manuscript, I thank Margaret Hedstrom, Robert L. Frost, Edward J. Hackett, and three anonymous reviewers.

Ann Zimmerman is Research Assistant Professor in the School of Information at the University of Michigan. Her research interests include scientific collaboration, the secondary use of scientific data, and the relationships between large-scale collaborations, policy, and research management.

Notes

¹ I also use the phrase *secondary use*, which I intend to be synonymous with the term *reuse*.

² The current convention in ecology is that the author listed first is the one who made the greatest contribution to the publication (Weltzin, Belote, Williams, Keller and Engel 2006).

³ All names are pseudonyms.

⁴ Sometimes colleagues pointed ecologists to individuals who might have data they needed. Often, however, they gained clues about data people may have collected by reading publications they had written.

⁵ Meta-analysis is a quantitative statistical tool used to combine and compare the outcomes of different research studies, often experiments, in order to achieve a larger effect in size (Michener 2000).

⁶ Even those ecologists who performed laboratory analyses possessed substantial field experience. In order to analyze plant nitrogen content, for example, one has to gather specimens from the field.