DATA SHARING AND SECONDARY USE OF SCIENTIFIC DATA: EXPERIENCES OF ECOLOGISTS

by

Ann S. Zimmerman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information and Library Studies)
in The University of Michigan
2003

Doctoral Committee:

Associate Professor Margaret Hedstrom, Chair
Assistant Professor Brian Athey
Associate Professor Paul N. Edwards
Professor Jeffrey K. MacKie Mason

ACKNOWLEDGMENTS

Many people contributed to the completion of my dissertation, and it gives me great pleasure to acknowledge the numerous forms of assistance I received.

Faculty and staff of the University of Michigan (UM) School of Information (SI) provided unlimited support, encouragement, and guidance throughout my studies. Initial thanks go to my committee members. My chair, Associate Professor Margaret Hedstrom, mentored me through each stage of the program. Her wisdom, her confidence in me, and her ability to offer comments of just the right kind at exactly the right time helped me to sharpen my thinking, to improve my writing, and to achieve more than I thought possible. Associate Professor Paul N. Edwards introduced me to the field of science studies and offered advice that enhanced the presentation of my research. Professor Jeffrey K. MacKie Mason and Assistant Professor Brian Athey accepted a methodological approach different from their own and offered comments that substantially improved my dissertation. Others at SI also assisted me. Associate Dean and Professor Gary Olson played a substantial role in making SI a wonderful place to pursue a doctoral education. Professor Judy Olson provided support at a critical time in my research. Associate Dean and Professor C. Olivia Frost met with me before I applied to the program and encouraged me throughout my studies. Professor George Furnas and Assistant Professors David Wallace and Beth Yakel showed an interest in my progress that helped to keep me

focused. Finally, I thank Sue Schuon, Doctoral Program Coordinator, who offered much sound advice and assisted me in negotiating myriad details.

This dissertation grew out of my position as Librarian at the U.S. Geological Survey's Great Lakes Science Center (GLSC), and I extend thanks to my colleagues and friends there. David Walsh, retired Assistant Director, raised the idea of my return to school, made it possible for me to pursue it, and encouraged me from beginning to end. My supervisors over the years--Anthony Frank, Dr. John Gannon, and Dr. Douglas Wilcox--were unfailing in the generosity that allowed me to combine job and school. I could not have completed this dissertation without their support. Among the many current and former GLSC scientists who discussed data sharing issues with me, encouraged my interests, and provided moral support, I extend special thanks to the following individuals: Dr. Mary Burnham Curtis, Gary Curtis, Carol Edsall, Dr. Mary Fabrizio, Patrick Hudson, Dr. Charles Madenjian, Dr. Bruce Manny, Dr. Jacqueline Savino, and Dr. Jeffrey Schaeffer. Scott Nelson, Computer Specialist, generously shared with me his wealth of knowledge regarding data management. My sincere gratitude also goes to Christine Schmuckal, Tatiana Schwartz, and Tracy Myers-Janevic who worked with me in the GLSC's John Van Oosten Library and whose skills and dedication kept operations and services running smoothly during my absences.

The cooperation and participation of the individuals I interviewed made this research possible, and I thank all of them for sharing their experiences. The knowledge and dedication of each person impressed me deeply and gives me hope that we will have the data, programs, and policies needed to achieve a better understanding of our environment and to deal with the threats that confront it.

I gratefully acknowledge the financial support I received. The GLSC supported a portion of my tuition and research costs, and SI provided funds for transcription and travel and awarded me the Virginia Ehrlicher Scholarship. The UM Rackham School of Graduate Studies One Term Dissertation Fellowship gave me time to write, and National Science Foundation Award #0214690 helped me to complete my dissertation.

Other doctoral students, especially Denise Anthony, Bill Aylesworth, and Deb Torres offered support that only those sharing the same experience can provide. Drs. Qiping Zhang and Bradley L. Taylor, recent graduates, set examples by their achievements. Brad Taylor's friendship, which began on the first day we entered the doctoral program, sustained me throughout and is a lasting benefit of my time at SI.

Several others contributed to my success in unique and important ways. Bonnie Francis transcribed my interview tapes with enthusiasm and skill. Ellen Cardwell, Meetings Manager, found me a room in which to conduct interviews at the 2001 ESA Meeting. Dr. Victoria Serbia provided expert counsel, and Dr. Simone Taylor inspired me by her own example. Finally, I offer special thanks to Nina Rush for all her help.

My parents encouraged my education throughout my life. I thank them for their foresight, for their support in all my endeavors, and for their love. Regrettably, my father died several months before I began my doctoral studies, but I know he especially would have been proud of this achievement. My mother encouraged me enough for two and showed unlimited faith in my ability to succeed.

My deepest gratitude is reserved for my husband, Tom, who lived with this project as long as I did, but never tired of it. As in everything I do, he supported me unconditionally and offered whatever help was necessary for my success.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

CHAPTER 1

DATA SHARING AND SECONDARY DATA USE: AN INTRODUCTION AND RESEARCH PROPOSAL

Satellites circle the earth collecting terabytes of information on the planet's atmosphere each year (National Research Council [NRC], 1995b, p. 23). The widely celebrated completion of the Human Genome Project, a map of our species' genetic code, resulted in a database that could fill more than 2,000 computer diskettes (Howard, 2000). Federal agencies spend billions of dollars each year to collect demographic, economic, environmental, health, and other data (NRC, 1999, p. 25). Academic institutions, private corporations, and nonprofit organizations conduct scientific research of all kinds. The data gathered are used to fulfill organizational missions, to compete in the business world, to advance national and international goals, to protect national security, and to contribute to decision and policy making. Recently, there has been a push to use the raw data gathered for one purpose to answer new and different sets of questions. The benefits of such secondary use are believed to be substantial.

> Prior to its public dissemination, the use of a database is limited to those involved in the collection of data or production, and therefore does not provide the opportunity to contribute broadly to the advancement of scientific knowledge, technical progress, economic growth, or other applications beyond those of the immediate group. It is only upon the distribution of a database that its far-reaching research, educational, and other socioeconomic values are recognized. One or more researchers applying varying hypotheses, manipulating the data in different ways, or combining elements from disparate databases may produce a diversity of data and information products. The contribution of any of these products to scientific and technical knowledge might well assume a value far greater than the costs of database production and dissemination (NRC, 1999, p. 34).

Sentiments such as these have led to policies and laws that favor data sharing and to information initiatives to make this sharing easier. Scientific journals are promoting, and in some cases mandating, that the data on which published articles are based be made available (McCain, 1995; Murphy, 1990; Sieber & Trumbo, 1995). Substantial investments have been made to document and preserve data and much effort has gone into the development of metadata standards to describe data sets. These activities have gained momentum as information technology has made it easier to collect, manage, and store research data (Sterling, 1988). The demands to share data have also increased in response to a push for interdisciplinary research. It is difficult to pinpoint where the focus on interdisciplinary research first arose, but it has been spurred by the belief that the solutions to today's complex, global problems are outside the realm of any one discipline to solve. There is a notion that data sharing must be a part of such research.

In spite of all this activity, very little empirical research has been conducted to evaluate the costs, benefits, and outcomes of data sharing or to test the assumptions made about its effects on the content, conduct, and communication of science. There are numerous cultural, legal, political, and technical obstacles that must be overcome in order for widespread data sharing to occur (Sterling, 1988). For example, it is difficult and expensive to organize, document, and maintain data so they can be used by others. The scientific reward system currently provides few incentives for scientists to share data, and in fact, competition to be the first to make a discovery discourages sharing (Sieber & Trumbo, 1995; Sterling, 1988). Some data have significant economic value, such as genetic data used to develop new drug therapies, and sharing these data can jeopardize a business's competitive advantage (Howard, 2000).

In this study, I address the lack of systematic research related to data sharing. Specifically, I analyze the experiences of secondary data users in order to provide information to test some of the assumptions made about data reuse and to evaluate mechanisms developed for sharing data. Through the use of in-depth interviews, I

describe the experiences of ecologists who used data they did not collect themselves. I chose ecologists because research directed toward environmental problem solving is one area where data sharing has been strongly encouraged and because ecological data have characteristics that make them difficult to share. The results of my investigation will refine our knowledge about the effects of data sharing on the practice, content, and communication of science. In addition, the information gained from my study can be used to characterize the needs of different types of users, to construct better directories for locating data, to design interfaces to data collections, to allocate limited resources to data with the greatest potential for reuse, and to formulate policy. My results will also be applicable to the management of other information resources that share characteristics similar to scientific data.

In the remainder of this chapter, I describe the demands on scientists, particularly ecologists, to share data. I also discuss the legal and policy framework that, for the most part, encourages data sharing; the existing and developing mechanisms in place to make this activity easier; the known costs that may be disincentives to data sharing; and the benefits that are believed to result. Throughout this chapter, it will become clear that although much energy has been expended to encourage data sharing, no one knows for sure how well the existing mechanisms work, or if the benefits and outcomes of secondary use match the assumptions that have been made about them. My study begins to address these questions.

Key Concepts and Definitions

In order to discuss the issues surrounding data sharing and reuse, it is necessary to introduce a few key definitions and concepts and to describe how they will be used in this study. These include *data, data sharing,* and *secondary use of data*.

Data

First of all, my study focuses on scientific *data*, defined as "scientific or technical measurements, values calculated therefrom, 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 or further calculation" (NRC, 1997, p. 198). This basic definition is only one way to describe the many facets of scientific data.

> There are several ways to characterize data: among others, by form, whether numerical, symbolic, still image, animation, or some other; by the way they were generated or gathered, that is from experiment, observation, or simulation; by level of quality; by the size or form of the databases that contain them; by the nature of support for their generation or distribution, that is, public or private, national or international; and, of course, by subject (NRC, 1997, p. 49).

There is much to learn about the secondary use of all types of scientific data. However, as I show throughout this chapter, there is an especially pressing need to learn about the sharing and use of small, observational, numeric data sets, such as those common to ecology. In addition, there are ways that learning about these data can help us to understand problems of data sharing generally.

Observational data result from observations of the natural world. Observational data are considered particularly important to preserve because they are a record of events that will not be repeated, and once the records are lost, they cannot be replaced. They can provide "a baseline for determining rates of change and for computing the frequency of occurrence of unusual events" (NRC, 1995b, p. 1). In terms of secondary use, observational data serve as fodder for new concepts that "may emerge--in the same or entirely different disciplines--from study of observations that led earlier to different kinds of insights" (NRC, 1995b, p. 1). Observational data sets may be very large, such as those from the field of space physics, which has generated more than 50 gigabytes of data

annually, or small enough to be stored and manipulated on a single personal computer (NRC, 1995b).

Numeric data are collections of data that are primarily numeric in nature. Specifically, "numeric data consist of a numeric value and one or more attributes of information about the numeric value," such as units of measurement and uncertainty and validity (Luedke, Kovacs, & Fried, 1977, p. 120). Numeric data exist in print and digital forms. In ecology, for example, it is not unusual to find raw data stored on original field data collection sheets (Committee on the Future of Long-Term Ecological Data [FLED], 1995, n.p.). The process of converting analog data to digital form so a computer can manipulate them is wrought with potential problems. The challenges confronted include the gap between when data were collected and when they are entered into a computer, sparse documentation of data, indecipherable handwriting, and inconsistent formats for field sheets through time. The quality of the resulting database is affected by all of these problems.

Data Sharing

Like *data*, the concept of *sharing* can be characterized in numerous ways. As one possibility, technological infrastructure issues might be emphasized. For data sharing, these could include matters related to computing and communications, specifications for defining common terms across different fields, or metadata standards to describe content and structure and to serve administrative functions (Dawes, 1991). Many of the current mechanisms developed for data sharing stress a technological approach. Alternately, social issues related to sharing might be highlighted. For example, sociologists have discussed incentives for sharing and analyzed reasons why scientists might not want to share data. By definition, sharing "implies that one as the original holder grants to

another the partial use, enjoyment, or possession of a thing" (Mish, 1993). This meaning is reflected in Robert Boruch's (1985) definition in which he described *data sharing* as "the voluntary provision of information from one individual or institution to another for purposes of legitimate scientific research" (p. 89). Finally, sharing can be characterized by the activity level at which it occurs. Authors who make documents available via their Internet home page are sharing that information even though they do not directly mediate or respond to each request for their work. In other situations, an author may share a work in progress, but only after querying the requester about the intended use. Some authors distinguish between the passive to active levels at which information can be shared (McClure, 1989; Sprehe, 1999). These writers note the important differences between production, access, disclosure, and dissemination. Sprehe (1999) observed that users assume various roles depending on the level at which the information is made available. Dissemination, defined as taking positive steps to place information in a user's hands, is the most active mode of information sharing, although it results in a passive role on the user's part (Sprehe, 1999). Access, on the other hand, demands an active role by the requester while the information holder is passive and waits for a request to be received.

For the purpose of my study, *data sharing* is defined broadly. The definition encompasses the technological and social aspects of sharing as well as the activity level at which it occurs. Included is a range of possibilities from one-on-one informal interactions between the holder and receiver of data to active dissemination or publication via formal mechanisms. There are times when distinguishing between the activity levels of sharing are important. One of the goals of my study, however, is to better understand the variety of ways in which data are shared, the tactics that individuals employ to find

data, and the influence of the data sharing mechanism on the secondary user's experience. Thus, all types of exchange are important. The main restriction placed on sharing in my study pertains to the ultimate use of the data. I investigate the use of shared data for research purposes. Although uses outside research are important, such as those related to policy formulation and decision-making, they are not analyzed in my project.

Secondary Use of Data

Sharing is coupled with *secondary use* because in order for data to be used by others, they must be accessible (Fienberg, Straf, & Martin, 1985). In this study, I define s*econdary use* as the use of data collected for one purpose to study a new problem. My definition includes data originally gathered to address a specific research question as well as data gathered to describe physical or biological phenomena (NRC, 1999, p. 4). In this investigation, I focus on the secondary use of data by ecologists. I use the term *reuse* in relation to data as a synonym for *secondary use*.

In a report on scientific and technical databases, a committee of the National Research Council (1999) described several ways that databases are used. The committee defined *end use* as "accessing a database to verify some fact or perform some job-related or personal task…" (p. 34). End use is not considered *secondary use* for the purposes of this study. *Derivative use* builds on a preexisting database by extracting information from one or more databases to create a new database that can be used for the same, similar, or entirely different purpose as the original database. Derivative use and other uses that combine or compare data to answer new questions are included in my definition of *secondary use*. For the purpose of my study, data do not have to be in a digital form.

A number of activities may precede or follow the provision of data, such as organizing and documenting data, educating others to use them, and making data available in forms that are manipulatable by different software programs. These activities are presumed to be important to the successful reuse of data. In particular, much effort has gone into the development of standards to make data sharing easier. The role of these activities, and their value to users, are also an important topic in my study.

Demands for Data

An investigation into the sharing and secondary use of scientific data is of little value if those with the most interest in it pay no attention to the topic. This is not the case, however, as the subject has become the focus of much activity, discussion, and interest. The demands for scientific data arise primarily from two areas. One demand comes from the scientific questions that researchers attempt to answer (Hesse, Sproull, Kiesler, & Walsh, 1993; Michener & Brunt, 2000). In some fields, secondary data use is the norm. For other disciplines, it is only recently that the research questions posed required data from one or more outside sources. The second type of demand for scientific data is comprised of a broad range of social influences. Although the two forces--scientific needs and social demands--are listed separately, they are often intertwined.

Demands Driven by Science

As mentioned above, one of the driving influences behind data sharing arises from the nature of the science being conducted, including the questions that researchers attempt to answer. Hesse, Sproull, Kiesler, and Walsh (1993) pointed out that "in all scientific disciplines, the phenomena that scientists study influence how their work is organized and carried out" (p. 92). For example, research in physical oceanography is

conducted using large research vessels carrying expensive data collection equipment. Data are gathered from remote locations, which requires coordination across long distances. Collaboration and data sharing are necessary to conduct physical oceanographic research since no individual and few institutions can afford to carry it out on their own (Hesse et al., 1993). The processes under study also dictate the data needs. Physical oceanographers, for example, require access to large databases on surface currents, salinity, and prevailing wind to study the fluxes in the world's oceans (Hesse et al., 1993). These databases must be available to physical oceanographers if they are to investigate the field's research questions.

In recent years, scientific disciplines related to environmental problem solving have been a special target of data sharing efforts. For example, funding arrangements for large, interdisciplinary research projects to study complex issues like global climate change often include arrangements and requirements for data sharing (NRC, 1995a). International bodies, such as Unesco, play an important role in data standardization for multinational environmental projects (Weingart, 1997).

Ecology as an Example

Ecology, the study of interrelationships between the earth's organisms and their environment, is one of the many disciplines that contribute to our knowledge of the natural world. In comparison to some fields, ecologists have little experience with data sharing and that which does occur is usually between close associates (FLED, 1995, n.p.). In fact, it has been said that ecology is one of the few scientific fields without coordinated efforts to share and preserve data: "Ecology and evolutionary biology stand virtually alone among the environmental and environment-related sciences in the lack of some agency- or community-mandated data archiving and data sharing policy" (Porter & Callahan, 1994, p. 195).

The causes for this are complex. Besides the nature of the questions that ecologists ask, the reasons include the discipline's culture and the characteristics of its data. There are indications, however, that the sharing of ecological data is becoming increasingly important. This change is influenced by scientific and social needs, which makes ecology an excellent example of the interrelatedness of the two types of demand. There is recognition by ecologists, driven by funding opportunities, that ecology must become a "problem-solving discipline" (Baskin, 1997, p. 310). This has influenced some ecologists to begin asking new questions and to look for ways to expand the scale of their science. The complexity of environmental problems, pressure from funding sources, and encouragement from various sectors are changing the research questions that some ecologists pose (Ben-Ari, 1998). Many of these topics require the combination of data from more than one ecological study or the interfacing of data from ecology with data from other fields (NRC, 1995a).

In 1991, the Ecological Society of America (ESA), the discipline's main professional society, convened a committee to articulate an ecological research agenda for the remainder of the decade. In the final report of their work, the Committee for a Research Agenda for the 1990's proposed a "Sustainable Biosphere Initiative" and outlined a research plan focused on three areas: biodiversity, global climate change, and ecological sustainability (Lubchenco et al., 1991). The three topics were selected because they are issues of great human concern and because research in these areas could contribute to fundamental ecological knowledge, which in turn could be used to work toward solutions to environmental problems. The ESA noted that the work of the Committee and the document they produced was unprecedented in its attempt to

formulate a research agenda for ecology. The report's content also shows the roots of what has become a campaign among some ecologists to increase the scope and scale at which their science is conducted. For example, the Committee noted that preserving biological diversity "requires a better understanding of how ecological processes operating on different spatial and temporal scales interact" (Lubchenco et al., 1991, p. 389).

Since the publication of this report, the ESA has tried to effect some of the changes in the practice of ecology that would contribute to the objectives identified in the Sustainable Biosphere Initiative report. The ESA actively supported the establishment of the National Center for Ecological Analysis and Synthesis (NCEAS), which was funded initially in 1995 for 5 years through a National Science Foundation (NSF) grant. In 2000, the NSF renewed NCEAS's funding for six more years.[1] NCEAS's purpose is to look at big questions in ecology without gathering any new data. Ecologists who believed it important to "scale-up" both the temporal and spatial ranges of their discipline's research spurred the creation of NCEAS. The "scaling up" notion was fueled partly by the findings of two separate literature surveys. In one survey, ecologists Peter Kareiva and M. Anderson (1988) searched the journal, Ecology, for all experimental ecology papers published from January 1980 to August 1986. They located 97 papers that satisfied their definition of experimental ecology, and for each of these papers they recorded the maximum linear dimension of experimental plots used in the research and the maximum number of replicates of any treatment. Kareiva and Anderson found that most studies were limited in physical scale and in the number of replications.

> Over 45% of the papers we looked at included at least one treatment that was replicated no more than twice. Nearly one-quarter of the studies used

> plots no larger than .25m in diameter; one-half of the studies used plots no larger than a meter in diameter. One has to wonder whether studies conducted at such a small scale are not missing key aspects of species interactions (p. 37).

The authors acknowledged that this situation exists because replications and scale require money and personnel that are beyond the resources of most ecologists. To address this problem, they recommended the use of theory to build models that "address dynamics at the scale of ten or hundreds of meters," and that with a "model firmly in mind, a single experiment can provide a powerful test of one's understanding of spatially-distributed interactions" (p. 37). As Kareiva and Anderson noted, however, and as others have also observed, ecology lacks solid theory (cf., FLED, 1995; Kwa, 1993; Roth & Bowen, 1999, p. 722; Slobodkin, 1988; Worster, 1994, p. 373). This situation adds to the challenges that ecologists confront in expanding the scope of their science.

In another survey of the literature, Tilman (1989) showed that 40% of ecological experiments lasted less than a year and only 7% lasted five or more years. The findings from the two surveys alarmed many ecologists:

> Such findings put numbers to a feeling shared by many influential ecologists at the time: Data from thousands of small studies on everything from predator-prey cycles to soil nitrogen levels were piling up, but too few ecologists were looking at big-picture questions such as how ecosystems respond to disturbances over time, or why some regions are more species-rich than others (Baskin, 1997, p. 310).

As illustrated above, the limited temporal and spatial scales at which ecology traditionally operates are believed by some ecologists to be inadequate to address today's critical environmental problems (Macilwain, 2000; Michener & Brunt, 2000). It is also true, however, that the practical aspects of conducting research in a particular scientific discipline limit the questions that can be asked. This is often true in ecology because field studies are labor intensive, which constrains the size of the physical area that can be

studied (NRC, 1995a, p. 84). The combination of data from multiple studies is seen as one way to address this limitation.

Cultural and Social Demands for Data

The scientific needs that lead to data sharing in some fields and not in others can be overstated, however, since data sharing is also heavily influenced by cultural and social factors, such as the ownership level of data (Sterling, 1988; Sterling & Weinkam, 1990). Joan Sieber and Bruce Trumbo (1995) pointed out that data sharing is not new, as "some government and academic archives have, for decades, made available to individual scientists massive sets of geophysical, demographic, attitudinal, health and economic data"; what is more recent is the demand for individual scientists to share data (p. 11). Scientific data that are gathered via satellite or remote instruments are not only too costly for a single private entity to support, they are also less likely to be "owned" by individuals who depend on the analysis and publication of the data to advance their careers. The latter is the case in ecology, which is characterized by single-investigator studies, and where there is a tendency for data sets to belong to the scientist who collected them (NRC, 1995b, p. 52). In addition, central files of organized, uneditable data are less likely to cause ownership disputes (Sterling & Weinkam, 1990). Research studies carried out by a single researcher gathering data in situ based on individual measurements have stronger ownership ties (NRC, 1995a, p. 16). Sterling and Weinkam (1990) saw the outlook for the sharing of this type of data as slim.

> Scientist-to-scientist cooperation depends on individual arrangements. On the whole, however, the prospects for that particular cooperation to flourish are dim. Economic motives, motives of personal power, possible disagreement, prospects of conflict, likely detection of bias or fraud...all combine to discourage data sharing (p. 119).

Demand for data among scientists who share similar disciplinary backgrounds is influenced by the cultural norms that exist in that field about when it is appropriate to

request data, when it is acceptable to deny access to data, and how to acknowledge the provider of data (Louis, Jones, & Campbell, 2002; McCain, 1991). When data requests come from researchers in other fields, different cultural norms and expectations can complicate sharing (Hilgartner, 1997; Van House, Butler, & Schiff, 1998). It is also possible that non-scientists will be interested in access to data, although we currently know little about the needs or requirements of such users (Neuhold, 1998).

Besides the cultural norms among scientists, there are other social forces, both within and outside the scientific community, that affect data sharing efforts. These influences come from several groups, including policy makers, an array of individuals and organizations with interest in data, the larger public, and from pressure within the research community itself (Sieber, 1988; Stanley & Stanley, 1988). The demands are shaped to some degree by perceptions regarding the appropriate role of scientists in society.

The traditional view of science is one of a group that sets and solves its own problems governed by the largely academic interests of a particular community and that determines its own criteria for judging quality (Gibbons, Limoges, Nowotny, Schwartzman, & Trow, 1994). This idea is consistent with the notion of "basic research." Some writers argue that the trend toward a more socially accountable and reflexive science effects the production of knowledge (Funtowicz & Ravetz, 1993; Gibbons et al., 1994). A change in knowledge creation would have ramifications in many areas, including the conduct and communication of science, judgments about scientific quality, the organization of institutions, and the formulation of policy. Others believe that the distinction between basic and applied research is diminishing and that what appears to be a significant change is merely part of the continuing ebb and flow in negotiations over the proper role of scientists in society (Godin, 1998). Donald Stokes (1997) argued that the prevailing view of a distinct line dividing fundamental understanding and use-based research provides an incomplete account of the relationship

between basic research and technological innovation (p. 89). He believed that this notion hindered dialogue between the scientific and policy communities and impeded the search for a new compact between science and society. Whatever their stance regarding changes to the production of scientific knowledge, most writers agree that there is increased pressure on scientists in certain fields to apply themselves to important human problems, and that such demands have consequences for the organization, content, and communication of research in these areas (Godin, 1998; Steele & Stier, 2000; Weingart, 1997). This is particularly true in fields related to the environment, the focus of this study, and to the areas of human health, privacy, and communication.

An increased demand, at least in some areas, for scientists to apply themselves to important human problems has led to a belief that many of today's complex problems require solutions that are beyond the scope of any one discipline to master. This conviction has resulted in the promotion of interdisciplinary research projects, which are believed to address larger problems than any discipline can undertake on its own. The successful translation of data and information across fields is a key component of interdisciplinary research. In spite of the pressures on scientists to conduct this type of research, little is known about its effectiveness. Steele and Stier (2000) noted that the benefits assigned to interdisciplinary research are based largely on faith and a few positive anecdotes. Since the secondary use of scientific data is an important aspect of interdisciplinary research, a study about the former can increase our knowledge of the latter.

A report by the National Science Board (2000) summed up the current thinking regarding the importance of data access to the conduct of interdisciplinary research, especially in addressing environmental problems.

> The growing frustration with the lack of adequate scientific information about environmental issues has led to a plethora of reports and suggestions. The majority of these focus on enhancing the disciplinary and interdisciplinary fundamental understanding of environmental systems and problems, improving

> the systematic acquisition of data, the analysis and synthesis of these data into useful information, and the dissemination of this information into understandable formats for multiple uses (p. 41).

This statement also illustrates the influence of funding sources on research agendas. For example, federal agencies that provide research grants, such as the National Science Foundation, increasingly require that the data collected by research projects be made available at the end of the study (Gershon, 2000; Palmer, 1996; Siang, 2002). Current U.S. policies and laws, which I discuss in more detail later, also require most federal agencies to make available data that they gather as part of their missions. Policies, law, and funding drive demand and also serve as strong mandates for scientists to provide access to their data.

Besides funding sources, scientists' professional and social organizations can encourage or discourage their behavior. Journal editors, for one, play a significant role in directing scientific norms and standards (McCain, 1995). For example, in the field of genetics data deposition is a frequent criterion for publication (Hilgartner, 1995; Howard, 2000; McCain, 1991). Professional societies, many of which are publishers, can also shape the culture of their disciplines. Since 1996, the American Chemical Society, the leading society in its field, has made available in electronic form supplementary material, including data sets, from one of its key journals (Glaze, 1996). The Ecological Society of America has promoted the preservation of ecological data and encouraged increased interaction between ecologists and researchers from other disciplines. For instance, in 1995 the society organized a symposium and invited participants from a variety of social science disciplines to explore their linkages with ecology to "begin to illustrate the means, barriers, and opportunities for promoting interdisciplinarity" (Haeuber & Ringold,

1998, p. 330). The ESA also started an electronic archive for authors interested in sharing data and other supplementary material associated with articles published in ESA journals.[2]

A scientific discipline that wants to increase data sharing must do more than recognize the value of this activity. Although recognition is an important step, a field does not change its practice based on the beliefs of some of its members. Data sharing is costly and scientists are currently not rewarded for it. Several other factors also discourage scientists from exchanging data. Security issues, the financial costs of duplication, the effort required to prepare data for sharing, uncertainty about the qualifications of data requesters, data set inadequacies, protection of graduate student or commercial interests, and ethical concerns are some of the deterrents (cf., Blumenthal, Campbell, Anderson, Causino, & Louis, 1997; Campbell et al., 2002; Ceci & Walker, 1983; Louis et al., 2002; Marshall, 2000; Stanley & Stanley, 1988). In addition, scientific reward structures may encourage secondary users to be overly negative in their evaluation of a data set (Ceci & Walker, 1983; Fienberg et al., 1985). Recognizing these issues, an ESA Committee studied scientific fields in which data sharing is practiced and identified common elements that led to successful data exchange mechanisms (FLED, 1995). Their report sums up the technical, scientific, and social demands that unite in some situations to make this activity work. The Committee found that success is achieved through a mixture of technical capabilities, such as free and easy software for data transfer, scientifically motivated needs, and socially influenced demands and incentives. Among the latter are those that emerge from the scientific community, such as leadership from key individuals and community acknowledgment of the importance of

data sharing. Social influences outside the immediate scientific community, such as support from key journals and external funding for data management are also important. Many of these elements have influenced views about the practice of ecology. As a result, the organization, preservation, and dissemination of data promise to be increasingly important issues for ecologists (Ingersoll, Seastedt, & Hartman, 1997).

Laws and Policies that Encourage Data Sharing

Arguably, the most motivating external social influences on data sharing are laws and policies. One reason for the increased demand for data is the development of a legal and policy framework that favors the open availability of scientific data gathered with federal funds. Historically, few U.S. laws and policies directly addressed access to or ownership of scientific data. This situation changed in recent years with the formulation of policy and legislation that generally favor unrestricted access to scientific data gathered with federal funds, with the exception of data related to national security or to the protection of personal privacy or confidentiality. These activities are predicated on the notion that wide availability to scientific data stimulate technical innovation and help solve problems, which in turn spur the economy and improve quality of life. However, the economic value of data has also led to more restrictive policies in the international arena and reduced the willingness of private enterprise to share data. In reality, the importance of data protection to economic growth, or the importance of an open policy to scientific advancement, is not the same for all scientific data (Hilgartner, 1997). In order to settle these arguments for the greatest benefit of society, further knowledge is needed about the contributions of data to science and the economy. In this section, I emphasize

U.S. laws and policies related to scientific data gathered with federal funds. I exclude discussion of complex intellectual property rights that are relevant to scientific data gathered by private industry, universities, and other non-governmental organizations in the United States and by foreign governments. These issues are important, but they are beyond the scope of my study.

Government policies related to the availability of scientific data gathered with federal funds have shifted over the years, but they currently reflect an active interest in seeing data collected to answer one set of research questions used to address other questions. For a long time, the federal government has invested in science for the public good. Its policies for achieving this goal as related to the dissemination of scientific and technical information gathered with federal funds have varied, however. In the 1980's, government granted commercial interests proprietary ownership in scientific and technical data, and agencies were actively discouraged from exploiting their information for public benefit (Sprehe, 1994; U.S. National Commission on Libraries and Information Science, 1984; Weil, 1988). These policies still exist to some degree in, for example, Cooperative Research and Development Agreements (CRADA), which provide commercial enterprise with sole use of government information (Reichman & Uhlir, 2001, p. 268). For the most part, though, recent policies have promoted the wider availability of scientific data, and government information generally, while trying to maintain the appropriate balance between public and private interests (McClure, Moen, & Bertot, 1999; Sprehe, 1999).

New policies, along with revisions to existing ones, vary in their specificity to research data, but overall, they promote greater access to all types of government

information, and they encourage the use of technology to achieve that access. The document that outlined the National Information Infrastructure (NII) (Information Infrastructure Task Force, 1993) was a key policy statement of the 1990's, and it set the tone for the formulation of new policies and revisions to earlier ones. The Clinton Administration established the NII as a broad policy effort intended to facilitate the nation's development of a network design and architecture to further its goals to provide improved access to government information (Bertot & McClure, 1996; Fletcher & Bertot, 1999).

The Freedom of Information Act (FOIA) and Circular A-130 are two key existing policies affecting access to government information. In the past, FOIA was not applicable to research data and Circular A-130 did not specifically mention scientific information. This situation changed in the 1990's.

Circular A-130 provides uniform government-wide information resources management (IRM) policies required by the Paperwork Reduction Act (PRA) of 1980. In 1995, the PRA was amended, resulting in substantial revisions to Circular A-130 (Office of Management and Budget [OMB], 1996). Although the Circular retained its focus on federal IRM, it also addressed agencies' responsibilities to disseminate and provide access to government information (Moen, 1996). Agency information was considered an asset that must be managed, and this required knowing what existed (McClure et al., 1999). OMB published guidelines for the electronic dissemination of information, and as part of this, each agency was mandated to "maintain an inventory of the agencies' major information systems, holdings, and information dissemination products…"(OMB, 1996, n.p.). This instruction led to the creation of locator services, such as the Government

Information Locator Service (GILS) (McClure et al., 1999).[3] GILS was intended to help the public find information about government products, including scientific data (Moen, 1996). Circular A-130 also referenced the importance of scientific information to foster "excellence in scientific research and effective use of Federal research and development funds" (OMB, 1996, n.p.). Although A-130 did not specifically mention scientific data, it set the tone for government agencies to make all types of information available electronically.

In late 2000, the PRA was augmented to include guidelines for the quality of information, including data, disseminated by Federal agencies (Reichhardt, 2002). Section 15 of the Treasury and General Government Appropriations Act for FY 2001 (Public Law 106-554) directed the OMB "to issue government-wide guidelines that provide policy and procedural guidance to Federal agencies for ensuring and maximizing the quality, objectivity, utility, and integrity of information, including statistical information, disseminated by Federal agencies" (OMB, 2002, p. 8458). OMB issued guidelines on an interim final basis on September 28, 2001. Due to concerns from scientific organizations over the requirement that data be "substantially reproducible," OMB requested additional comments and issued final guidelines early in 2002 (Cohn, 2001; OMB, 2002; Reichhardt, 2002). By October 1, 2002 each federal agency was instructed to have procedures in place for guaranteeing information quality and for addressing complaints. OMB's final guidelines promote reproducibility of results, but they also recognize that replication is not always practical or feasible. Decisions about reproducibility of data are determined by each agency, but OMB instructed that agency "guidelines need to achieve a high degree of transparency about data even when

reproducibility is not required" (OMB, 2002, p. 8456). OMB stated clearly that the guidelines do not apply to federal scientists or grantees who publish their findings under the academic model unless the agency uses the information to support an official position. Despite this exception, a wide variety of information disseminated by federal agencies, such as endangered species lists and research supporting water pollution regulations, will have to meet the new standards (Reichhardt, 2002). This law, along with other controversial policies, stimulated numerous discussions within the scientific community. For example, in late March 2002, the National Academy of Sciences (NAS) held the first of a series of workshops on the implications of the data quality guidelines (Reichhardt, 2002). The previous month, scientists and journal editors met at the NAS to attempt to develop straightforward rules to compel scientists to share data (Marshall, 2002).

Recent revisions to the Freedom of Information Act speak more directly to access to data gathered with federal funding. FOIA provides the public with access to the records of its government. Until recently, FOIA was not applicable to research data because they did not fit the definition of a federal record (Nelkin, 1984). However, Alabama Senator Richard Shelby sponsored a one-sentence amendment as part of the Omnibus Appropriations Act for FY1999, Public Law 105-277, which provided unprecedented public access to research data produced with federal funds. The amendment grew out of Senator Shelby's frustrating attempt to obtain data associated with a Harvard University study that connected air pollution and health effects, and which the Environmental Protection Agency (EPA) cited as part of its justification for tightening air pollution standards (McGinley, 1999; Shelby, 2000). Once passed by

Congress, this law mandated OMB to revise Circular A-110 to "ensure that all data produced under an award will be made available to the public through the procedures established under the Freedom of Information Act" (NRC, 2002, p. viii). The final guidelines, issued by OMB in late 1999 after the receipt of thousands of comments, limited the data that must be disclosed to published or cited data used to develop legally binding agency actions (NRC, 2002, pp. 2-3).

Reactions to Public Law 105-277 were strong, but mixed, and they demonstrate the complexity of formulating policy in this area and the range of issues to be considered. In general, the business and industrial communities, with the exception of biotechnology and pharmaceutical firms, were delighted with the law. Industries viewed the legislation as an opportunity to scrutinize the data on which federal policies and regulations are based. Biotechnology and drug companies, who often work with university researchers funded by federal grants, were concerned that the provision would negate their substantial investments in scientific research. Scientists, including those who in principle support data sharing, feared “that corporate or political interests will use the law to hamper research on controversial subjects, tie up scientists in red tape, circumvent confidentiality agreements, and thwart government regulations” (Hilts, 1999, p. 1). In summary, although both sides of the debate welcomed the goal of the law to provide open access to data, some opponents "argued that the Shelby amendment was too blunt and cursory to fully address an issue as complex as that of data access," and critics felt the final OMB guidelines interpreted the amendment too narrowly. At this time, the controversy over the revisions to Circular A-110 continues (NRC, 2002, p. 4).[4]

The law that changed FOIA's influence on access to research data was a significant step toward mandating the broader availability of research data, and the controversy it raised provides a glimpse into the complex issues that surround data sharing. Political and scientific controversies related to data sharing influence the willingness of scientists to share data. The results of research on secondary data use can inform law- and policy-making activities by increasing our understanding of the outcomes of data sharing.

A less legally binding, but key act related to scientific data occurred in 1994 when President Clinton signed Executive Order 12906 creating the National Spatial Data Infrastructure (NSDI). The order mandated all federal agencies to document spatial data they collect in fulfillment of their missions (Executive Office of the President, 1994). The purpose of this Order was to encourage innovation and to stimulate the economy through access to data. The NSDI provided an umbrella under which government agencies, private companies, and nonprofit organizations could work together to leverage the billions of dollars each spend to collect, manage, and disseminate spatial data (Domaratz, 1996). Even though the Order did not extend to data without a strong spatial component, it raised awareness about all types of data. Some ecological data sets are spatial, meaning they are associated with multiple geographical locations; these are the ones most affected by the creation of the NSDI. All ecological data have some spatial component, but for the most part, it is unimportant to the success of a project (Michener, 2000, p 96).

In summary, current U.S. laws and policies demonstrate a belief in the public good aspects of scientific data gathered with federal funds. Open access policies directed

toward publicly funded scientific and technical data have been adopted in recent years, especially in environmental and earth science research (NRC, 1999, p. 55). Internationally, and among private interests in the U.S. and elsewhere, these benefits are not supported as strongly (Lopez, 1998; NRC, 1997). In 1996, the Commission of the European Communities adopted its Directive on Legal Protection of Databases, providing unprecedented protection to any systematically arranged collection of data or information (Reichman & Uhlir, 2001, p. 271). The European Directive polarized U.S. parties with a stake in this issue. The U.S. Congress has drafted bills to provide protection for databases outside the Copyright Law, but to date no legislation has passed (Gasaway, 2002). Although the proposed legislation does not address data gathered with federal funds, Reichman and Uhlir (2001) perceived the public domain for scientific data in the U.S. shrinking based on these activities, and they warned that they portend negative results to the economy and to science. Obviously, the true power of scientific data to provide economic benefit is difficult to measure, but the more knowledge we have about the outcomes of data access, the better we will be able to judge the circumstances under which open availability to data should be encouraged.

Mechanisms to Share Data

The benefits attributed to data sharing, in combination with legal, political, scientific, and social demands, have led to a number of mechanisms to share data. In some fields, methods to share data are well established. Even in these areas, however, there is a push to make data available to those working in other disciplines and to expand the amount of data that is accessible. Digital libraries, electronic data archives, and

national and international information initiatives focused on data access are some of the institutional means under development to enable data sharing. These projects depend heavily on standards, a technical mechanism for sharing data. Standards include those to organize and describe data and to import and export data across multiple platforms. Although the institutional and technical mechanisms I describe in this section represent a proactive stance toward information sharing, they emphasize a technological infrastructure approach. Their compatibility with the cultural and social nature of research is largely untested. One goal of my study is to investigate these aspects.

Institutional Mechanisms

The wider availability of a diversity of information, including numeric data, sparked the interest of librarians in providing access to resources that formerly were not part of most library collections.

> The evolving digital library may play a key role for scientists by providing a unified environment for information discovery and access. In particular, the digital library can go beyond the traditional library, and provide direct, immediate location and access to both literature and data (McGrath et al., 1999, p. 188).

Previously, a minority of librarians encouraged their peers to make data collections available to their users, and those who did focused on centralized databases in the social sciences (Heim, 1987; Rinderknecht, 1991). Today, a greater number of librarians are interested in cataloging and providing access to numeric data. For example, the goal of the Alexandria Digital Library (DL) was to provide spatial data in a distributed environment. Researchers at the University of California-Berkeley DL built an electronic library of environmental information that included numeric data sets. In addition, a

number of special journal issues, books, and articles that appeared in the library literature in the last five-ten years reflect this growing interest (e.g., Hernon, 1995; Lutz, 1995; McGrath, 1999; Neuhold, 1998; Smith & Gluck, 1996; Trybula, 1997). Many of these publications emphasized spatial data due to the creation of the NSDI and because this type of data have practical applications in a diverse array of areas. The primary emphases of the recent attention by librarians have focused on educating the profession about scientific data, on strategies for incorporating scientific data into library collections, and on adapting cataloging standards for the description of digital data. Less attention has been paid to the users of these resources, especially the ways they utilize the new types of information they find in digital libraries, or even whether they discover these resources.

Archivists have a longer standing, if not widespread, interest in the preservation of scientific data. In this way, they have served a role as an institutional mechanism for access to data. Archivists have contributed a number of important insights to the documentation, maintenance, and use of data. These contributions include literature on the challenges of preserving digital data for the long-term, the problems of hardware and software obsolescence, the difficulty of supporting data collections, and the processes by which researchers create, store, and use data (e.g., Elliott, 1983; Geda, 1979; Haas, Samuels, & Simmons; Loewen, 1991-92; Long, 1995; NRC, 1995b; Warnow-Bluett & Weart, 1992). Historically, archivists preserved materials for their long-term value as records that document the activities of an organization for accountability and legal purposes. However, increasingly, they have promoted the maintenance of data because of their informational value. The National Archives and Records Administration

(NARA), in particular, has attempted to play a lead role in facilitating the preservation of scientific data gathered by federal agencies and has worked with others to develop mechanisms to accomplish this (National Academy of Public Administration, 1992; NRC, 1995b).

Specialized organizations that provide access to scientific data sets are presently the most recognized way to find and obtain data. There are many of these, and they vary greatly in terms of the subjects they cover and the services they offer. Some widely known data resources in the environmental arena include the National Oceanic and Atmospheric Agency's (NOAA) National Weather Service (meteorology), the NOAA National Oceanographic Data Center (oceanography), and the Carbon Dioxide Information Analysis Center (atmospheric science) (NRC, 1995a; NRC, 1995b). Organizations such as these provide the infrastructure to make data available to a large number of users. The field of ecology lacks a similar infrastructure.

There are also directories, many of them available on the World Wide Web, to help individuals locate data. The primary purpose of these resources is to connect a user with a data source, although raw data is also obtainable from some of these sites. NASA's Global Change Master Directory and the U.S. Long-Term Ecological Research Network are examples of data directories of interest to environmental scientists.[5]

World governments have also zeroed in on the value of sharing data and information and have developed mechanisms to do so. One example is evident in relation to complex, global environmental problems where, as discussed previously, the secondary use of data is believed to be of particular importance. Airborne pollutants from one country affect the health of another nation's aquatic organisms, and habitat

destruction in the rainforests changes the weather patterns in other parts of the world. Governments have taken an active interest in these problems because they have socioeconomic, health, and quality of life implications. Several national and international initiatives to share environmental information, including data, have resulted from cooperation among the world's governments. Biodiversity data, such as specimen collections found in the world's herbaria and natural history museums, are a special focus of several of these projects. A substantial amount of information already exists about the earth's biodiversity, but it is scattered around the world and remains inaccessible to users unable to travel to the relevant repositories (Edwards, Lane, and Nielsen, 2000). This problem led the science ministers of 29 industrial countries to create a Global Biodiversity Information Facility (GBIF).

> The virtual facility hopes to convert a growing tower of biodiversity Babel, replete with incompatible databases, confusing terminology, and uncataloged material, into a transparent source of information that is accessible to anyone, anywhere (Redfearn, 1999, p. 22).

The multi-million dollar project, intended to provide one-stop shopping for biodiversity information, is still under development (Edwards et al., 2000; Redfearn, 1999).[6] Since multiple names for a species often exist, one of the first challenges confronted by projects such as GBIF is the need for a standardized terminology. So far, much of the effort has gone toward projects such as Species 2000, the International Plant Names Index, and the Integrated Taxonomic Information System to help sort out differences in nomenclature (Edwards et al., 2000). Eventually, GBIF hopes to "foster interoperability with those domains outside of species diversity, such as sequence and other molecular data, geospatial and climate data, and ecological and ecosystems data" to probe questions not possible before (Edwards et al., 2000, p. 2313).

In the United States, the National Biological Information Infrastructure (NBII), an initiative of the U.S. Geological Survey's Biological Resources Division, is the U.S. node

for GBIF and hopes to serve as *the* virtual resource for data and information on biodiversity and ecosystems (Sepic & Kase, 2002). A 1998 report by a panel of the President's Committee of Advisers on Science and Technology (PCAST) called for funding to support the creation of a "next generation" NBII, known as NBII-II (PCAST, 1998, p. 69). Like the creators of GBIF, the PCAST panel's vision for NBII-II is an ambitious one. The dream is accompanied by the recognition that a great deal of research is required to make the NBII-II vision a reality. It is imagined that NBII-II will enable users to search through a multitude of data sets, combine the data found in new ways, analyze and synthesize them, and present the resulting information in a coherent manner (PCAST, 1998, p. 65).

The PCAST report identified specific research needs in communications technology, computer science, and library and information science vital to the success of the NBII-II. Some of the areas in which more knowledge is required include new statistical pattern recognition and modeling techniques; strategies for sampling and selecting data; data-cleaning methods to automatically correct database errors; visualization techniques that scale to large and multiple databases; and mechanisms to efficiently search through terabytes of data (PCAST, 1998, pp. 63 & 73). The Panel also noted the need for "ongoing, formative evaluation, detailed user studies, and quick feedback between domain experts, users, developers, and researchers" (PCAST, 1998, p. 73). The Panel emphasized that knowledge of user behavior is as important as technological breakthroughs to the design of flexible and useful systems. As the PCAST report noted, there is much to learn about all aspects of data sharing. So far, the focus has been on technical mechanisms to share data. Very little effort has been devoted to evaluating the effectiveness of these mechanisms, understanding the use of digital data resources, or analyzing the products produced through data reuse. Therefore, research in these areas is especially needed.

The ultimate success of projects such as GBIF and the NBII-II depends on more than technological developments and computer and information science research. These virtual facilities also rely on some change in the way science is practiced and reported and in the way data and information are shared among research communities. As discussed previously, the incentives for scientists to share data do not currently match the demand. In many cases where data sharing does occur, it is among communities that have shared understandings about the data, including appropriate uses, possible limitations, and specialized technical knowledge, although data exchange within the same community can also be challenge (Schiff, Van House, & Butler, 1997; Van House et al., 1998). Currently, federal funds awarded for scientific research come attached frequently with strings contingent on supplying access to data. Law and policy also require or encourage provisions for data access. These developments led to the creation of mechanisms like those described above, yet we know little about how well they work.

Technical Mechanisms

Institutional and governmental initiatives to share scientific data are dependent on standard ways to describe data and to transfer them across multiple platforms. If data are to be reused, they must be available in a form that can be accessed, exchanged, and manipulated, and they must be described and documented in a way that makes sense to a secondary user. Thus, the main technical mechanisms for sharing data, or what Bowker (2000a) referred to as "technical fixes," are documentation and computing standards (p. 649). Below, I describe the main types of standards relevant to scientific data. Further details about specific standards applicable to environmental data, especially metadata standards, appear in the next chapter.

The area of documentation that has received a lot of attention is metadata. Metadata is the term used for the documentation that describes data. One purpose of metadata for scientific data is to help potential users to locate data, to determine if they

meet their needs, and to provide information for accessing and using the data. Metadata is said to be the key to data retrieval and reuse, and much energy has gone into creating metadata standards for various user communities (Ercegovac, 1999; Michener & Brunt, 2000; Michener, Brunt, Helly, Kirchner, & Stafford, 1997; Milstead & Feldman, 1999b; Vellucci, 1998). Besides its role in data discovery, metadata is employed to describe the structure of data and to administer their use. We know very little about the effectiveness of metadata for its intended purposes, however, and researchers have identified the need for systematic research into the behavioral, technical, and sociological aspects of metadata (Fraser & Gluck, 1999; Goodchild, 1995; Thiele, 1998). Since a secondary user might work in an area similar to the disciplinary home from which the data were gathered originally or be from a completely different discipline, the structure of the metadata, the level of granularity of the description, and the fields required are not the same for all users. As Goodchild (1995) observed, “the more widely distributed the data, the more difficult it is to anticipate uses and thus to determine what to include as metadata" (p. 418).

Ecological data present particular challenges in terms of documentation for secondary use (cf., Bowker, 2000a; Bowser, 1986; Michener et al., 1997; NRC, 1995a; NRC, 1997). The lack of a sharing tradition in ecology can be attributed partially to these qualities. Data sets tend to be small and highly diverse, and the methods and techniques used to obtain and manage the data vary. This variability makes it difficult to describe ecological data adequately enough for others to use them. Ecologist William Michener and his coauthors summarized some of the characteristics that complicate the secondary use of ecological data:

> Ecological data sets are often extremely complex. Missing values, midcourse modification of sampling or laboratory procedures, addition or deletion of study parameters, personnel turnover, plot or habitat modification by disturbances (natural or anthropogenic) or changing

> environmental conditions, and numerous other factors leading to data anomalies are commonplace (Michener et al., 1997, p. 332).

The difficulties presented by these characteristics are not trivial, and they complicate the realization of the positive outcomes that many believe will result from the secondary use of ecological data.

Metadata standards are used to describe data in a uniform way. Other standards were created to transfer information between parties using different computer systems. For example, the Spatial Data Transfer Standard (SDTS) provides a mechanism to exchange data across multiple computer platforms (U.S. Department of Commerce, 1992). The MARC (MAchine-Readable Cataloging) standard is used to transfer bibliographic records between cataloging systems (Frank, 1997; Larsgaard, 1996).

Costs of Data Sharing

Making data available through the mechanisms described in the previous section is an expensive process, and nearly all discussions of data sharing acknowledge the numerous costs of preparing and maintaining data for secondary use. Many of these same discussions also note that there are costs for the user as well as for the provider of data, such as investing time to understand the data well enough to use them. For the provider, the main costs are the expenses to organize, store, and preserve data and to support their use. Since few systematic studies of data sharing exist, there may also be costs that we do not recognize or existing costs that are not fully understood. Whatever the expenses, they can serve as disincentives to sharing.

The organization, documentation, and preservation of data to aid retrieval and use are necessary to secondary users. These activities can also be helpful to the original data

collector, and in fact, this argument has been presented to scientists as an incentive to document their data (Ingersoll et al., 1997). Scientists and technicians, computer scientists and data managers, and archivists and librarians are among those who presently perform these tasks. There is a great deal of existing data in both paper and digital forms that are not documented properly, and new data are collected all the time. The vast amount of data makes it impossible for any one community to organize it all (Milstead & Feldman, 1999a). Regardless of who is responsible, the costs to organize, document, and preserve scientific data are significant. A recent report estimated that the cost of data systems to support intensive modeling efforts could amount to a quarter of a project's total budget (NRC, 2000, p. 62).

Large observational data sets, such as those common to astronomy and meteorology, are challenging and expensive to store and preserve. Even though the price of digital storage keeps falling, databases continue to grow in size, and so the expenses for storage remain a factor (NRC, 1995b; NRC, 1997, p. 4). Added to this is the fact that the long-term reliability of today's digital storage media is inadequate, which means that data must be refreshed or migrated to ensure that they remain readable. However, it is often not only the data that must be preserved, but the hardware and software on which they depend. The preservation of digital resources, including scientific data, is complicated and made more expensive by hardware and software obsolescence (Rothenberg, 1995). Preserving the software in which data were stored originally is a challenge for both large and small data sets. The expenses of a long-term commitment for equipment and staff to support preservation can also be quite high.

The cost of storage for small data sets from the observational sciences is not an obstacle. The greatest expense of preserving small, observational data sets is the time-consuming and labor-intensive process of assigning metadata and otherwise documenting the data so they can be reused. In spite of these costs, some authors have recommended that all observational data should be preserved (FLED, 1995; Michener et al. 1997; NRC, 1995b). Proponents of this view argue that all observational data should be saved because it is impossible to determine their future value. Such recommendations are costly and impractical to implement, however, and they may not be necessary. If we had more knowledge about the use of data sets, we could identify the characteristics of data worth preserving and invest resources only in those that are most valuable.

The focus on interdisciplinary research brings about the need to describe data so they are understandable to users from many disciplines; this complicates the organization and delivery of information and adds to the cost. Most information systems and organizational structures are built around disciplines, and the focus on interdisciplinary research challenges these paradigms. Members of a discipline share common terminology and research methods and their own channels for disseminating research results (Klein, 1996; Pierce, 1990). Most metadata standards evolved around the needs of particular communities, and therefore, they reflect those requirements in their design (Michener & Brunt, 2000; Milstead & Feldman, 1999b). This is positive from the standpoint of meeting the requirements of a certain discipline, however, the elements in a particular metadata standard may reflect too much, too little, or simply the wrong information for users from other fields (Hill et al., 2000). Thus, the commonalities that bind disciplines together can hinder the exchange of information with other fields. In

order to share information effectively across disciplines, information must be described in a way that is meaningful to those working in other areas. Several approaches to this problem have been suggested, such as linking together existing controlled vocabularies and dictionaries, creating ontologies, using keyword mapping, and employing other mechanisms to provide concept-based searching and semantic interoperability (Cortez, 1999; Harvey, Kuhn, Pundt, Bishr, & Riedemann, 1999; Pundt & Bishir, 2002). These solutions, many of which are still under development, add to the costs of data sharing.

A metadata scheme for nongeospatial ecological data was developed to meet the needs of users from different backgrounds and with varying levels of technical knowledge (Michener et. al, 1997). A Level 1 user with technical expertise in the subject area, including familiarity with data collection and analysis procedures, might require only a basic description of the data set. A Level 2 user, searching a metadata catalog and using the data without direct contact with the data holder, would require much more detail about the data set. In this scenario, the cost of creating metadata rises as the user levels increase. Michener and his colleagues admitted that it would be excessive to assign metadata at the highest level to all data sets. Decisions about the level of metadata to assign would be easier if we knew more about the secondary users and uses of scientific data. For example, what standards are important? How are metadata used, if they are? And, how do users judge data quality? Answers to these questions would also make it possible to test Michener et al.'s user categories and their hypotheses about the amount of metadata required by different levels of users. Finally, knowledge about use would help determine which data are worth the cost to document and preserve. As Sprehe (1999) noted, “Once information is accessible, dissemination revolves around the

question of whether the intrinsic merits of the information are such that it is worth investing the resources to add value and disseminate the information" (p. 343).

Negotiating disciplinary divides is especially difficult when it comes to scientific data. Besides differences in methodology and terminology, data are gathered under the theoretical assumptions of a particular field (Bowker, 2000a; Edwards, 1999; Kuhn, 1970 [1962]). The technical knowledge required to appropriately use and evaluate scientific data is reflected in data holders' apprehensions over the misuse of their data. These concerns apply to the use of data both within and outside a particular discipline. Ecology, for example, is comprised of many subfields, each with varying views and methodologies (Bowler, 1993, p. 19; Kwa, 1993, p. 216). Bowker (2000a) described the data documenter's task with an illustrative and humorous analogy.

> In essence, the record-keeper is being asked to abstract the record set from the historical flow of time – to provide enough information so that a limnologist from Mars (who presumably has been out of work for several million years) can come along and from the dataset and a sufficient command of English interpret the data (p. 664).

Technical knowledge is also necessary to evaluate data quality, and a reliance on poor-quality data can lead to inaccurate results (Goodchild, 1995). This is an unseen cost that is potentially significant.

Benefits of Data Sharing

In spite of all the obstacles, including the significant costs, the sharing and secondary use of scientific data are promoted because the educational, scientific, and socioeconomic benefits are thought to be substantial. This is clear from the mechanisms being created to support data sharing and from the laws and policies that encourage or

mandate access to data. All these activities are intended to achieve certain desired outcomes. Recent endeavors are meant to increase access to the information and records of the government, to stimulate the economy, and to advance science.

The ability of open access to data to stimulate the economy is based on a view of scientific research as a public good. This philosophy has been most evident in regard to federal funding for research. Whether public good benefits actually apply to data, however, is a topic of debate. It is accepted generally that scientific research itself has the attributes of a public good: nondepletability and nonexcludability (NRC, 1997, p. 112). Nondepletability refers to the fact that a product cannot be used up and is therefore available for others to use. "Nonexcludability means that the good in question produces benefits from which others cannot be excluded and which cannot easily be constrained only to those who pay" (NRC, 1997, p. 112). Scientific data have certain characteristics of a public good, and therefore some believe that open access is important. Others point out that scientific journal articles also have public good aspects, and they have been copyrighted by scholarly journals for years without impeding the flow of science (NRC, 1997, p. 113). Our current knowledge about the benefits of scientific data and to whom they accrue and under what circumstances is insufficient to sort out this critical policy issue. This situation provides yet another reason to investigate the process and results of secondary data use.

In addition to economic gains, substantial positive outcomes from data sharing are thought to accrue to education, scientific practice, and social goals. Fienberg, Martin, & Straf (1985) summarized these benefits.

- reinforcement of open inquiry
- verification, refutation, or refinement of original results

- the promotion of new research through existing data
- improvement of measurement and data collection methods
- development of theoretical knowledge and knowledge of analytical technique
- encouragement of multiple perspectives
- provision of resources for training in research
- protection against faulty data
- encouragement of more appropriate use of empirical data in policy formulation and evaluation

Other social scientists reiterated these benefits and discussed additional positive outcomes that might result from data sharing, such as better quality data and greater accountability (Sieber & Trumbo, 1995). Of particular value are the new insights that are believed to be possible when data are reused for a purpose different from that for which they were gathered. Although there is anecdotal evidence that such positive outcomes can result, few empirical studies have tested these benefits (Bowser, 1986; Whillans, Regier, & Christie, 1990).

The existing knowledge about the benefits of data sharing comes mostly from the use of large data files (NRC, 1995a). Data depositories are well-established in some fields, such as agriculture, astronomy, genetics, and meteorology, and some data sets in these disciplines may be used many times. As mentioned previously, the existence of data sources can be attributed partly to the nature of conducting research in these areas. Many of these resources also exist in areas that are the responsibility of federal agencies that gather data in support of national problems and concerns. Although these institutions may understand the needs of their primary clientele, they know little about new user communities, and they could benefit from a better understanding of the process of secondary data use. One of the concerns for these sources, and for large data infrastructures generally, is that the accessible data to a large degree determine the problems that are possible to study (Bowker, 2000a; Rockwell, 2001). For example,

Richard Rockwell (2001) noted that if Robert Putnam and his colleague Henry Brady had not resurrected a series of surveys conducted by Roper Starch Worldwide, a research company, this collection on social and political trends would have gone unused. Instead, it is now a popular data series.

We know little about the challenges, benefits, and outcomes of the sharing of small, observational data sets gathered by a single researcher or a small team of scientists. As mentioned previously, ecology lacks a formal infrastructure for sharing data, and one of the several reasons for this may be the fact that most ecological data sets have a small potential for reuse because of their limited scope and scale and the challenges they present to potential users. Science studies can help shed light on the social aspects of data sharing among scientists in all disciplines through its insights on the relationship between theory and data, the replication of scientific studies, and an analysis of sharing behaviors as part of complex systems of exchange (Hilgartner, 1997).

Thomas Kuhn gave primacy to theory in science (Cole, 1992, p. 7). Kuhn believed that scientists with different theoretical views could interpret the same data differently (Kuhn, 1970 [1962]). More recently, scholars have pointed out that scientists' interpretive differences occur for all sorts of reasons.

> Scientific disagreements are not uniform. Some involve disagreement about the value of a specific parameter; others, differences over which theory or model to accept; still others, about what variables and relationships are worth studying, and what data and methods can be employed legitimately (Roberts, Thomas, & Dowling, 1984, p. 113).

Disputes based on epistemological views are difficult to resolve (Edwards, 1999; Roberts et al., 1984; Von Schomberg, 1993). A scientist's epistemology reflects the fundamental paradigm through which he/she views the creation of knowledge. Arguments between

researchers over scientific objectives and methods are stumbling blocks to knowledge exchange. For example, the disagreements among "high-proof" and frontier researchers over the proper role of data in global climate change models represent epistemological differences that scientists are unable to settle (Edwards, 1999). Bowker (2000a) noted that it is impossible to separate data and theory. When data are rolled into theory, the loss of data is not a problem, but when data are separated from theory it becomes difficult to use data in multiple ways, especially across disciplines. The important point to remember is that scientific data are not neutral or one-dimensional. Data alone, in the hands of scientists, do not have the capability to solve difficult problems. In fact, as climate change research illustrates, they may create more complexities

Latour and other contemporary sociologists believe that science is socially constructed. This theoretical outlook has consequences for the sharing and reuse of data. Social constructivists hold that the presentational methods of scientists are tools meant to persuade others to share a particular view and that replication is not as simple as following published results (Collins, 1992 [1985]; Latour & Woolgar, 1979). Therefore, published studies do not provide other scientists with the information they need to reproduce a study or to understand the data associated with it; social exchange is an integral part of scientific understanding. The implications of this for data sharing are that misinterpretation of data might increase the greater the social distance between data users and data producers and that metadata, while helpful, is unable to resolve this problem for anything very complex.

Social constructivists Stephen Hilgartner & Sherry Brandt-Rauf (1994) argued that the form of the data determines when they are ready for sharing and that different

access policies are needed for various forms. The authors noted many interesting areas for further research (pp. 358-366, 369).

- research on the actual practices of scientists regarding access to data
- notions about what can be packaged in a form that can be "published" or transferred, particularly in light of the fact that data do not have a well-defined or stable meaning
- investigation of data access practices that contribute to research productivity while promoting social goals
- an understanding of the audience and market for data

The topics they posed could be used to begin to test the stated benefits of data sharing. It is only recently that a few researchers have pursued these important questions (e.g., Campbell et al., 2002; Louis et al. 2002). Hilgartner and Brandt-Rauf's primary interests surrounded access to data; they paid little attention to what occurs after data are obtained, which is a main interest of my study.

While sociological theories offer some insight into the feasibility of data sharing as cultural practice in science, more practical concerns have dominated alternate views of the rosy picture of benefits painted by data sharing proponents. Stanley and Stanley (1988) voiced concern about the negative effects of a shift from voluntary to mandatory sharing. In general, they and others noted that these effects include the costs of sharing data, the lack of reward, loss of control over use of the data, and concern about the qualifications of secondary users (e.g., Marshall, 2000; Stanley & Stanley, 1988; Van House et al., 1998; Van House, 2002). Stanley and Stanley saw data sharing as a continuum with voluntary sharing at one end and mandated sharing at the other. They believed that the intended use affected the primary researcher's willingness to share data. They recommended that data sharing be voluntary, that there be a balance between the sharing and development of data, and that guidelines be created for secondary use. Many

authors believe that the benefits outweigh these concerns and that these problems can be addressed through education, metadata, and policy (Ceci, 1988; Fienberg et al., 1985).

Even though they are in the minority, the opinions of Stanley and Stanley are worth consideration and systematic investigation because embedded in their concerns are questions that have repercussions for the best use of resources and scientific talent. Since the efforts devoted to data sharing divert time and resources from other activities it is important to ask several questions. What is the appropriate amount of activity that scientists should invest in sharing? What degree of control should investigators expect to have over data that they share? Do the benefits outweigh the financial and human costs of sharing? Should all data be subject to the same sharing policies? For example, should data gathered by an individual investigator working in a field setting be exchanged under the same guidelines as the large amounts of data gathered by remote instruments?

Finally, it is useful to ponder if it is important to protect to some degree a scientist's ownership in data. The benefits attributed to data sharing may be realized in most cases, some cases, or rarely. At this time, we have little evidence to test assumptions about the positive outcomes and to judge how frequently and when they occur since we lack answers to the above questions.

Proposal for a New Study

Scenarios describing how data will be located, retrieved, and reused often present an idyllic picture of the ease and simplicity with which secondary use will occur. For example, it is imagined that natural resource managers will quickly solve complex environmental problems without leaving their computer workstations; controversial

decisions will be obvious and indisputable when all sides have the opportunity to examine the data on which environmental policies and regulations are based; and the cross-disciplinary linkages needed to solve today's most perplexing natural problems will occur once data are available to anyone who wants them. These scenarios persist despite the fact that little is currently known about the secondary use of scientific data. What is known suggests that many challenges must be surmounted before a realistic picture of the possibilities can develop.

Existing research has focused on the social aspects of data sharing, especially the mechanisms under which scientists grant or deny access to data, and on the effects of databases on scientific communication and on the work of various communities of users (cf., Campbell et al., 2002; Hilgartner, 1995; Hilgartner & Brandt-Rauf, 1994; Louis et al., 2002; McCain, 1991; McCain, 1995; Van House et al., 1998). My study goes beyond these investigations to tell us much more than we currently know about the experiences of secondary users of scientific data. Little attention has been paid to what motivates scientists to look for data, how they find and obtain them, and how they use them. These are the main interests of the study I propose here. In addition, the problems and issues that surround the wider availability of scientific data are similar to those of other digital resources, and thus, the results of my study are broadly applicable. For example, it has been noted that publications are becoming more like databases (Cameron, 1998). Traditional print publications were permanent, citable, and accessible. Now, similar to databases, they are becoming increasingly dynamic.

As I mentioned previously, much of what is known about the secondary use of scientific data concerns large data sets without strong ownership ties and in fields with

established infrastructures for sharing data. A particular need has been identified for interdisciplinary research directed toward the solution of global environmental problems (Macilwain, 2000; PCAST, 1998; Steele & Stier, 2000). Some of this research will depend on large, standardized data sets from areas such as meteorology, hydrology, and remote sensing. Other work, however, will require small, observational data sets from fields such as ecology. Still other projects will be built on combinations of multiple data sets with differing characteristics. In this study, I focus on the sharing and secondary use of data to address ecological research questions. I investigate all types of data obtained and used by ecologists, but I pay particular attention to ecological data, which present significant obstacles to reuse and where little is known about how users overcome these challenges.

Digital library researchers have noted that it is challenging to study the use and impact of digital libraries because baseline data about use prior to the advent of digital libraries is lacking (Bishop, Neumann, Star, Merkel, Ignacio, & Sandusky, 2000; Bishop & Star, 1996). Thus, it is difficult to determine what effect, if any, digital libraries have on work practices, productivity, and the creation and communication of knowledge (Kaplan & Nelson, 2000). One portion of my research project provides critical baseline knowledge regarding the existing use of data by ecologists. Although centralized data archives and a formal tradition of sharing are not typical in ecology, there is enough activity to analyze the sharing and secondary use of scientific data within ecology and related disciplines. My study goes beyond the provision of baseline data, however, to provide a rich description of the experiences of secondary data users.

Research Questions

Specifically, my study will investigate the following research question:

- What are the experiences of ecologists who use shared data?

The following subquestions define the specific areas that comprise ecologists' experiences for the purpose of my study.

- How do ecologists locate data?
- What are the characteristics of the data ecologists collect?
- What information about the data do ecologists receive and/or depend on to use the data?
- How do ecologists assess the quality of the data they receive?
- What challenges do secondary data users face, and how do they overcome them?

My primary data collection method will be semi-structured in-depth interviews with researchers who reused data. Recent instances of secondary data use by ecologists will be gathered from two prominent ecological journals and used as a springboard to identify informants. Interviews will provide information about motivations for looking for data, methods of data discovery, the characteristics of the data obtained, issues related to understanding the data in order to use them, and attitudes about data sharing and secondary use. I provide further details about my research methods in the third chapter of this dissertation.

Significance of the Study

The current data-sharing environment is characterized by several opposing trends. There is an increased interest in access to data as evidenced by contemporary U.S. laws and policies and national and international information initiatives, including standards development. On the other hand, on a global scale, intellectual property rights for data are being strengthened. Knowledge of the benefits and outcomes of data sharing are

confined to certain areas, such as meteorological and space science data, while very little information exists about other types of data. Yet, such knowledge is crucial to the development of future intellectual property rights related to data and is necessary to gain a deeper understanding of the contribution of scientific data to the economy and to scientific content and practice.

Early digital library conferences made clear how little is known about the results of information seeking and the role of documents, in a range of genres and media, in meeting the information needs of users (Bishop & Star, 1996). This lack of knowledge points to many opportunities to understand diverse users, both in public settings and across scholarly disciplines. Robert Boruch (1985) noted over 15 years ago that it is reasonable to expect a variety of outcomes, both positive and negative, from data sharing.

> Apart from political incentives, the problem of understanding how to evaluate the products of data sharing systems, how to improve them, and when to encourage or terminate them seems a reasonable intellectual problem (p. 113).

Despite all the work and research that have gone into building digital resources, we still know very little about the use of scientific data and similar forms of digital information (Kaplan & Nelson, 2000). Yet, much energy has gone into mechanisms to enable data sharing, and these activities show few signs of slowing down. Scientists are mandated to make data available as a condition of funding, and the notion that data sharing and interdisciplinary research are key parts of solutions to the world's problems are firmly entrenched.

Lastly, one of the byproducts of this study may be to tame the rhetoric optimistically spoken about the power of data sharing to solve a host of complicated issues and problems. Combining ecological data from multiple studies or mixing

ecological data with that from other disciplines may, indeed, be a necessary component of solutions to the world's ills. However, as Rob Kling has stated several times, the power of information technology to increase labor production, to revolutionize scholarly communication, and to reshape education has rarely come to pass as predicted by technological pundits (Kling, 1999; Kling & McKim, 1999; Kling, Rosenbaum, & Hert, 1998). These prophecies would be harmless if they did not divert attention from important attempts to gain a deeper understanding of the issues. Changes to complex phenomena are influenced by social contexts and relationships, as well as by technical advances. The sharing and use of scientific data are part of an intricate web that includes scientific culture and practice, political and social demands, funding pressures, varying epistemological views, and technological issues. In addition, complex matters, like ecology in the broadest sense, cannot be reduced easily to quantifiable elements that behave the same across all scales. Combining data from many small-scale studies to predict changes on a larger scale is not a simple matter. In fact, the problem of pattern and scale in ecology is arguably "the central problem in ecology" (Levin, 1992, p. 1943). Thus, the time is right to begin filling the void in our knowledge about the role of data sharing in environmental problem solving.

Notes to Chapter 1

[1]Information on the award renewal is available at https://www.fastlane.nsf.gov/servlet/showaward?award=0072909.

[2]Ecological Archives, a publication of ESA, is available at http://esapubs.org/archive.

[3]Recently, the GILS acronym was changed to stand for *Global Information Locator Service*. The focus was also changed from a tool for discovering federal information to a mechanism to find all information. See http://www.gils.net.

[4]See the NRC (2002) report for a cogent discussion of the issues.

[5]Further information on these data sources is available at http://gcmd.gsfc.nasa.gov/ and http://lternet.edu/data/.

[6]As a service, GBIF is not yet operational.

CHAPTER 2

LITERATURE REVIEW

There is a rich body of literature to draw from related to data sharing and reuse, although few empirical studies exist. The advent of digital libraries and archives, the availability of electronic databases in fields such as genetics and molecular biology, the implementation of policies that encourage sharing, and the increasingly problem-oriented approach of some sciences provide topics ripe for research to improve our understanding of the retrieval, use, and outcomes of scientific data sharing and its effects on the content, conduct, and communication of science. These topics are important to investigate because scientific databases exemplify many of the characteristics of digital resources generally whose wider availability promise to upset our notions of the practice of science, as well as the retrieval, support, collection, and use of information. Research is needed that goes beyond the prescriptive and anecdotal nature of most of the existing literature.

An investigation into the sharing and secondary use of scientific data must be grounded in several important domains. These areas, especially as they pertain to scientific data sharing, are the subjects of this chapter. In the first section, I review literature that provides useful background for understanding the context of data sharing. Second, I discuss the social and cultural aspects of scientists' work with an eye to how these are, or might be, changing due to technological advances and political and social demands. In the third part of the chapter, I focus on the characteristics, organization, and

standards of institutional and technical mechanisms to support the information needs of scientists and the scholarly work that addresses the adaptations being made in these areas to meet proposed changes in scientific work. Finally, I conclude the chapter with an analysis of how the existing literature applies to what we know or do not know about the field of ecology.

Data Sharing in Context

The sharing of data became an identifiable topic of discussion when computers began to play a more common role in research. In the 1980's, social scientists produced a number of articles and books on data sharing. For the most part, these authors encouraged their peers to make their data available for others to use by explaining to them the educational, scientific, and socioeconomic benefits. Social scientists also highlighted the obstacles to data sharing. Although my study focuses on the secondary use of observational scientific data, these writings remain applicable to research data from many fields, and thus, they merit some discussion here.

One of the first reports to discuss secondary data use was Sharing Research Data, which was mentioned briefly in chapter 1 in regard to the benefits derived from data sharing (Fienberg et al., 1985). Although the report focused on positive outcomes, it also emphasized the challenges and obstacles to data sharing in the social and behavioral sciences. It discussed the technical hurdles, such as incompatible hardware, software, and data structures as well as the costs to store, document, transfer, and use the data. The authors noted that the use of computers facilitated sharing, but that they also led to more need for standardization and better documentation. At the time, Fienberg and his

colleagues observed a growing concern over privacy and confidentiality, an issue especially pertinent to data from behavioral science research. They also foreshadowed the growing profitability of data and its affect on increasing proprietary restrictions. Despite frank discussion of the challenges and obstacles, the book encouraged social scientists to make their data available. Sharing Research Data was one of several publications by social scientists that together addressed the cultural, financial, legal, political, and technical obstacles to data sharing (cf., Ceci, 1988; Ceci & Walker, 1983; Nelkin, 1984; Sieber, 1988; Sieber, 1991; Stanley & Stanley, 1988). These writings were important in articulating many factors that remain issues today.

The 1990's were characterized by government studies on data sharing, particularly data from the biological and physical sciences. Driven by recognition of the considerable government investment in data collection, the resulting reports consisted of investigations into the range of databases available in federal agencies, the state of data sharing in particular disciplines, the uses of shared data, recommendations for encouraging secondary use, and barriers to sharing, especially those related to intellectual property rights (cf., National Academy of Public Administration, 1992; NRC 1995a; NRC 1995b; NRC, 1997; NRC, 1999).[1] Together, these reports summarized the primary issues related to data sharing, including many of those discussed previously by social scientists, and they signaled greater activity and interest in data sharing. Although they noted many similar issues, government studies from the 1990's expanded on certain topics only touched on by social scientists in the previous decade. In the intervening years, technology advanced considerably, and the biological and physical sciences struggled to keep up with the large amounts of data gathered through improved

techniques and instrumentation. Solutions to the technical impediments of data sharing became more important. A growing need was recognized for better data management tools and interfacing methods to deal with large volumes of data and for the interoperability of hardware, software, and data management technologies. The focus on interdisciplinary research presented problems for locating and describing data, and this led to a call for research into new approaches to aid diverse audiences in locating information and in improved system design methods (NRC, 1995a, p. 95). In addition, societal problems seemed to grow beyond the ability of any one institution, discipline, or nation to address. One result of all these factors was the development of "big science," "a funding umbrella for multiple individuals and institutions to conduct coordinated data acquisition, investigation, and publication" (NRC, 1995b, p. 16). The Global Change Data and Information Program and NASA's Mission to Planet Earth are examples of this approach.

The current decade shows promise of the federal government's continuing interest in promoting the sharing of scientific and technical data and in its sponsorship of symposia and workshops that bring together diverse groups of stakeholders to explore solutions to new and old issues.[2] Intellectual property rights, which are always contentious, have taken on a new urgency in light of controversies over the sharing of human genome data, an increase in partnerships between academia and private enterprise, and what some see as signs of an erosion in access to public domain data (Marshall, 2002; NRC, 2002; Reichman & Uhlir, 2001; Siang, 2002). Privacy of data and closer examination of the sharing responsibilities of scientists are two other issues that have moved to the foreground (Marshall, 2002; NRC, 2002).

As a topic, data sharing has not been limited to government reports and the writings of social scientists. Other authors, many of them researchers themselves, have written about the cultural, legal, political, and technical issues that can affect the sharing of data. They have covered a number of the same topics discussed previously, but they have also offered unique insights on the potential and challenges of data reuse based on personal experience. For example, Whillans, Regier, and Christie (1990) described the multiple uses made of fisheries data gathered under the supervision of Ontario biologist F. E. J. Fry. Bowser (1986) made real many of the challenges of secondary data use in the account of his experience reusing data on the biology and chemistry of Wisconsin lakes. The difficulty of combining data from multiple institutions in order to investigate the impacts of dams on stream fishes was the subject of a paper by Robert McLaughlin and his colleagues (McLaughlin et al., 2001). The authors' integration of data was complicated by variability in the information collected by each institution and by the diverse means those organizations used to gather, organize and store the data. Other authors mentioned difficult to resolve issues: that data can be interpreted in multiple ways to support different arguments; that the results of a research project can be called into question by poking at the credentials and expertise of the investigator; and that the methods and techniques used to collect the data can be criticized (Hilts, 1999). Impediments to data sharing, such as intellectual property, quality and liability concerns, and the scientific reward structure, were also described. Data quality, in particular, has been a concern of scientists who have written about data sharing. Goodchild (1995) argued that the lack of "suitable methods for formulating and communicating information on quality currently forms a significant impediment to sharing..." (p. 414). Others have

made this same observation and have noted that although information technology makes it possible to handle large volumes of data, it does not help determine quality (Averch, 1985; Bikson, Quint, and Johnson 1984).

Government studies, anecdotal descriptions, and the writings of social scientists have done much to illuminate the broad range of issues that surround data sharing. Systematic research on data sharing to help address these issues is lacking, however. Exhaustive searches of the literature in the social and life sciences turned up few writings that went beyond descriptions of the challenges, costs, and benefits of data sharing and reuse. It is only recently that issues of data access began to be addressed by a few scholarly researchers. Their work was driven largely by debates over access to and ownership of scientific data. The economic value of some data has been increasingly recognized, and this has led to disputes over ownership and battles for intellectual property rights. Nelkin (1984) observed almost 20 years ago that the knowledge, not just the products, generated from research was growing in economic and political importance. The increasing number of ownership debates was also spurred by the blurring between basic and applied research, the growing public demand for information, and more recently, by the increasing number of partnerships between academia and private enterprise (Louis et al., 2002; Nelkin, 1984). These forces became the jumping off point for the research by sociologists and information scientists that I describe later in this chapter. For the most part, though, data sharing has moved ahead through laws and policies and social and technical mechanisms under the assumption that the benefits and incentives will ultimately overcome the obstacles and disincentives.

Scientific Culture, Practice, and Communication

A number of potential changes to scientific culture, practice, and communication are embedded in the demand for data sharing. One of the keys to successful data exchange is the ability to make explicit the implicit knowledge held by the source of a scientific data set. Metadata is one attempt to span the distance between the collector and the secondary user of data. However, metadata can only partially bridge this gap because, as Harry Collins (1992 [1985]) illustrated vividly in his description of the replication of the TEA laser, social exchange is an important part of scientific exchange, even within the same discipline. A scientific article is supposed to provide the information that another researcher needs to duplicate an experiment, but Collins discovered that visits to labs to work with other scientists were necessary to duplicate the TEA laser. Much of what we know about how scientists negotiate distance is found in studies that investigate the practice of interdisciplinary research, especially the ways in which scholars locate, understand, and use information from other fields. This knowledge is applicable to understanding the use of shared data. In this section, I review the literature in this area and then discuss how it relates to the sharing and reuse of scientific data.

The Nature of Disciplines

Interdisciplinary research is offered as a panacea to many of today's complex problems (Klein, 1996; NSB, 2000; Pierce, 1999; Steele & Stier, 2000). Yet, we have little evidence of its effectiveness to develop solutions to these concerns. We comprehend a little more, but we still lack a great deal of knowledge, when it comes to

understanding and serving the needs of interdisciplinary researchers. Any attempt to do so must be preceded by an understanding of how interdisciplinary research is different from the traditional disciplinary isolation of most scientific fields.

All disciplines share several characteristics centered on control. Disciplines exercise control by identifying the questions that will be attended to and by framing the context in which they will be investigated (Klein, 1996). Gibbons et al. (1994) remarked that a discipline's "cognitive and social norms determine what shall count as significant problems, who shall be allowed to practice science and what constitutes good science" (pp. 2-3). Attaining competence in a discipline requires mastery not only of relevant intellectual content, but also of what is not written down (Pierce, 1990). Members of a discipline share common terminology and methods and their own publication channels for disseminating research (Klein, 1996; Pierce, 1990). Disciplines build cumulatively on commonly accepted knowledge to construct new work (Klein, 1996; Pierce, 1990). Pierce (1990) combined these concepts into a succinct definition.

> To work ... from a disciplinary perspective means one's work is in some way (through the problems addressed, the previous work acknowledged, the publication channels used) connected to that of others in the relevant discipline. There is no better way to define what *discipline* means, because any more abstract mapping of disciplinary boundaries leads to inconsistency and conflict (p. 51).

Obviously, the discipline approach has its merits, and even with all the attention given to interdisciplinary research, no one argues seriously for its demise. It is not always the best method, however.

> If the separate bodies of knowledge of different disciplines provide strength, they are also a weakness. Disciplines build their exclusive communities by stressing social participation over subject matter. At some point, work on the same subject done in different disciplines must be recognized and dealt with as a whole, or the insights originating in a discipline are lost to those outside its circles (Pierce, 1990, p. 58).

Klein (1996) summed up the factors that led to the emphasis on interdisciplinary research: "Interdisciplinary approaches arise because of a perceived misfit among needs, experience, information, and the structure of knowledge embodied in conventional disciplinary organization" (p. 134). This reflects the situation in most sciences that have a role to play in solving problems, although the rhetoric overemphasizes the ease with which scientists from the same discipline communicate with and understand each other.

Interdisciplinary Research

Several terms are used to describe the concept of research that extends beyond the bounds of a particular discipline, including multidisciplinary, interdisciplinary, and transdisciplinary (Palmer, 2001, pp. ix-x). A multidisciplinary approach is the most conservative one. Researchers from different fields work together on a problem, but they do this by dividing up the work according to their respective expertise; no attempt is made to understand other disciplines in detail. Interdisciplinary work is a step beyond this and occurs when knowledge, experience, technology, or expertise is transferred among the worlds via borrowing, collaboration, and/or boundary crossing (Pierce, 1999). Gibbons et al. (1994) used the term transdisciplinary to describe a step beyond interdisciplinary work where “the shape of the final solution will normally be beyond that of any single contributing discipline” (p. 5). This is an intriguing but difficult to prove transformation, and they have been criticized for a lack of evidence for what they say is occurring (Godin, 1998; Weingart, 1997). For now, *interdisciplinary* is the most commonly used term in the literature, and I use it in this chapter.

Scientists who conduct interdisciplinary research face a number of challenges because the disparities between disciplines make it difficult to communicate information across them (Palmer, 1996; Pierce, 1999). These differences include the expectations of those involved in the peer review process, the models or paradigms on which research is based, and the distinct stylistic and presentational features that exist in each field (Pierce, 1999). Boundary crossing via publication outside one's discipline is the most direct form of interdisciplinary information transfer (Pierce, 1999; Steele & Stier, 2000). Boundary objects play a key role in the successful translation of information between different communities. Star and Griesemer (1989) described in detail the boundary objects that were central to developing and maintaining coherence between scientists, sponsors, and amateur naturalists during the early years at the University of California-Berkeley Museum of Vertebrate Zoology (MVZ).

> In natural history work, boundary objects are produced when sponsors, theorists and amateurs collaborate to produce representations of nature. Among these objects are specimens, field notes, museums and maps of particular territories. Their boundary nature is reflected by the fact that they are simultaneously concrete and abstract, specific and general, conventionalized and customized (p. 408).

For example, Star and Griesemer found that the state of California, which both amateur collectors and scientists shared as a common referent served as a boundary object at the MVZ. Standardized field forms, another boundary object, ensured that amateurs collected the same data whenever they obtained an animal and that this data would suit the needs of professional scientists. More recently, Van House (2002) argued that digital libraries are boundary objects because they are used by diverse communities and are created by coalitions of users, information owners, and technologists. Palmer (2001) extended the notion of boundary objects to investigate all the activities and elements that

researchers employ to cross boundaries. She found that some of the most important "trajection elements are people, data, methods, and words" (Palmer, 2001, p. 11).

Star and Griesemer observed that boundary objects have “different meanings in different social worlds but their structure is common enough to more than one world to make them recognizable, a means of translation” (p. 393). The authors also commented on the importance of standardization of methods to make information compatible across different social groups (p. 407). Unfortunately, boundary objects are not well understood or easily identified, so their use as a translation tool is not widely implemented.

Although most writers believe that crossing discipline boundaries is difficult, Pierce (1999) argued that the lines are less restrictive than the literature suggests. She conducted a study to describe the characteristics of authors who publish outside their own discipline in order to determine if their publications are read by researchers in the discipline in which they are published. Not surprisingly, her results showed that authors who engaged in boundary crossing were most likely to come from neighboring disciplines, which she defined as “disciplines likely to be working on similar research topics” (p. 278). Interestingly, she also discovered that citation rates for boundary crossing articles showed that they are used by disciplines further afield. This suggested that boundary crossing articles result in complex patterns of interdisciplinary information transfer. Steele and Stier (2000) reported a similar finding. Pierce speculated that the effort required to communicate results across one set of disciplinary boundaries made the articles more understandable to others, too. Some believe the government should provide incentives for those who cross disciplinary lines and are able to use scientific information from diverse sources (Averch, 1985).

If the interpretation and production of written knowledge from other disciplines is challenging, it might be hypothesized that the technical, tacit, and theoretical knowledge required to understand and use data collected from other fields present even greater difficulties. This could be especially true in disciplines, such as ecology, where methods and terminology are not standardized and where they can change over time (Bowker, 2000a; Collins & Pinch, 1998; Latour, 1999). Biodiversity sciences are illustrative of the difficulties that exist when multiple scientific domains, including ecology, converge to use data from multiple sources. The main challenges arise from complexity related to the biological diversity of the organisms themselves, to ecosystems that are in flux, to data gathered at different spatial and temporal scales, and to the sociological diversity of the agencies and groups of people involved (Maier et al., 2000; Schnase, 2000). In regard to the last factor, Bowker (2000b) stated that "a major part of the task of building robust databases in biodiversity is facilitating interdisciplinary communication" (p. 695). Further, he noted that the creation of databases requires attention to the work practices of the communities involved in biodiversity research.

Science as We Know It

Research agendas are driven largely by the questions scientists pose and by the multitude of political and social expectations that impinge on researchers. Societal demands are particularly influential in policy-oriented fields, such as environmental science and medical research (Godin, 1998; Weingart, 1997). A number of potential changes to scientific practice and communication are embedded in the demands for data sharing and for interdisciplinary research. In order to comprehend fully what a new research agenda means to those who practice and those who support science, it is first

necessary to understand the recent past, the current situation, and the projections about the future. For a study of data sharing and secondary use, this requires delving into theories about scientific communication, culture, and practice, and analyzing the factors that might be influencing change in these areas.

William Garvey (1979) described science as a social structure of which communication is the salient feature. By definition, scholarly communication includes social processes as well as scientific outputs, such as publications.

> By scholarly communication, we mean the study of how scholars in any field ... use and disseminate information through formal and informal channels. The study of scholarly communication includes the growth of scholarly information, the relationships among research areas and disciplines, the information needs and uses of individual user groups, and the relationships among formal and informal methods of communication (Borgman, 1990, p. 13).

Garvey described a public and private aspect to science, or what he referred to as formal and informal domains. Communication entering into the creation of science takes place in the informal domain. Here, scientists discuss their work with their colleagues and present it at scientific meetings. Based on feedback and reactions from their peers, they revise or withdraw their ideas. The formal, or public, domain of scientific communication establishes priority and serves as a reliable system for sharing information. Peer-reviewed publications, particularly journal articles, are the centerpiece of the formal arena. The system of scientific communication was set up to fill specific needs related to the creation and reliable transfer of scientific knowledge and has taken a long time to develop. Because it serves such an important purpose, the system is resistant to modification, even by scientists (Garvey, 1979). These qualities make potential changes of great interest and consequence to many parties--a concern that is reflected in

the amount of current writing devoted to the topic. Scientists rely on the system to disseminate and share information, to certify quality, and to dole out rewards. Alterations will affect the nature and practice of science, the formulation of policy, and the organization of institutions, such as libraries, that serve the needs of scientists.

Science as It Might Be

The growth of information technology is a key factor that has led to speculation about the potential for significant changes to scientific culture, practice, and communication. How technology performs this function is unclear. The concept of technological determinism gives precedence to computers as the driving force behind changes in scientific collaboration and communication (Hurd, 2000; Kling et al., 1998). Another thesis is that "the way scientists do things modifies the technology that is used to support the things scientists do" (Sterling, 1988, p. 50). Social informatics is founded on the belief that information technology is best viewed as a socio-technical system in which technology and society influence each other (Bishop & Star, 1996; Kling et al., 1998; Kling, 1999). Societal forces include a problem-oriented research agenda, the belief in the power of interdisciplinary research to solve problems, and the globalization of science (Crawford, 1996; Gibbons et al., 1994). Current discussions on scholarly communication center on the question of whether, and if so, how much, technological and social factors will change the format, conduct, and content of scholarly ideas (cf., Hurd, 2000; Hurd, Weller, & Crawford, 1996; Kling & McKim, 1999; Lindquist, 1998; Lyman, 1999; Lynch, 1993). Writers have prophesized about possible adjustments to both the formal

and informal realms of the communication system. In the section that follows, I review the scholarly literature on technological and social influences to both spheres.

Computer-assisted communication has created the opportunity for scientists to interact with each other regardless of their physical location. Electronic mail, listservs, and collaboratories hold the potential to expand the number of participants, to increase productivity, and to democratize science through greater access to information and to elite researchers, especially for those scientists located at less prominent institutions (Finholt & Olson, 1997; Hesse et al., 1993; Hurd, 1996; Walsh & Bayma, 1996). At the moment, it is difficult to predict if these changes will occur, but some preliminary investigations have been made.

In a study of scientists from four disciplines, Walsh and Bayma (1996) concluded that computer-mediated communication led to an increase in collaboration that was facilitated, if not caused, by computer networks. The authors also observed a rise in the amount of communication, which they believed could help reduce the isolation of scientists who lack colleagues in their institutions. Communication technology also increased participation, but this was expressed as an expansion and not as a leveling. Major researchers and institutions continued to benefit from their status, although networks opened activities to more participants. Lastly, Walsh and Bayma observed that fields differed to the extent in which they were changing, a result that others have confirmed (Kling & McKim, 1999). Hesse et al.'s study of physical oceanographers reached conclusions similar to Walsh and Bayma, although they were more optimistic about the effect of computer networks on researchers at isolated and less prestigious institutions.

On the down side, computer networks might prove to be more of an imposition than a benefit to elite scientists, who may continue to maintain the restrictive boundaries that currently define invisible colleges (Finholt & Olson, 1997). Van Alstyne and Brynjolfsson (1996) painted a bleak picture of the capability of technology to collapse barriers based on geography. New barriers based on interest or time may arise, as obstacles centered on physical location fall. This situation has consequences for the serendipitous sharing of information across discipline boundaries. Van Alstyne and Brynjolfsson noted that the insulation of subpopulations slows the speed at which new ideas propagate through an entire population. If this occurs, it is a significant problem in an era when interdisciplinary interactions are believed to be important.

Although it is assumed to be beneficial, the democraticization of science could profoundly change the production of scientific knowledge. Influential thinkers in the sociology of science have noted that a few scientists do the most important work, and the rest "mop up" (Kuhn, 1970 [1962]; Garvey, 1979; Lotka, 1926). Will more participants necessarily produce more of the information we need to solve problems? Or, will they divert funds from the work of the few scientists who make key breakthroughs (Weller, 1996)? What level of resources should be aimed at supporting those who perform secondary data analysis? These are only a few of the many questions raised by technologically-driven changes to the culture and practice of science.

As many, if not more, speculations have been made about how information technology will affect the system of scholarly communication. The characteristics of print and digital information are different. This has led a number of authors to surmise

that the transition from one to the other will cause the communication system to undergo significant change.

Print publications are permanent, citable, and accessible, and it is possible to make some judgments about quality based on the reputation of a journal or a publisher (Cameron, 1998). Digital information, on the other hand, is ephemeral and dynamic, and quality control is ill-defined (Cameron, 1998; Lynch, 1993). Additionally, the lines between creation and distribution are more difficult to define. Referring to documents that are "born" digital, Peter Lyman (1999) echoed the sentiments of Geoffrey Nunberg (1993) who noted that the computer is not limited to a single role in the production and dissemination of information, and in fact, it tends to erase distinctions between the separate processes of creation, reproduction, and distribution. The user has much greater control, although it is too early to tell if users will be allowed to appropriate digital work and reuse them in new contexts (Lyman, 1999). Roosendaal and Geurts (1999) added that in an electronic environment, the communication process becomes more intertwined with the research process.

Hurd, Weller, and Crawford (1996) saw the shift from print to electronic effecting nearly every participant and component of the scholarly communication process, including manuscript preparation and submission, peer review, publication, information economics, information needs and uses, and the future of the library. Others have hypothesized that the unit of distribution might switch from journals to articles and even to components of articles, such as data, due to the potential for information to become more modular (cf., Bishop & Star, 1996; Bishop et al., 2000; Hurd et. al., 1996; Lindquist, 1998; Roosendaal & Geurts, 1999).

Writers disagree over how much the system will change, but there is a consensus that what we are seeing currently is different from what is likely to happen, and all participants admit that it is difficult to guess what the future will look like. Some authors believe that the system will be revolutionized eventually even though what we are witnessing currently is a modernization of existing processes (Lyman, 1999; Lynch, 1993; Hurd, 1996; Roosendaal & Geurts, 1999). For example, many print journals now have electronic equivalents, but there are few truly electronic journals (Kling & McKim, 1999; Lyman, 1999). In a summary of long-term data, Tenopir and King (2001) observed that the journal system has exhibited surprising stability, and so it remains to be seen whether the system will merely evolve or whether it will undergo a true revolution. Whatever the future holds, most writers believe that the function of peer review to certify quality will remain important and that any changes must incorporate this and other key functions of the system (Hurd, 2000; Tenopir & King, 2001). Additionally, as Tenopir and King (2001) noted, innovations can be a supplement to existing practices and do not have to replace them.

> History shows that a diversity of channels for distribution and publication increases the value of scholarly information. Publishers should not object to web archives, and authors should not abandon journals. Researchers should use multiple distribution channels, including self-archiving and publishing in traditional journals. Journals provide a stable archive of the literature, quality filters and other valuable aspects; web e-print servers allow quick access to more sources of information. Together, they serve the needs of today's scientists for more knowledge from a wider variety of sources (p. 673).

Garvey and Griffith observed that the current scheme of scientific communication is centered about serving the needs and supporting the activities of the productive scientist (Griffith, 1990, p. 40). Thus, changes in this system have important consequences.

Some authors have added another twist to the discussion through their observation that “one size doesn’t fit all” when talking about scholarly communication because acceptable practices vary among disciplines (Hurd, 1996; Kling & McKim, 1999; Kling & McKim, 2000). Hilgartner (1997) made the same observation about scientific data. Personal and organizational goals, reward systems, and sources of financial support differ among participants in the communication system. "These disparities complicate setting directions and resolving problems and assure varying perspectives on controversial issues" (Hurd, 1996, p. 12). They also add to the challenges facing those who provide information services to scholars.

Data Sharing and Scientific Culture, Practice, and Communication

The literature on data sharing has followed several courses since computers first came to play a role in science over 20 years ago. Much of the writing appeared in the government-sponsored reports summarized earlier in this chapter and discussed the challenges, obstacles, and benefits of data sharing. It is only recently that the few systematic inquiries into the effects of data access on the culture and practice of science were undertaken. The lag might be attributed to the fact that the technology to facilitate data sharing has improved greatly in recent years, and laws, policies, and other political and social mechanisms have directly encouraged it. The recent empirical research has focused on the social aspects of data sharing, especially the mechanisms and conditions under which scientists grant or deny access to data, the role of databases in scientific communication and practice, and the function of journals in mandating sharing.

Previously, in many disciplines, raw data were unavailable to people outside the institution or agency that collected them. With the exception of certain government-funded data collections, such as centralized archives of weather and population data, scientific data are not considered "published" and have not been easily available. Under most definitions, data sets are part of the private or informal side of scientific communication. The audiences who might know how to obtain data were also those who had the expertise to understand and use them. If outsiders requested data, access was controlled through an intermediary who could explain the limitations of the data and discuss possible uses (Van House et al., 1998). Unmediated access to scientific data has led some researchers to speculate that databases have already changed, or will soon change, the nature of scientific communication. A proliferation of digital databases could reconfigure sociotechnical networks by reshaping scientists' private and public boundaries (Hilgartner, 1995). Cameron (1998) foresaw a convergence between publications and databases whereby each entity becomes more like the other.

Stephen Hilgartner (1995) wondered about the effect of electronic collections of data on scientific communication. He explored the nature of data banks and speculated on the ways in which biomolecular databases represent a departure from traditional publication practices. Hilgartner believed that bimolecular databases had already created new "communication regimes," the phrase he coined to describe "a sociotechnical regime that constitutes a particular means of scientific communication," such as the journal and the scientific meeting (p. 244). Traditional print publications contribute to regularity in science communication and reinforce a particular social order, whereas novel regimes, like bimolecular databases, have the potential to reconfigure sociotechnical networks. He

compared the stable character of traditional communication mechanisms to the range of roles played by bimolecular databases. For example, print journal articles are public and fixed knowledge that require little maintenance once they are published. The scientific reward process is built largely around this peer-reviewed, print communication. Databases, however, can span the range between published and unpublished, public and private knowledge. Their contents can be corrected and updated, a benefit that can increase the costs to maintain the information. Finally, in many sciences, there is little incentive to contribute data since rewards are not based on this activity. In fact, due to competition for priority, there is often a disincentive to share data that someone else might then use to produce a publication. In the end, Hilgartner noted that database designers are confronted with these issues everyday in the course of their work, and they cannot wait for definitive answers to the theoretical questions posed by databases. Builders of digital libraries and data banks, policy makers, and scientists also need insight into the ways in which the wider availability of data will shape scientific work.

In earlier work, Hilgartner and Sherry Brandt-Rauf (1994) defined data in a broad sense to include "biological materials, reagents, novel instrumentation, and other scarce or limited resources" (p. 356). They argued for a "data stream model," stating that "data should be conceptualized not as the end-products of research, but as part of an evolving data stream" (p. 359). Among other things, they believed the form of the data determines when they are ready for sharing and that different access policies are needed to respond to the different forms. Similarly, Stephen Ceci (1988) pointed out that all data are not the same. For example, numeric data can be copied indefinitely, but cellular material, such as tissue culture, might be finite. Another important factor in a scientist's willing to share

is whether the data are associated with published research results or whether they are unpublished (Campbell et al., 2002; Louis et al., 2002). Goldberg (1997) reported that the temporal phase of the research process--from idea phase to completion of the work--was a key factor influencing the sharing decisions of chemistry, engineering, and physics researchers. All the elements described above are important to consider when developing sharing policies.

A recent series of reports from surveys of academic faculty in the medical and life sciences provided further insight into the intricacies of data sharing behaviors. These studies investigated the withholding of research data, defined broadly to include research results, methods, and biomaterials (Blumenthal et al., 1997; Campbell, Weissman, Causino, & Blumenthal, 2000; Campbell et al., 2002; Louis et al. 2002). In the first study, Blumenthal et al. (1997) asked academic life scientists about their own participation in data withholding in an attempt to discover the extent of data withholding among this group and to analyze the factors that increase the possibility that life scientists will engage in this behavior. In subsequent surveys, the authors pursued further the level of data withholding in this field, probed scientists about their reasons for not sharing, investigated the characteristics of scientists most likely to be denied access to data, analyzed the sharing of pre-publication and published data, studied the influence of research relationships between academia and private industry on a scientist's willingness to share data, and explored the impact of data withholding on the work of scientists (Campbell et al., 2000; Campbell et al, 2002; Louis et al., 2002). Together, these studies illustrate the complicated social and political factors that influence the willingness of scientists to share data.

The work of these writers is important because it described the complex nature and role of data in science, outlined potential motivations for granting or denying access to data, investigated the consequences of data withholding on scientific work, and highlighted the practical and academic reasons that make this a rich and important area for study.

Anecdotes about increasing secrecy in science and an unwillingness to share data led information scientist Katherine McCain (1991) to investigate the exchange of "research-related information" (RRI) among geneticists. She undertook to understand the factors affecting researchers' behavior and expectations as information requesters and as information providers. Like previous researchers, McCain defined RRI broadly, and included in her definition are raw data, computer programs, extensive tables and text too lengthy to include with the published article, craft knowledge necessary for validation and extension of the research, and physical research products (p. 493).

McCain study's showed that most geneticists believe that research-related information associated with published research "should be available to all, with the recognition of the researcher's right to practice private science" (p. 511). Acknowledgment is the expected and acceptable return when an article is published based on someone else's experimental material or technique. Hilgartner (1997) criticized McCain's theoretical window as too simplistic. He argued that a view of science as open or closed, private or public, overlooks nuances present in the system of exchange. McCain's work remains useful, however, since it is among the few studies to explore the sharing practices of scientists. The results of a survey by Stephen Ceci (1988) also found that researchers from a variety of physical and social sciences endorsed "the principal of

data sharing as a desirable norm of science," although the results of a second survey showed that scientists often refused requests for data (p. 54). Recent surveys confirm McCain's findings that geneticists endorse communal behavior, but they also help explain Ceci's conflicting results by showing "evidence of continuing normative ambivalence about openness and disinterestedness" among geneticists (Campbell et al., 2002; Louis et al., 2002, p. 304). Joan Sieber (1988) noted the variety of circumstances that governed data exchange practices and described the need for "a comprehensive taxonomy of data-sharing situations" to help formulate policy (p. 200).

Together, these studies shed light on the attitudes toward the sharing of research byproducts among scientists. They also demonstrated the wide range of research information that scientists share, the conditions under which they grant or refuse access to data, and the expected return for sharing.

My research takes both a narrower and a broader focus than previous work. While it is important to understand the cultural issues related to data sharing, it is time to focus attention on understanding the processes of data sharing and reuse and on learning if the results of each are all they are proposed to be. There is enough voluntary sharing of data to begin considering these questions because in spite of the challenges and obstacles, data sharing does occur and anecdotal evidence and case studies reveal the potential for substantial benefits (Bowser, 1986; NRC, 1995a; Whillans et al., 1990). There is currently a dearth of research, however, that addresses a scientist's motivations for seeking data, that evaluates the existing informal and formal mechanisms that support data discovery, and that analyzes the experience of the secondary user.

Institutional and Technical Mechanisms that Support Data Sharing

The research I discussed in the previous section shows that scholars are beginning to take an interest in the sharing and reuse of scientific data. While researchers have been slow to address the multitude of interesting questions that the topic affords, others have been busy building mechanisms to make data more accessible. I described some of these efforts in the first chapter. In this section, I discuss the scholarly work directed toward some of the institutional and technical mechanisms for sharing data: digital libraries and methods and standards for data organization and retrieval.

Digital Libraries

An assessment of information needs in eight science disciplines showed that the task of keeping up with the staggering growth of research information in all fields would be a significant challenge in coming years (Gould & Pearce, 1991). The major areas of growth identified were journal literature, unpublished information, and primary data. In preparation for the future, librarians were urged to provide better information about primary data and to develop more integrated information environments. Librarians responded to this challenge through digital libraries. Digital libraries are being created and are evolving in the midst of the unknowns associated with the current and future processes of scholarly communication. Librarians have been reminded, particularly so in recent years, that they exist to serve the needs of users and not to build collections (Garvey, 1979; Lyman, 1999; Poland, 1994). Serving the informal information needs of researchers has always been a challenge for librarians. Today, the complexity of doing so has grown, but so has the need.

> The digital library is generally conceived of only as an information resource, as if the library were only a collection, rather than a shared intellectual resource and site for a community. The social functions of the library are not easily measured in terms of outcomes, but are an element in the productivity of faculty and students. Libraries must begin to define and measure their role in productivity (Lyman, 1999, p. 377).

The definitions of a digital library are as varied as the prognostications about the future of the communication system of the scholars they will serve. In summarizing some of the existing definitions, Bishop and Star (1996) observed that many of them emphasize a digital library's intermingled social and technical aspects, which fits with the idea presented above that digital libraries must become places for social interaction as well as spaces for collections. They saw the following elements present in a digital library (p. 308).

> 1.) There is some sense of a collection. An unstructured, unindexed aggregation of documents does not constitute a library.
>
> 2.) The collection is not exclusively bibliographic or exclusively a set of pointers to other materials but includes full-form online material encompassing a range of media and intended uses, such as articles, books, simulations, formulas, datasets, images, etc.
>
> 3.) As in a physical collection, there is a concern to link audience, group, patron, or community with attributes of the collection. However, because of the unique characteristics of online media, collection development may also include group or community development or at least provide a virtual space for linking those with common interests.
>
> 4.) There is in some sense a set of services (human and computer-based) that links collections to those using them and links people to one another.
>
> 5.) The technologies involved are those that support document creation, retrieval, transfer, dissemination, manipulation, and management of the DL, as well as social interactions.
>
> 6.) There is in some sense an institution in which DL collections, services, and social interactions are embedded.

Like electronic and print publications, traditional and digital libraries differ from each other in important ways. For one, authors are likely to become publishers in the

digital environment (Hurd et al., 1996, p. 100). The blurring between the creation and use of information means that librarians must reach out to the sources of information (Lindquist, 1998; Nunberg, 1993). This is a difficult task since the information to be made available is diverse, and some of it is highly complex. The user community is also increasingly heterogeneous, and it is often difficult to know who they are. Librarians may need to relax their quality standards if they are to successfully serve user needs in the digital environment. Research is dependent on many different types of exchange and incompleteness and uncertainty may be part of the process of using digital libraries (Bishop & Star, 1996, p. 315; Palmer, 1996).

The emphasis on the social aspects of digital libraries spurred researchers to develop an area of study called social informatics, the "study of social influences, processes, practices, and effects related to how knowledge is structured and communicated in DLs" (Bishop & Star, 1996, p. 309). Social informatics is an interdisciplinary body of research that has been employed to investigate the design, use, and consequences of technology within many types of cultures and institutions (Kling, 1999). A hallmark of social informatics is that it studies the influence of information technology and society on each other without giving one aspect priority over the other. The rhetoric surrounding digital data sharing projects often employs words such as *informatics* and *bioinformatics* to describe the tools and techniques for storing, handling, and mining scientific data, particularly the large data sets that grow out of genomics research (Sugden & Pennisi, 2000). These terms emphasize a narrower and more technical view than the phrase *social informatics*, which was intentionally worded to signal its inclusion of social aspects into the structuring of knowledge, the building of

infrastructure, and the use of digital resources. (Bishop & Star, 1996). Mechanisms for making scientific data available should be built and improved with social aspects in mind.

Bishop and Star (1996) reviewed several concepts fundamental to their discussion of digital library social informatics. Two of these--genres and the mutability and integrity of documents in a heterogeneous usage environment--are especially pertinent to the study of data in digital environments, including libraries. Genres are conventional forms of structuring knowledge. For example, encyclopedias, journal articles, and poetry represent traditional ways of organizing information. Data sets, too, are a genre. In the past, with some exceptions, such as large, centralized data depositories in the social and physical sciences, data sets have been part of the private side of science. For the reasons discussed throughout this dissertation, scientific data are expected to become more accessible. When data sets are widely available as a separate genre, they may become more recognizable, movable, and mutable. "As technologies for producing, ordering, and disseminating documents change, social practices and institutions will be affected" (Bishop & Star, 1996, p. 335).

The concept of *communities of practice* pervades research on the social informatics of digital libraries (Bishop & Star, 1996; Van House et al., 1998). This idea comes from social theory and refers to the shared understanding, practices, technology, artifacts, and language of a particular group (e.g., Lave, 1988; Wenger, 1998). Each group of users has its own community of practice that affects the way it uses and interprets the information in a shared digital space. In addition, the communities have varied levels of technical skills and expertise. In order to build digital libraries that aid people in their work, it is crucial to understand the general nature and practice of a

community's work so digital libraries can be designed that do not inadvertently "break" those practices (Schiff et al., 1997). When tools and technologies change, as they do frequently in a DL, the practice and artifacts of a community can also change. It is a challenge to continually revise the design and capabilities of the digital library as users alter their work processes. Part of the notion behind the communities of practice concept came from the results of earlier research that demonstrated "one size does not fit all" when it comes to DL interface design (Bishop & Star, 1996, pp. 313-316; Hesse et al. 1993). Thus, researchers have attempted to understand how communities, such as teachers, scientists, and the general public might use the same digital library (Hill et al., 2000). A digital library's content will be different if selected to satisfy a particular community (Hill et al., 2000). Designing for a specific group has the advantage of giving DL creators something to aim for, and it provides the potential for a high degree of success. Hill and her colleagues cautioned, however, that experts in a particular area could have needs that cannot yet be satisfied by the library. On the down side, DLs designed around specific communities can embed the idiosyncrasies of that group into their design. Bowker (2000a) questioned the "communities of practice" trend for understanding large-scale infrastructures, such as biodiversity databases. He cautioned that "classification systems are increasingly being yanked out of their institutional and political contexts, and applied in other fields with different ontologies and associations" (p. 648). Thus, the practices of communities are lost within the databases that are created.

Environmental Data Collections in Digital Libraries

Several projects in the first round of the NSF/NASA/ARPA Digital Libraries Initiative focused attention on scientific data.[3] Of particular interest to my study is the research conducted as part of the University of California-Berkeley Digital Library (UC-B DL). The goal of the UC-B DL was to develop a digital library of environmental information, including text, images, maps, numeric datasets, and multimedia documents to support environmental planning. One objective of the user needs assessment and evaluation portion of the Berkeley project was to provide information about the library's users and their work in order to aid other project researchers with their decisions about content, functionality, and user interface design. The Berkeley researchers examined the use of data sets, one type of artifact, by different communities of practice. This is only part of what they studied, and their findings were preliminary, but their investigation stands alone as an empirical investigation into the use of scientific data, including ecological data, by different communities of users. This made their findings of special interest to my study.

Environmental planning requires the use of numerous types of information, such as text, images, maps, and data. It also involves diverse groups of people, including private land owners, environmental activists, government officials, and scientists from biology, ecology, and hydrology. Each of the groups involved in environmental planning had its own community of practice that affected the way it used and interpreted the information in the shared digital space (Van House et al., 1998). In addition, the communities had varied levels of technical skills and expertise.

Monitoring data are of particular interest to those involved in environmental planning because they can help assess current conditions, detect changes over time, and aid in the development and evaluation of policies and actions. Through their interviews with the various DL user groups, Van House and her colleagues found that data holders were particularly concerned about two potential consequences caused by the greater availability of data: distribution of poor-quality data and misuse of data. Goodchild (1995) noted that in GIS applications "data quality impacts both the communication to the user and the process by which decisions are reached based on that information" (p. 422). The UC-B researchers noted that there are valid reasons for these concerns:

> For example, one source of data on fish populations is boat trips that net fish and measure their catch. The time of year and day, the stage in a species' spawning cycle, weather conditions, type of net, the depth at which the net is trolled, and many other factors affect the catch. A comparison with "last year at this time" and conclusions about trends in the fish population that did not account for these factors could be seriously misleading (Van House et al., 1998, p. 339).

This excerpt illustrates the characteristics of ecological data that make secondary use of them so challenging. The technical knowledge required to work with environmental data was part of the concern over misuse of data. Van House and her coauthors noted that metadata only partially solved this problem. In general, metadata is most useful for an 'insider' who understands the capabilities and biases of particular data (Van House et al., 1998, p. 338). Technical knowledge is also necessary to evaluate data quality. A reliance on poor-quality data can lead to inaccurate results. The Berkeley researchers found that these technical barriers acted as social barriers. They noted that, "Part of the tension in environmental planning is the conflict among communities of practice--people with not only different disciplinary bases and ways of working…but with different priorities and ways of seeing the world…" (Van House et al., 1998, p. 340). Finally, the

UC-B researchers pointed to fertile areas for further study related to the design of digital libraries

> Alternately, this analysis also points to rich areas for development ... How can we structure in several standard views of the same sets of data, so that those with varying needs, interests, and capabilities are assisted? What are the documents, databases, images and other materials that are of central use to both narrow and broad groupings? (Schiff et al., 1997, n.p.).

The researchers at UC-B also attempted to answer whether it was possible to design a digital library to serve multiple communities of practice (Van House, Butler, & Schiff, 1995). This remains an open question and is one that Van House continues to explore in her recent analyses of data from the UC-B DL project (Van House, 2002; Van House, in press). The results of Hesse and his coauthors' (1993) research on computer networks in oceanography suggested that "a generic scientist/user model is too simple to describe network needs and success" (p. 99). Further, they noted that if a "more differentiated view of scientists is necessary, then more attention should be paid to how scientists in the disciplines actually work–what their social structure looks like and what kinds of access they need to access what kinds of resources" (Hesse et al, 1993, p. 99). Van House and her colleagues focused on the latter question. I seek to extend that work by looking closely at the reuse of data by one group of scientists.

Organization and Retrieval of Scientific Data

Institutional mechanisms, such as digital libraries, depend on technical mechanisms to simplify their work. In the case of scientific data, these consist of approaches to the documentation, organization and retrieval of numeric data and to the development and revision of standards to facilitate these activities. Until recently, these

topics received varying levels of scholarly attention. As a result, many of the areas that researchers identified 10-15 years ago remain issues today. The demand for access to scientific data, combined with advances in information technology, spawned a renewed and wider interest in research on these topics. Computer scientists, librarians, archivists, and scientists have historically approached the organization of data in different ways. Today, there is both a convergence and a reanalysis of the methods various groups have used, and a concerted effort is being directed toward the development of standards to describe data and to facilitate their sharing. These efforts build on the groundwork laid by past researchers.

Numeric databases--computer-readable collections of data that are predominantly numeric in nature--present a number of challenges in terms of organization, storage, and retrieval that are different from bibliographic data (Luedke et al., 1977). Writers who believed that numeric databases were a natural extension of library and information services first described these challenges in the pre-Internet era of the 1970's and 1980's. Over 20 years ago, Luedke, Kovacs, and Fried (1977) wrote what is still the key review paper on numeric databases. At the time, they observed a growing interest in numeric data spawned by the use of computers coupled with declining storage costs and improved data gathering through real-time monitoring, and they foresaw a "seemingly endless array of problem areas requiring multidisciplinary solutions and access to broad data compilations" (p. 119). Today, their predictions have largely come to pass. In the late 1970's and most of the 1980's, however, the discussions were limited mostly to research scientists and computer scientists and to librarians with an interest in data dissemination

and access. There have been a few collaborative efforts, but for the most part, the two groups have approached the topic separately.

In 1980, Carter (1980) described national and international activities surrounding numerical data dissemination. Her article noted that the National Research Council's involvement in the development of data compilations for physics, chemistry, and technology dated back to the early part of the 20th century. She also described the Committee on Data for Science and Technology (CODATA), which was formed in 1966 to collect, collate, evaluate, and disseminate international data from the biological, chemical, geological physical, and technological sciences.[4] Most interesting, however, were Carter's descriptions of problems in data retrieval, which she assessed from both the data provider's and the user's view. Many of the challenges she articulated remain problems today. From the secondary user's view, these included adequate documentation to judge the quality of the data, to understand the units of measurement, and to determine the methods used to gather the data. The data provider was challenged by the tasks of educating users and understanding their needs, by the maintenance of the data, and by the need to determine the best methods of dissemination. Many of these issues have not disappeared in the 20 years that have passed since Carter brought them to our attention.

> Determination of end use is important. What are the user needs? Will the user span a broad range of interests or disciplines? How foolproof should the system be? Or how elementary should the steps be in which the data files are to be queried? Is the user going to know what the tolerance or error bar means? There is a tendency for data users to take what is flashed on the screen as gospel, even more so at times than from the printed page (Carter, 1980, p. 150).

Before potential users can begin to make judgments about data, they must find them. Techniques and services to assist users to locate data have been a long, but

sporadic topic of discussion. Citation practices and indexing methods have generally ignored the data content of articles (Luedke et al., 1977, p. 144). White (1982) found that the Social Science Citation Index (SSCI) covered citations to data sets, but the lack of consistency in citing them and the absence of authority control by the SSCI made it difficult to pull these citations together. Over the years, writers called upon journals and authors to follow standard forms when citing data sets (Dodd, 1979; Howard, 2000). Unfortunately, these pleas went largely unheeded.

Numerical data indexing was one solution offered to the problem of finding data (Luedke et al., 1977; Murdock, 1980). Murdock (1980) described numerical data indexing as an attempt to identify, through an index, those publications that contain numerical data. He also discussed "data flags," which an index could employ to indicate the presence of numeric data in an article. Unfortunately, data indexing was too expensive and difficult to be used widely, although Murdock felt that if the benefits could be quantified the chances that data would be indexed would increase. More recently, researchers in library and information science have investigated the retrieval and use of document components, such as passages of text, images, and data (e.g. Bishop, 1999; Buckland and Plaunt, 1998). Another problem that troubled Murdock was the difficulty of standardizing a discipline's vocabulary and symbols so they would be understandable to a wider range of subject fields. Information professionals and data managers continue to grapple with this issue. Ontologies are being explored as a means to overcome the problems of data sharing that arise from semantic diversity. Pundt and Bishr (2002) defined an ontology as "a specification of a conceptualization," and they described the use of ontologies for environmental data supported by geographic information systems

(p. 96). Others observed that the term is often used to refer to common ground for communication within a group (Bénel et al., 2001). Ontologies help users locate information or data about a single concept that can be described using various terms (Adams, 2002, p. 22). Bénel and colleagues (2001) noted that is difficult, however, even in well-defined and formalized fields like medicine and zoology, to get experts to agree on concepts. Ontologies are one component of the Semantic Web, an extension of the current Web in which information is given well-defined meaning, creating a setting whereby software agents can carry out sophisticated tasks for users (Adams, 2002; Berners-Lee, Hendler, & Lassila, 2001). At this time, the implementation of approaches that facilitate communication across disciplines to assist the retrieval of data and information remain elusive.

In the 1970's and 1980's, recognizing the difficulty that users confronted in locating data, a few vocal librarians urged their colleagues to provide access to numeric databases, which they saw as a natural extension of library collections and reference services (Chen & Hernon, 1984). Luedke, Kovacs, and Fried (1977) discussed some of the limited earlier efforts by librarians to catalog numeric data (pp. 150-151). Librarian advocates also recognized, however, that numeric databases were more difficult to use and to support than bibliographic systems because they lacked standardization in structure and software, were more localized and specialized, and quality was difficult to discern (Gray & Dodd, 1984; Luedke et al., 1977). In addition, librarians who worked with these systems often required knowledge of statistical techniques and had to be able to understand the subject content and structure of the database (Chen & Hernon, 1984; Dionne, 1984; Gubiotti, Pestel, & Kovacs, 1984). Gray and Dodd (1984) encouraged

librarians not to be dissuaded by these challenges since they could choose the services they wished to provide ranging from information *about* the database, to direct access to data, to the collection of data, including preservation functions. Gubiotti, Pestel, and Kovacs (1984) predicted that future numeric databases would be required to go beyond retrieval and display to become a tool to help solve problems. Nearly 20 years later, that prediction is now accepted as truth, although our understanding about how data perform this function, if they do, is not well understood.

Bibliographic and Computer Science Approaches

As can be seen from the above, organizing and describing data for access is not a new effort, although it has certainly received more widespread attention in recent years. The wider availability of a diversity of information, including numeric data, has sparked the interest of librarians in providing access to resources that were not formerly part of most library collections. Digital libraries consist of heterogeneous collections of information types, including raw data, which must be unified and linked together in a single resource (Ercegovac, 1999). Yet, it is also necessary to be able to differentiate between these resource types (Frank, 1997). With data, it is not only necessary to locate them, but then once found, users need to know the concepts represented in the data, their quality and reliability, which software systems are able to display and analyze the data, and cost and use restrictions, if any (Frank, 1997). Library science and computer science have different approaches to the description and organization of information, and each method has advantages and drawbacks when it comes to meeting the needs of users seeking to locate, retrieve, evaluate, and use data (Frank, 1997; Burnett, Ng, & Park,

1999). In the section that follows, I discuss the affect of each method on the description and organization of scientific data, and I describe recent efforts to combine the best from each approach.

The bibliographic method has its origins in cataloging and the data management approach has its roots in computer science (Burnett et al., 1999). Library science has been based on a format (the book) instead of a concept (information) (Larsgaard, 1996, p. 18). Traditional library catalogs have served to inform users whether an item exists in the library, and if so, where it can be found (Frank, 1997). Catalogs rely on surrogates and serve to distinguish individual physical items from each other (Burnett et al., 1999; Frank, 1997; Larsgaard, 1996). The descriptive information needed by many data users goes beyond what can be found in library catalogs (Ercegovac, 1999).

The library community has become aware of the limitations of its approach to assist users to locate and use data. To address this weakness, librarians extended the Anglo American Cataloging Rules (AACR2), the cataloging rules standard for libraries, in order to present more useful descriptions of digital files and other networked resources (Frank, 1997). Dodd (1982) discussed the attributes of machine-readable data files (MRDF), including numeric data, which make them difficult to describe under existing library rules. She noted that MRDF lack internal user labels, such as a title page, which traditionally served as the chief source of information for catalogers. In addition, the elements of MRDF are not equivalent to those found in bibliographic works. For example, some MRDF have no recognizable title, and others have more than one. Additionally, production is the rule rather than publication; data are changed easily, so identifying editions can be almost impossible; and physical description is difficult since

MRDF are so different from traditional library materials. In summary, Dodd noted, "the value of cataloging is ultimately proved not by how well each MRDF is uniquely defined, but by how efficiently the user is directed to the needed resource" (p. 157).

Another key library standard is MARC (MAchine-Readable Cataloging). Librarians in the United States follow the USMARC standard. In contrast to AACR2, a cataloging rules standard, MARC is a database format for transferring catalog records between bibliographic systems (Frank, 1997; Larsgaard, 1996). The MARC standard does not fare well in regard to the management needs (intellectual property, preservation) or evaluative needs (authenticity, user profiles) of digital information and data (Ercegovac, 1999).

Computer science approaches to data management and organization, unlike traditional bibliographic methods, "aim not only to store, access, and utilize data effectively, but also to provide data security, data sharing, and data integrity functions" (Ng, Park, & Burnett, 1997, p. 338). Metadata is the term historically used by computer scientists to refer to the documentation that describes data. In its broadest definition, metadata is defined as "data about data," or "information about data" (Goodchild, 1995, p. 418). Metadata does not apply only to data in digital form, although this is how the term is generally interpreted. The purpose of metadata is to help potential users locate data, to determine if they meet their needs, and to provide information for accessing and using the data (Frank, 1997; Milstead & Feldman, 1999a). Recently, metadata standards have evolved that embody the concepts of both metadata and cataloging (Frank, 1997; Larsgaard, 1996; Ng et al., 1997). Bishop and Star (1996) predicted that the variety of information genres present in digital libraries might lead to the merging of systems,

functions, and organizational traditions that were previously separate. Therefore, it makes sense for the traditions of computer science and library science to come together, although it is unclear which approach ultimately will take the lead.

Metadata: A Merging of Approaches

Much energy has gone into creating metadata standards for various user communities (Ercegovac, 1999; Milstead & Feldman, 1999b; Vellucci, 1998). Ercegovac (1999) noted that metadata standards vary in terms of their specificity, structure, and maturity largely because each standard developed based on the needs of a particular user community. Most metadata standards are comprised of data elements, with part of the purpose of the standard being to describe the structures and relationships among the data elements (Frank, 1997). For example, the Dublin Core is a simple 15-element standard; whereas the Content Standard for Digital Geospatial Metadata (CSDGM) is a complex scheme comprised of 300 elements. Milstead and Feldman (1999a) noted that the plethora of standards is the biggest obstacle to its orderly development.

The CSDGM is one of the key metadata standards for scientific data and other data with a spatial orientation. The standard was developed by the Federal Geographic Data Committee (FGDC), an interagency group established by the Office of Management and Budget, in response to the Executive Order that created the National Spatial Data Infrastructure (Federal Geographic Data Committee [FGDC], 1994). The purpose of the CSDGM is to provide a common set of terminology and definitions for the documentation of digital geospatial data in order to coordinate the development, use, sharing and dissemination of geographic data (FGDC, 1994; Frank, 1997; Mangan, 1995;

Milstead & Feldman, 1999b). The FGDC realized that if federal agencies were going to coordinate data collection and sharing, they needed a common standard to discuss and describe those data sets, and so the FGDC began work on such as standard in 1992. The use of the standard is mandatory for geospatial data collected by federal agencies (Mangan, 1995). The standard was developed with the intent that data producers would be responsible for documenting their own data (Frank, 1997).

The CSDGM is an important standard for those working with environmental data. Even though all environmental data, including much ecological data, do not have a strong spatial component, the CSDGM has been used as a base to produce standards that are more suitable to the data collected from fields such as botany and ecology. The biological "profile" was developed to expand the standard to provide elements appropriate for the documentation of nongeospatial, biological resources data and information (Milstead & Feldman, 1999b). Since adequate documentation is considered the biggest obstacle to the secondary use of ecological data, the biological profile might provide a useful standardized method to describe these data. At this point, however, little evaluation has been done to determine the effectiveness of metadata standards.

Recognizing library catalogs as a source of information about data, the FGDC worked with librarians to design the CSDGM to be compatible with USMARC (Mangan, 1995). Upon completion of the CSDGM standard, it was found that USMARC lacked some of the fields needed for digital data, such as the method of storage, the requirements to make use of the data, and the information related to obtaining the data (Mangan, 1995, pp. 100-101). The USMARC standard was modified to include this information, but this

has not eliminated the challenges of adapting elements designed for bibliographic information to digital data (Larsgaard, 1996).

Both MARC and the CSDGM are criticized for being difficult to use (Frank, 1997). One challenge of describing scientific data is deciding on the level of granularity, or in other words, the appropriate level of description (Vellucci, 1998). Michener et al. (1997) provided guidelines based on the knowledge of the secondary user, but these recommendations have not been tested. Data from different scientific fields require descriptive elements that are unique and complex and that are not satisfied by Dublin Core or MARC (Ercegovac, 1999; McGrath et al., 1999; Michener & Brunt, 2000; Milstead & Feldman, 1999a). For example, planetary data require fields related to orbital positions and descriptions of atmospheres and clouds, whereas metadata for astronomy must support searches by wavelength and sky position (McGrath et al., 1999).

Given the plethora of standards that now exist, some attention has been directed to creating crosswalks or maps between the different standards. For example, the CSDGM standard was developed to be compatible with MARC because the FGDC saw library catalogs as a key point for locating information about data. The Warwick Framework is intended to aggregate multiple sets of metadata (Milstead, 1999b; Vellucci, 1998).

The vast amount of electronic information to be organized and described has led some authors to conclude that librarians alone cannot possibly catalog the massive number of existing and yet to be created electronic objects and that the producers of information will need to be responsible for documenting their work (Frank, 1997; Milstead & Feldman, 1999a; Milstead & Feldman, 1999b). This approach presents a number of challenges. Key among these impediments are first to convince authors to

catalog their creations and then to get these millions of non-information professionals to take responsibility to catalog to a certain level and standard (Milstead & Feldman, 1999a). A study by O'Neill, Lavoie, and McClain (1998) assessed the use of Dublin Core metadata, one of the simplest standards, in Web documents and found that most sites were not described adequately nor did they adhere to a well-defined set of metadata elements. Automated cataloging methods to speed up and simply the process of describing data are hoped for in the future (Frank, 1997; Michener & Brunt, 2000).

Despite all the effort that has gone into their creation, little is known about the effectiveness of metadata in helping users locate, evaluate, and use electronic resources. In one of the only investigations, Fraser and Gluck (1999) conducted a usability study to explore how users determine the relevance or potential value of geospatial information from metadata. They analyzed three metadata standards, FGDC, GILS, and MARC, and focused on issues such as the interface, layout, presentation features, and other aspects of display formatting. Fraser and Gluck found that participants placed primary emphasis on ease of use. Data sets that might potentially be relevant were overlooked if the metadata were not usable. For example, if key information, such as that found in abstracts, was buried among other fields, users became frustrated and their assessment of the value of the data was influenced negatively by the poor metadata format. It is recognized that much work needs to be done to evaluate the effectiveness of metadata (Goodchild, 1995; Thiele, 1998). Because the documentation and organization of scientific data sets can be time-consuming and expensive, it is important to evaluate the effectiveness of existing standards in meeting their objectives.

Application to Ecology

The purpose of this section is to compare what applies and does not apply to ecology based on the broad overview of the literature presented in this chapter and on the discussion in the previous chapter. Much of the information that I cover is sprinkled throughout this chapter and the previous one; this section collates and summarizes this material. My comparison addresses data-sharing infrastructure, scientific methods and standards, characteristics of ecological data, and data ownership. These are the factors that make ecology different from fields where the exchange and secondary use of data are more common. I also speculate on the ways in which the questions posed by ecologists are altering their practices and how this might affect the pursuit of ecological knowledge. My discussion blends two ways of looking at ecologists in the context of data sharing and secondary use--ecologists as users of many types of shared data and ecologists as sharers and users of ecological data.

How is Ecology Different?

Upon close inspection, few sciences can be reduced to generalities. However, without generalities it is impossible to draw contrasts between one science and another. Therefore, my discussion compares ecology with other disciplines that generate observational data and that have established infrastructures for data sharing. Many of these fields, such as meteorology and oceanography were mentioned previously. It is also difficult to reduce ecology to generalities because it is comprised of numerous subdisciplines using different methods and working at different scales. Some ecologists perform experiments in a controlled laboratory environment, others conduct experiments

in a field setting, and still others observe phenomena in the natural world. In all cases, though, ecologists attempt to understand living organisms and their relationships with the environment. Whether they work in a field or laboratory setting, ecologists are confronted with variance among individuals within a population and with the unpredictable character of the natural environment.

Ecologists do not have an established infrastructure for sharing data. Sharing occurs on an ad hoc basis and relies heavily on social interaction. The lack of infrastructure is due to the characteristics of ecological data, to a reward structure that does not encourage sharing, and to the fact that many ecological data sets are of interest to a limited number of potential users. Proprietary rights are also an issue. A single ecologist often gathers data, and therefore ecologists tend to feel that they own the data they collect. Observational sciences that do have a foundation for sharing frequently depend on instruments to collect data. This factor diminishes issues of data ownership. Scientific methods are also more standardized in these fields, and the questions posed by scientists depend on shared data resources.

Unlike many other sciences, ecology is not theory rich. At first glance, this might appear to make it easier to share ecological data since they are not as intertwined with theory as data in other disciplines. This difference, however, is negated by the difficulty of working with ecological data because of the lack of methodological and measurement standards. This factor also makes it expensive to document data for reuse. In other observational sciences, storage expenses are greater than documentation costs. The characteristics of ecological data are also part of ecologists' concerns about misuse. In some respects, ecologists have an advantage as users of shared data in that they recognize

the challenges that exist. This may make them more sensitive to possibilities for the misuse of data that are shared with them.

How Will Ecology Change?

Ecologists are subject to political and social demands, and these demands have started already to change the research questions that some ecologists pose. Many questions now require the interfacing of ecological data with data from other fields or with data from more than one ecological study. New statistical techniques for performing meta-analysis and improved methods for modeling and predicting environmental change will also depend on access to data. These factors portend a change for ecologists as both users and providers of data. The characteristics of ecological data will continue to make them difficult to share, and ecologists may never be able to rely heavily on unmediated access to raw data as users or sharers. Ecologists are facing issues that researchers in other disciplines confronted earlier, and they have the opportunity to learn from these experiences. Whether they will accept fully this opportunity remains to be seen. At the same time, other "small science" fields characterized by an independent-investigator approach and heterogeneous and unstandardized data, such as microbiology, soil science, and anthropology, stand to benefit from what ecologists learn (Reichman & Uhlir, 2001, p. 248). Therefore, lessons from ecology have broad implications.

Summary

In this chapter, I reviewed the scholarly literature in areas relevant to data sharing and secondary data use. Much prescriptive knowledge exists about the important

cultural, political, social, and technical issues that affect data sharing, but empirical research is lacking. I employ the available information to inform my investigation of the reuse of shared data by ecologists, an area about which little is currently known, but one which has important implications for the development of data resources and for policy formulation. The availability of previously inaccessible information, speculation about revolutions in scholarly communication and scientific practice, and the need for libraries to reach out to the users of information leads to a desire to add to what is known about the nature and use of research products, such as scientific data, that fall outside the traditional, formal realm of scientific communication. This desire is met by a void in several areas: limited or no baseline knowledge of processes and products, incomplete knowledge about basic variables, and little evaluation of existing data sharing mechanisms. How I address these gaps is the topic of the chapter that follows.

Notes to Chapter 2

[1]According to the Committee on Environment and Natural Resources, the Federal government spends approximately $5 billion per year on environmental research and development (R & D). This includes the areas covered in many of the government reports issued in the 1990s. In the mid-1990s, the majority of these funds were competitively awarded, with about half going to support extramural R & D efforts.

http://www.nnic.noaa.gov/CENR/programguide/Intro.html

[2]There is currently a lot of activity related to issues that surround scientific data sharing. For example, the *Symposium on the Role of Scientific and Technical Data and Information in the Public Domain* took place on September 5-6, 2002. The proceedings are scheduled for publication in early 2003.
See http://www7.nationalacademies.org/biso/Public%20Domain%20Symposium.html.

See also the *Institute of Medicine Conference on Seeking Access to Research Data in the 21st Century*.
http://books.nap.edu/nap-cgi/srchnax.cgi?term=%22seeking+access+to+research+data%22

In April 2002, the National Institutes of Health (NIH) sought comments on a proposed data sharing policy that would apply to all grants, contracts, and cooperative agreements. The draft was finalized on August 1, 2002 and went into effect on January 1, 2003.

[3]In the first part of the Digital Libraries Initiative (DLI), which ran from 1994-1998, the NSF awarded grants to 6 projects. The DLI is currently in phase 2, and it includes several interesting projects related to numeric data. Among these are the University of Pennsylvania's project to record and track data provenance, Harvard University's digital library of social science data, and the University of South Carolina's on-line library of experimental software and data for research in economics and sociology.

http://www.dli2.nsf.gov/

[4]CODATA remains active and has sponsored many interesting conferences on data access. Information on its current activities is available at http://www.codata.org/

CHAPTER 3

CONCEPTUAL FOUNDATIONS AND RESEARCH METHODS

So far, we have seen that in spite of the costs and challenges, data sharing is being encouraged because the benefits of secondary data use are believed to be substantial. A number of mechanisms have been created to make it easier to share data, and U.S. laws and policies promote or require access to research data gathered with federal funds. In places where access to data is restricted, such as within the commercial sector and in some countries outside the U.S., the limitations are based on the economic value of data (Lopez, 1998). As Dawes (1991) summarized, in government, information tends to have more worth when it is shared and used, whereas in the private sector, information is more valuable as a competitive weapon when it is kept proprietary (p. 6).

The force of the claims made about the value of scientific data and the benefits of their reuse to environmental research might lead one to believe that the processes by which these data are located, incorporated into work, and turned into new information are well understood. This is not the case, however. Our knowledge about the secondary use of much scientific data is limited, and the predictions about its benefits are based largely on faith. We understand little about how users locate data, how they blend them into their work, what obstacles they face in doing so, and how data are used to test novel hypotheses and to create new information. These knowledge gaps hinder the formulation

of policies and the design and evaluation of mechanisms intended to help individuals locate and use data.

It has long been recognized that it is important to understand scientific practice and communication in order to support the information needs of scientists and to effectively disseminate the results of research. This is truer than ever in light of the diverse genres of information that are now available. Some of these forms are new, such as interactive, electronic publications, and some of them are traditional, yet more are public manifestations of previously private information, such as scientific data. The availability of formerly inaccessible information, the creation of new, uncertified genres, and the speculation about revolutions in scholarly communication have led to a need to understand more about the nature and use of research products, such as scientific data, that fall outside the traditional and formal realms of scientific communication. Therefore, the time is right to study the experiences of secondary data users.

Conceptual Foundations

The main conceptual foundations for this investigation come from the history, philosophy, and sociology of scientific knowledge. Of primary importance are historian Theodore Porter's (1995; 1999) theory of measurement as a social technology and similar ideas embodied in French anthropologist and social constructivist Bruno Latour's (1999) concept of *circulating reference*. The concept of *communities of practice* from learning theory and the notions of *inscriptions* and *boundary objects* from the sociology of scientific knowledge also play important roles within this framework (Latour & Woolgar, 1979; Star & Griesemer, 1989; Wenger, 1998).

Social constructivists, such as Latour, have attempted to understand the content of science by focusing on how social variables cause scientists to develop certain ideas rather than others. Social constructivists believe that "nature does not determine science; instead…the social behavior of scientists in the laboratory determines how the laws of nature are defined" (Cole, 1992, p. 5). The vastness of the theories of the sociology of scientific knowledge and the variety of views, from positivism to social constructivism, are evident in a review article by Steven Shapin (1995). Porter did not take sides in the argument between positivists and social constructivists about whether science can get at the real nature of things. Like Cole, however, Porter believed that nature poses some restrictions on what the content can be. Social processes might influence the foci of attention, the rate of scientific advance, and the production of new contributions in the lab, but not the actual content of scientifically accepted knowledge (Cole, 1992). In spite of their different beliefs about the construction of scientific knowledge, Porter and Latour identified strikingly compatible notions about the power of standard measurements as communication and reference devices.

Measurement as a Social Technology

Theodore Porter explored the history of quantification. In doing so, he formed a theory of measurement as a social technology. 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.

According to Porter, rigor and standardization in measurement first arose in areas such as calendar and clock time and measures of length, weight, and volume. Rule-

bound, standard measures were a response to the inadequacy of local knowledge for dispersed social and administrative purposes. Ken Alder's (2002) detailed account of the chaos of measures that existed in eighteenth century France and the creation of the meter as the standard measure of the earth, which Alder discovered is in error, illustrates the strength and lasting power of standards. In the industrial age, standard measurements became more important because administrative and social processes took place increasingly outside local contexts. Standards and rules were necessary for transporting information across distance. 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.

Quantification also has the appeal of impersonality, discipline, and rules. Porter defined objectivity to mean the rules of law and not of men. "Objectivity has come to be distinguished first of all by what it leaves out, by the absence of subjectivity" (Porter, 1999, p. 402). This definition implies the subordination of personal interests and prejudices to public standards -- a concept that has become synonymous with the scientific ideal. Porter noted that mechanical objectivity could serve as an alternate to personal trust. For example, standard statistical measures promote confidence where personal knowledge is lacking.

Porter proposed that scientific knowledge came to be expressed in objective and rational language for two reasons. First of all, scientists are subject to external social and political pressures. Rigorous standards help secure the faith of outsiders in scientific results. Applied fields are more open to scrutiny because there is less distance between inside and outside worlds. The more vulnerable a field is to outside criticism, the more

likely it will be to insist on standardization, even where it violates expert judgment. Secondly, the rigid rules for writing research papers and analyzing data are a way to establish a common discourse and to unify weak communities. Numbers, graphs, and formulas are strategies of communication that make it possible for scientists to reach consensus. Of course, the work of scientists is also judged by other formal standards, such as educational degrees and professional practices that exclude amateurs. Additional knowledge, such as personal habits, methods, and background also play an important role in determining a scientist's credibility.

Where standards do not exist, credibility would seem to be paramount. Although this may be true, objectivity by itself cannot settle disputes in conditions of pervasive distrust. The impersonal nature of objectivity has often been confused with objectivity as truth (Porter, 1995). As noted in the previous chapter, scientists disagree for all sorts of reasons. Some of these disagreements are difficult to resolve. For example, basic epistemological differences can lead to different interpretations of the same data (Kuhn, 1970 [1962]). In addition, Porter noted that in some fields, such as agriculture and medicine, informal working methods are nearly impossible to harmonize. However, he saw this as a difference of degree and not of kind (Porter, 1995, p. 32).

According to Porter, most, but not all disciplines rely on objective, standard measures for communication and credibility. In physics, for example, the dynamics of research activities are so self-contained that interactions within the community are responsible for the certification of knowledge. This type of openness and absence of rigid rules is a rare occurrence, however, and takes place only under special circumstances. Porter believed that most disciplines are weak communities that respond

to pressures through objectivity. Kwa (1993) made this point in regard to ecology. Ironically, physics is often used as a model to forecast potential changes in the scholarly communication of other disciplines. If one agrees with Porter, physics is quite unique among scientific communities in that its intense socialization, combined with close personal contacts, allows physicists to operate with much less formality. Therefore, physics may serve as a poor predictor for changes to the communication systems of other scientific disciplines, a situation that others have also begun to recognize (Kling & McKim, 2000; Service, 2000).

Reducing and Amplifying the World

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 Bruno Latour (1999) referred to as *reduction* (p. 70). Reduction allows for "much greater compatibility, standardization, text, calculation, circulation, and relative universality…" (p. 70). The companion to reduction is *amplification*. By reducing the natural world to inscriptions, local knowledge becomes public knowledge, and in this way it becomes amplified. Latour (1999) 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 (p. 30).

As noted above, inscriptions are one means of reducing the natural world into a language that can be transferred outside the local context. Inscriptions are created by

inscription devices, which Latour and Woolgar defined as "any item of apparatus or particular configuration of such items which can transform a material substance into a figure or diagram which is directly usable by one of the members of the office space" (p. 51). This definition reflects the fact that inscriptions have most often been used to describe work in a laboratory setting. This is not surprising since this is the environment in which the majority of science and technology studies have been conducted. However, the concepts that grew out of these studies are applicable to field studies. In fact, Latour (1999) made this point in his written account of field scientists from various disciplines working together in the Amazon forest. "For the world to be knowable it must become a laboratory" (p. 43).

Latour and Woolgar (1979) identified several important features of inscriptions. First of all, once the end product, an inscription, is available “all of the intermediary steps which made its production possible are forgotten” (p. 63). This may be true for the creator of the inscription. An outsider’s faith in the credibility of the inscription would, however, seem to depend partially on the belief that those steps were carried out rigorously by a skilled and trusted scientist. Secondly, inscriptions are seen as direct indicators of the substance under study. This accounts for their ability to bring the world closer while also pushing it away. Finally, inscriptions are viewed as confirmation for or evidence against “particular ideas, concepts, or theories” (p. 63). Others have pointed out that data gathered under one set of theories might be interpreted differently under another set of ideas (Kuhn, 1970 [1962]). Thus, the interpretation of data is not fixed.

In this study, I treat data as *inscriptions*. Suchman and Trigg (1993) summarized the beliefs of Lynch and Woolgar (1990) who proposed that "occasions of scientific

practice distributed in time and space can be aligned, through the juxtaposition of inscriptions from one occasion with those produced on another" (p. 157). The potential for this to occur would seem to be greatest when inscriptions are standardized. Standards help inscriptions act as boundary objects. Van House et al. (1998) noted that "to share data sets is to grant them the status of boundary objects; to determine that they are sufficiently robust to be used across sites and malleable enough to be adapted to local needs" (p. 340). However, the results of their interviews with communities engaged in cooperative work to solve environmental problems questioned the status of data sets as boundary objects. The dissociation of data sets from the assemblages under which they were created and for which their use was intended made it difficult for them to span the distance between and among communities of practice. Van House et al.'s discussion does not mention explicitly whether standards were present, and what, if any, difference they might have made to the sharing of information between communities. Star and Griesemer (1989) found that standards did make it possible for different communities to work together and to share information under a particular set of circumstances present at the Berkeley Museum of Vertebrate Zoology (p. 408). It is not always clear, though, in what situations standardized measurements are effective in granting credibility and spanning distance.

As seen above, theories about communities of practice and situated action have been used productively to understand the knowledge-making activities of scientists (e.g., Suchman, 1987; Suchman & Trigg, 1993; Van House et al., 1998; Van House, in press). Etienne Wenger, who along with Jean Lave, coined the phrase *communities of practice*, described it as a group of people who share an interest in a domain of knowledge and

who develop a set of approaches that allow them to deal with that domain successfully (De Cagna, 2001). The main focus of this theory is on learning as social participation; Wenger (1998) summarized its key points.

> Participation here refers not just to local events of engagement in certain activities with certain people, but to a more encompassing process of being active participants in the *practices* of social communities and constructing *identities* in relation to these communities. … Such participation shapes not only what we do, but also who we are and how we interpret what we do (p. 4).

I draw from *communities of practice* theory to represent the shared knowledge and practice of ecologists. As a broad framework it is useful; I also remain open, however, to its potential limitations. For one thing, like the replication of experiments, it can be difficult to share data within the same community of practice (Collins, 1992 [1985]; Bowser, 1986; Michener et al., 1997). Currently, it is unclear whether this is due primarily to a lack of standards or to some other phenomenon. For another, biodiversity and environmental sciences bring together multiple disciplines, including ecology, and make it impossible to assume a "one-to-one mapping between a classification system and its setting" (Bowker, 2000a, p. 648). Additional research is needed to tease apart the role of the different factors that define successful information sharing both within and outside a community of practice.

Relevance to Scientific Data

The conceptual foundations used in this study are relevant to the sharing and reuse of scientific data in several ways. First of all, the theories of Porter and Latour highlight the importance of standardization to the reuse of data across distance. The use of data outside their original context implies distance. Therefore, standards are important

because they can help span distance and overcome lack of trust. The use of standards is a key ingredient to the success of data depositories that exist already in certain disciplines (NRC, 1995b). These data resources also demonstrate that the word *standard* has many possible meanings. One possible interpretation concerns personal or institutional quality control practices related to data collection and analysis; another pertains to metadata that satisfactorily conveys sufficient information about the data to locate them, to judge their suitability for a particular need, and to use them if they fit that need. One goal of this study is to investigate the range of standards that play a role in the reuse of data.

Likewise, the word *distance* is subject to a variety of interpretations. Most commonly, distance is intended to refer to something outside the local sphere of activity. Examples of this definition include the space between the assumptions and methods of one discipline and another, or the gap among scientists and the general public. Distance can also exist within a community, however, for reasons such as personal or institutional status, subspecialty, or epistemological view. Additionally, the word *distance* can be defined in a temporal sense. For example, there can be a time lag between the original data collection and reuse. Scientists have noted that without adequate documentation to jog the memory, it can be difficult to remember the details of one's own studies (Michener et al., 1997). In a more sobering scenario, the scientist may be deceased, making it impossible to obtain a firsthand account of data collection methods and procedures (Bowser, 1986). Over time, scientific methods, theories, and terminology are also subject to change. For instance, the meaning of basic terms related to water quality measurements altered significantly between the 1940's and the 1980's, which made it difficult for contemporary researchers to reuse earlier data (Bowser, 1986).

The lack of standards has been identified as one obstacle to the sharing of ecological data. This deficiency has resulted generally in a call for greater standardization in the methods used by ecologists and specifically in the development of a metadata standard to describe ecological data sets. As in all sciences, though, tacit knowledge and subjective expertise are important to the practice of ecology (cf., Collins & Pinch, 1998; Latour, 1999; Roth & Bowen, 1999; Roth & Bowen, 2001b). Porter (1995) noted that the focus on objectivity naturally negates the role of subjectivity in the creation and transfer of knowledge. It is unknown how, or if, scientists rely on subjective knowledge in the reuse of data. As Porter acknowledged, standards are only one way in which trust and credibility are established among scientists. If the full range of quality judgments were understood, it might be possible to incorporate some of these means of assessment into information systems.

It is even more difficult to understand what role subjective information plays in addressing environmental problems. Some ecologists believe it has a role, however, as evidenced by a series of articles on Traditional Ecological Knowledge (TEK) published in a recent issue of an influential ecology journal (Ford and Martinez, 2000). Advocates of TEK believe there is a place for the knowledge held by indigenous peoples in helping ecologists to understand the environmental dynamics of a particular locale.

In addition to a lack of standards, the limited and local scale of ecological studies has been criticized, and it is believed that ecology must "scale up" for political, social, and scientific reasons (Baskin, 1997). At the same time, ecology and other field sciences are constrained by the nature of the phenomenon they study (Kuklick & Kohler, 1996; Roth & Bowen, 1999). Ecologists have several options to expand the scale of their

work. They can attempt to increase the development of large-scale, standardized, multidisciplinary monitoring and research programs; they can improve techniques for meta-analysis; or they can create models to help mimic or predict environmental processes. The research agenda of ecology, driven partially by funding, indicates that all these approaches are becoming increasingly popular (Macilwain, 2000). The first option generates large amounts of data while the latter two depend on data for their implementation. Regardless of the approach ecologists choose, these options point to the fact that the organization, preservation, and dissemination of data to fuel ecological research will continue to be a growing issue of concern.

The local nature and limited scale of many field sciences, such as ecology, combined with a dearth of standards, make reuse of data from these disciplines interesting to study because the obstacles to secondary use seem especially difficult to overcome. If data depositories are to be used, they must be effective at spanning distances. In this study, I describe the ways in which ecologists, as members of a community of a practice, currently overcome these challenges, whether through the help of standards, through personal relationships with data providers, or through some other means. My investigation also adds to our knowledge about the information, such as standardized metadata and methods, which might enhance the use of observational data.

Research Questions

The conceptual foundations provided by the work of Porter, Latour, and others were combined with the existing literature, especially predictions about the challenges and incentives of secondary data use, to form several research questions. The obstacles,

incentives, and benefits of data sharing were described in detail in the first chapter of this dissertation. To summarize, obstacles to the sharing and reuse of data can be cultural, financial, legal, scientific, or technical. Incentives to overcome these challenges are provided by funding sources, social pressures, policy and law, research needs, and encouragement from key institutions and individuals. The positive outcomes of data sharing are considered primarily to be socioeconomic, scientific, and educational. Although I do not systematically measure the benefits of data sharing and data reuse, my results provide additional knowledge about the research products that follow from these activities. In doing so, my study paves the way for future research.

My study uses a qualitative approach to address the following question:

- What are the experiences of ecologists who use shared data?

The following subquestions define the specific areas that comprise ecologists' experiences for the purpose of my study.

- How do ecologists locate data?
- What are the characteristics of the data received?
- What information about the data do ecologists receive and/or depend on to use the data?
- How do ecologists assess the quality of the data they receive?
- What challenges do secondary data users face, and how do they overcome them?

For the purposes of this study, I define s*econdary use* as the use of data collected for one purpose to study a new problem. Further, the secondary user may not have collected the data, although he/she may have incorporated some of their own data into the study. In addition, only secondary uses for the purpose of ecological research are examined. This study does not examine the reuse of data for policy or decision-making

purposes. Qualitative research methods are used to investigate the study's research questions. A rationale for this approach is described in more detail in the next section.

Secondary use of scientific data is contingent on sharing. Thus, some of my research questions are directed toward gaining a better understanding of the variety of ways in which data are shared, the tactics that individuals employ to find data, the influence of the data sharing mechanism on the secondary user's experience, and the information that ecologists need to understand the data. Others of my research questions address the importance of standards to establish trust and to overcome distance. In the remainder of this chapter, I describe the methods I employ to investigate the preceding research questions.

Research Methods

Assumptions and Rationale for a Qualitative Study

I selected a qualitative research method for this study for several, inter-related reasons (Creswell, 1994; Powell, 1999; Taylor & Bogdan, 1998). First of all, qualitative approaches are suited to the investigation of topics about which little is known. Thus, they are effective when important variables are unclear and the researcher wants to focus on the context that shapes understanding of the phenomenon being studied. Little direct research or theory exists on the sharing and reuse of data, and this makes it difficult to identify variables or to state research hypotheses. Quantitative methods, on the other hand, are most effective when variables are known and theories exist that can be tested. Secondly, the inductive nature of qualitative research allows categories to emerge rather than being identified *a priori* by the researcher ahead of time. This characteristic makes

it possible to identify patterns or to develop appropriate theories to help understand or explain a phenomenon. This is important in areas where empirical information is scarce. Lastly, qualitative researchers are especially interested in the process of how things occur as well as the product or outcome of the activity. The design and evaluation of scientific data repositories depend on a better understanding of how users locate and make use of genres, such as scientific data. Access to scientific databases in areas such as genomics and biomolecular ecology is believed by some to have changed already the nature of scientific communication and practice (Hilgartner, 1995). It is difficult to test this hypothesis when so little is known about the process of secondary data use.

Like any research method, a qualitative approach also has weaknesses. These shortcomings include imprecise measurement and weak generalizability of findings, vulnerability to several sources of bias, and data overload (Miles & Huberman, 1994). I addressed these limitations in the design of my study by focusing my research questions, by clearly defining the population I studied, and by planning for ways to avoid data overload. I provide more detail on these issues in the remaining sections of this chapter.

Interviewing, the primary data collection method I use in this study, also has strengths and weaknesses. Interviewing is an effective method of data collection when informants cannot be observed directly, when the researcher wishes to study past events, or when the researcher cannot gain access to a setting (Creswell, 1994; Weiss, 1994). In this study, it was not feasible for me to observe ecologists reusing data. Van House and her colleagues (1998) confronted this same problem in their study of data sharing, and subsequently they, too, relied on interviews instead of direct observation. Interviewing is also well suited in cases where the research interests are clear and well-defined and the

purpose of the study is to describe a process (Weiss, 1994). The conceptual foundations I draw from in this study, combined with the discourse about the incentives and obstacles to data sharing, provide a clear focus for my investigation into the process of data sharing and secondary use.

There are several weaknesses to interviewing as a research method (Taylor & Bogdan, 1998). An interview is a particular type of situation, and it cannot be assumed that respondents' words would match their future actions. This limitation is not of particular concern to my study since informants shared their experiences of a past event. The memories of informants can be unreliable, however, and this is a potential weakness for my investigation. Finally, some informants are not as articulate as others, and even if they are, I may be limited in my ability to understand their language since I do not have the opportunity to study in it in its everyday context. I have worked with field biologists for over fifteen years, however, and this experience has provided me with useful insight into the culture and practice of ecologists. In spite of the weaknesses, interviewing is an effective method under the right circumstances. Taylor and Bogdan (1998) recommended lessening the limitations of interviewing by getting to know people well enough to understand what they mean, by creating an atmosphere in which they are likely to talk freely, and by spending time with people "on their own turf" (p. 92). I followed their advice by familiarizing myself with subjects' biographical information, including their professional accomplishments, by attempting to establish rapport through an initial, introductory contact, and by taking advantage of opportunities to visit with subjects in person.

In the past, researchers using qualitative methods often had to contend with the criticisms and suspicions of those who preferred quantitative approaches. Today, it is recognized generally that both paradigms have strengths and weaknesses; these have been described by others and are not repeated here (Miles & Huberman, 1994; Taylor & Bogdan, 1998). What is perhaps more surprising is the degree to which qualitative researchers disagree amongst themselves about basic assumptions and appropriate methods of data collection and data analysis (Miles & Huberman, 1994). Some practitioners of qualitative methods prefer tightly controlled data collection and analysis procedures, somewhat akin to the quantitative paradigm except that the data are made up of words instead of numbers (Miles & Huberman, 1994). At the other end of the spectrum is the open stance advocated by phenomenologists who believe that firm rules are unnecessary because there is no social reality to be accounted for (Miles & Huberman, 1994). In phenomenological studies, human experiences are examined through extensive and prolonged engagement with the people being studied and no preconceived theories, expectations, or frameworks guide researchers as they analyze data (Creswell, 1994, p. 12, 94).

The methods I employ in this study, although leaving lots of room for an inductive approach, most closely follow recommendations made by Miles and Huberman (1994). There are several reasons for this. First of all, the authors noted that an unfocused foray into qualitative research could result in time spent gathering data without a clear idea of what to look for. This is a danger particularly for the less experienced researcher. Therefore, Miles and Huberman recommended the use, if available, of existing theory on which to base research questions or to form initial hypotheses.

Although knowledge about the process of secondary data use is limited, writers such as Porter, Latour, Wenger, and Star and Griesemer have posed theories about the transfer and use of information across distance that is made possible by community membership and by the use of standards, information reduction, and boundary objects. These theories, along with writings about the incentives and obstacles to data sharing and reuse, provided a frame for my study that was useful in deciding who and what would and would not be investigated.

In addition, the purpose of my study was to gather data that would be useful to providers of scientific data and information; the objective was not to conduct a purely sociological study. Therefore, it was important to try to ensure that my research questions would contribute to this goal. This did not leave out the possibility that new questions would emerge along the way or that existing ones would be deemed less relevant. Heuristics are one of the assets of a qualitative method, and Miles and Huberman do not suggest that this strength be overlooked. Finally, my personal bias follows more closely the beliefs of Miles and Huberman than the phenomenologists when it comes to objective reality--processes are socially constructed, but there is some objective reality that can be measured.

Participants

The topics to be explored in a study constitute its substantive frame (Weiss, 1994, p. 15). The substantive frame directs the selection of individuals to be interviewed and the questions to be asked. The research questions I pose in this study served as the frame and guided my method of data collection and the selection of study participants.

My primary method of data collection was semi-structured in-depth interviews with ecologists who reused data. For the purposes of this study, individuals were defined as ecologists if they are members of the Ecological Society of America (ESA), if their institutional affiliation or professional title contained the word *ecology* (or variant of it), or if they identified themselves as ecologists. I identified potential subjects to interview by searching 1999, 2000, and 2001 issues of Ecology and Ecological Applications to locate research articles based wholly or partially on secondary data use.[1] I selected these ESA journals because they have among the highest impact factors for the journals in their field.[2] Ecological Applications publishes research papers that integrate ecological science and concepts with their application and implications, and Ecology reports on research that develops new concepts in ecology or that tests ecological theories.[3] I selected the three-year time period somewhat arbitrarily 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 use experience could be remembered. The lag between the conduct of research and the submission of a manuscript and its subsequent publication means that sometimes several years passed since the data were obtained and used. The best way to counteract this situation was to focus on recently published articles and on research that depended heavily on shared data. Further, I selected papers in order to obtain diversity in the types of data reused, in data sources accessed (i.e., individuals, organizations, and through web sites), and in experience levels of ecologists. I made determinations about data reuse by reviewing the methods and acknowledgments sections of published papers. References to data that appeared in other sections of an article (i.e., introduction, results, and

discussion) primarily supported a statement or verified a fact; they did not constitute reuse.

I interviewed 20 respondents. My main respondents were 13 ecologists who reused data and published the results of that work in an issue of Ecology or Ecological Applications. Although all 13 papers published by ecologists consisted of two or more authors, I did not find it necessary to interview more than one author associated with each paper. In all cases but one, the interviewee was also the first author of the published paper.

Data managers comprised a second group of study participants. This group, which I selected in order to obtain another view of ecological data, consisted of four individuals. I identified these participants based on my own knowledge of ecological data management programs. Two data managers were employed by government agencies, and two worked in research centers associated with academic institutions. Finally, I interviewed three individuals to achieve a breadth of perspectives on the topic (journal editor, NSF program manager, and an ecologist).[4] The results I present in the next chapter focus on my analyses of data from interviews with ecologists and data managers.

Unit of Analysis and Sampling Scheme

The sampling scheme used in this study is purposive. Miles and Huberman (1994) recommended this approach to increase the generalizability of findings from qualitative studies, especially when studies are conducted at multiple sites (p. 37). The ecologists that I interviewed fall into two main categories--experienced ecologists and

less-experienced. Experienced ecologists are defined as those with 15 or more years or more years of experience, and less-experienced ecologists are defined as those with 14 or less years of experience.[5] I divided ecologists somewhat arbitrarily into two categories after the interviews were complete. My rationale was based on three main factors. First of all, it made sense based on the data. My initial plan was to separate ecologists based on years since they received a Ph.D., but as I discuss in more detail in the next chapter, ecologists' identities are tied strongly to their own experiences in the field or laboratory. Therefore, educational level, particularly the number of years since an ecologist obtained a Ph.D., is not the most useful indicator of experience level.[6] Secondly, this division resulted in two somewhat even groups, with a large spread between them. In comparison to those with more experience, less-experienced ecologists clung to notions of scientific norms and were less comfortable with "bending the rules"; I discuss this further in the next chapter. Under this classification, 5 ecologists qualify as experienced (average = 26.8 years), and 8 are defined as less-experienced (average=9.125 years). Finally, computers began to play a more significant role in science in the mid-1980's, approximately 15 years ago. All the ecologists I interviewed, however, were computer literate, and so this rationale was less useful in segregating participants. This is not surprising due to the computer-intensive nature of secondary data analysis work.

Data Collection

There are a number of texts that provide theoretical and practical advice to researchers embarking on a study utilizing qualitative interviews as a method of data collection (e.g. Arksey & Knight, 1999; Kvale, 1996; Rubin & Rubin, 1995; Weiss,

1994). These texts, along with considerations for subsequent data analysis, provided guidance for the methods I employed in this study.

Following my guidelines for participant selection, I made initial choices about individuals to interview for pragmatic reasons. I started my data collection shortly before the 2001 ESA annual meeting, and so I was able to use the conference Web site to identify potential respondents involved with the meeting (i.e. participating in a session, presenting a paper or poster, or serving as a member of a committee). Since I planned to attend the conference, I contacted these individuals first to find out if they would be there, and if so, if we could arrange an interview. Consequently, I interviewed 5 individuals at the meeting; I also met 2 others, whom I later interviewed over the telephone.[7] Of the 5 people I interviewed, 1 was an ecologist who had published a paper in Ecology, 1 was a data manager, and the others were the 3 participants I mentioned previously whom I interviewed to gain a broad perspective on the topic.

I sent a letter to subjects that described my study and requested their participation. Appendix A is an example of the letter sent to ecologists, and Appendix B is a version of the letter to data managers. A short time later, I contacted subjects by telephone. I was rarely successful in reaching potential interviewees by telephone, but I left messages explaining my reason for calling and providing my phone number and e-mail address. Often, subjects would respond to my phone message by sending me an email message. If they did not, I sent an email message to them. Except for the interviews themselves, most of my contact with subjects occurred over email. Once respondents agreed to an interview, I faxed them a consent form, which informed them about the purpose of my study, described confidentiality measures, and explained potential benefits and harms

related to participation in my project. The consent form given to the respondents in my study appears in Appendix C. This form follows the regulations of the University of Michigan Internal Review Board Behavioral Sciences Committee and suggestions made by Weiss (1994). All participants signed the form and agreed to be audiotaped. After each interview, I wrote a letter of thanks to the respondent.

I conducted 20 interviews between June 2001 and February 2002. The cost of travel, combined with the dispersed geographic locations of the interviewees, made it difficult to conduct all interviews face-to-face. However, I arranged to interview half the respondents in person by attending the ESA Annual Meeting in August 2001, by taking advantage of other personal and professional travel to arrange interviews, and by meeting locally with two respondents. I conducted the remaining 10 interviews over the telephone, although as I noted above, I had met two of these participants at the ESA meeting. On average, interviews lasted 90 minutes. The shortest interview lasted 30 minutes and the longest was over two hours. I conducted one interview with each respondent, and I interviewed one author associated with each journal paper. In all cases expect one, the ecologists I interviewed were also the first authors of the published papers. There was no appreciable difference in length or quality of phone versus face-to-face interviews. I taped each interview and had transcriptions made from the tape.

The questions I asked of each group of respondents were based on interview guides; the guides for ecologists and data managers appear in Appendices D and E.[8] The interview guides were based on my key research questions, and they served as the framework for the topics I covered in my semi-structured interviews. The interview guide for ecologists focused on locating, accessing, understanding, and judging data and

on general attitudes toward data sharing.[9] I conducted two pilot interviews to refine my interview guide with ecologists; the second interview subsequently became part of the data analyzed in this study. The interview guide for data managers explored their role within their organization, the types of data they managed, and the challenges these data presented in terms of documentation, storage, and dissemination. It also gathered data on their experiences in working with data collectors and data users and their thoughts about data sharing, standards, and issues of quality.

The researcher is an instrument in qualitative research. I have worked with ecologists for over 15 years in my capacity as a librarian at two federal research centers. Therefore, I have specialized knowledge of the field and of its culture. I prepared for each ecologist interview by reading closely the Ecology or Ecological Applications article and by familiarizing myself with each person's larger body of work and his/her background. I began interviews by preparing subjects for the level of detail I would be seeking through my questions. In addition, I presented myself as an interested and knowledgeable lay person by summarizing briefly the Ecology or Ecological Applications paper that was the subject of the interview. My intent in doing so was to convey my general knowledge of ecologists' work and to provide interviewees with a sense of my scientific knowledge, so they would discuss the data they reused at a somewhat technical level. I checked my perceptions with each interviewee to learn if my assumptions were correct. I felt comfortable asking questions as needed about the detailed aspects of an ecologist's research. During my years of working with scientists, I have found that most are willing and enthusiastic to answer questions about their research.

Data Analysis

The data in my study consisted of transcripts from interviews, various types of documents (curriculum vita, scientific articles, correspondence, etc.), and my own notes and observations. A professional transcriptionist transcribed 19 of the 20 interview tapes following guidelines that I provided (Appendix F). It was often difficult for the transcriptionist to recognize the specialized language used by ecological researchers, and so after the first several interviews I provided a list of terms to accompany each tape; this improved her ability to transcribe scientific vocabulary. When I received a transcript, I listened to the interview and made corrections to the transcript. In some cases, this changed the meaning of the transcribed information. Listening to the tape also refreshed my memory of the interview and familiarized me with the data.

Coding

I developed an initial coding scheme for ecologist interviews based on my research questions and conceptual framework and applied it to transcripts from my first seven interviews. I coded these transcripts manually by going through the interview text and writing the abbreviation for each code next to relevant sections. Based on this preliminary coding, I determined that my codes were too detailed. In my second round of coding, I retained most of my initial code list, but I collapsed subcodes into categories. For example, my initial scheme included 8 attributes for the code related to the challenges ecologists encountered in locating data to reuse. This plethora of possibilities made it difficult to see patterns; essentially, the data were "too spread out." For the final

coding, I imported all transcripts into QSR International's N5 software for qualitative analysis and indexed the data according to the list of codes shown in Appendix G. I also imported and coded the text from each Ecology and Ecological Applications paper that described the data that were reused and the methodology ecologists used to collect them.

Much of the data indexing was relatively straightforward because my codes were descriptive or directly related to my research questions. My remaining concepts were built around my theoretical framework, and although I employed codes related to these ideas in my preliminary indexing of transcripts, the material that ultimately became part of these topics only emerged after extensive note-taking, memoing, and visualization of data. Miles and Huberman (1994) recommended these approaches as means to move from description to higher levels of abstraction, and I employed them all. I also kept a journal to record emerging themes, hunches, and interpretations.

I did not develop a coding scheme in advance to index my interviews with data managers. Instead, I created codes once I saw how data manager interviews functioned within the context of my interviews with ecologists. Miles and Huberman (1994) noted that this approach leaves the analyst more open-minded and context-sensitive, which at times, is advantageous (p. 58). They also observed that since the objective is to match observations to theory, the process is not completely unstructured. I selected data managers as a group to interview in order to obtain a contrast in perspectives, but I did not know at the start of the study what comparisons might emerge. The interview guide for data managers reflected my theoretical framework, and thus it shared concepts covered by my ecologist interview guide. This ultimately made it possible to compare and contrast the data from each set of interviews without creating a coding scheme in

advance. As I analyzed all my data, the topics that emerged as most important were standards, data quality, and documentation. For the final coding, I imported all transcripts into QSR International's N5 software and indexed the data according to the list of codes shown in Appendix H.

Methods of Verification

Methods to determine the reliability and validity of qualitative data differ from those used for quantitative studies (e.g. Arksey & Knight, 1999; Boyatzis; Creswell, 1994; Kvale, 1996). Qualitative researchers do not have a single stance on this topic, although most agree that positivist notions of these concepts are not directly transferrable to qualitative data, and thus, they cannot be determined by the same methods (Creswell, 1994, p. 157).

Creswell defined reliability as the limitations in replicating the study (p. 159). Others have referred to it as "consistency in judgment" (Boyatzis, 1998, p. 144; Kvale, 1996). Qualitative studies work toward reliability by providing detailed descriptions of the selection of subjects, data collection, and data analysis; by the reporting of the researcher's biases, values, and central assumptions; and by checking with the subjects that the researcher captured the essence of their experience (Arksey & Knight, 1999; Creswell, 1994). One of my aims in this chapter is to present detailed information about these topics. To this end, I supplied copies of my interview guides and codebooks in the appendices; together these provide a detailed view of key aspects of my data collection and analyses. I discuss member checks below. Earlier, I related my background in terms of my experience in working with ecologists. My interest in this research topic grew

directly from this experience. Over the years, I noted an increase in policy- and science-based demands for data sharing. I observed that some scientists I worked with were interested in unmediated sharing, but most were concerned about potential misuse. At the same time, federal agencies and other organizations were building systems to share data. I saw this dichotomy, and I became interested in examining the questions it raised. I approached this research with few preconceived notions about possible answers to these questions, and those assumptions I started with were informed by the literature I covered in the previous two chapters.

Validity refers to the accuracy of the information and whether it matches reality (Creswell, 1994, p. 158). Among the most common methods to determine validity are triangulation and member checks. Triangulation attempts to find convergence among other sources of information. Arksey & Knight (1999) describe four types of triangulation: methodological, data, investigator, and theoretical (p. 23). Methodological triangulation, the use of a research design that draws on a variety of methods to collect and interpret data, is the most common. Multiple methods of data collection were difficult to employ in my study. I could not discern any obvious methods to use in conjunction with interviews to address my research questions. The documents gathered as part of data collection were useful, but they could not be said to serve as a method of triangulation. To ensure external validity, I aimed to provide thick description and a comprehensive description of the methods used so other researchers can determine whether the findings can be compared with their own studies. I also employed negative case analysis to test mini-theories I developed during data analysis. Taylor and Bogdan (1998) noted that this method of analyzing outliers helps to refine interpretations (p. 152-

154). In my analysis, for example, I observed a seeming contradiction in the way ecologists spoke about their methods for understanding and judging the data they sought to reuse and the means by which they described their overall approach and perspective. Data comprehension was based largely on informal knowledge gained through ecologists' own fieldwork, but they framed other aspects of their experiences in a formal scientific manner. When viewed within the frame of formal and informal knowledge or the public and private sides of science, this contradiction was resolved. Ecologists recognize the importance of informal knowledge they gain the field, but they rely primarily on formal notions of scientific practice to frame and direct their approach because informal knowledge is not acknowledged publicly in the context of "real science."

Data triangulation, the use of diverse sources to explore the same phenomenon, and theoretical triangulation were the types most applicable to this investigation. I conducted interviews with a number of ecologists and with others involved in the secondary use of data. Multiple perspectives of the same experience help add validity to my findings. For example, my interviews with data managers, who work closely with scientists, helped to confirm what ecologists told me their experiences. Additionally, I tested my interpretations with the findings of others and against my conceptual framework; these portions of my analyses are presented in detail in the next chapters.

Member checks give informants a chance to react to what has been written (Taylor & Bogdan, 1998, p. 159). As mentioned above, they also improve study reliability. I selected a subset of my interviewees to carry out member checks. This gave the subjects of this investigation an opportunity to read and respond to my interpretation of their experiences. The member checks did not significantly change my results, but

they did enable me to ask additional questions, to clarify points I was uncertain of, and to help validate the facts of an individual's experience. Member-check documents were 5-6 pages long and were shared via email.

Summary

In this chapter, I reviewed the conceptual foundations and research methods used in my study. My main conceptual foundations are taken from the history, philosophy, and sociology of scientific knowledge and from learning theory. Of particular importance are *communities of practice* theory and concepts of *measurement as a social technology*, *circulating reference, inscriptions, and boundary objects*. Although these specific terms and phrases were not part of my interviews, I employed these concepts to form research questions directed toward achieving an understanding of the experiences of ecologists who reuse data; I used qualitative research interviews to examine these questions. The results of my investigation are reported in the chapter that follows.

Notes to Chapter 3

[1]Complete descriptions of the editorial policy for each journal are available at:

http://www.esapubs.org/esapubs/journals/applications_main.htm
http://www.esapubs.org/esapubs/journals/ecology_main.htm

[2]The Institute for Scientific Information's 2000 Journal Citation Report: Science Edition ranks Ecology seventh and Ecological Applications ninth among journals in the ecology subject category as measured by impact factor.

[3]In 2000, both journals increased the number of issues published. Ecological Applications grew from 4 to 6 issues. Ecology expanded publication from 8 to 12 issues.

[4]I interviewed one ecologist about his views on data sharing. I mistakenly identified a paper he published in Ecological Applications as being based on data reuse. This interview was enlightening because it dealt with historical ecology, which relies on data that are somewhat unique, and because the respondent is a leader in the field.

[5]Years of experience are calculated based on each ecologist's answer to the following question: Approximately how many years have you been an ecologist? Interviews with the 13 ecologists took place between June 2001 and January 2002.

[6]All ecologists have doctoral degrees except for one participant who is currently working toward her Ph.D.

[7]I had made arrangements to interview these 2 ecologists at the meeting, but I was unable to do so because of schedule changes.

[8]The questions covered the same topics, but each interview varied slightly to match individual situations. For example, each ecologist reused different types of data that they obtained from varying sources.

[9]Based on advice presented in several qualitative research texts, I included several "quantitative" questions in my ecologist interview guide. Except for the question about an ecologist's years of experience, I asked interviewees these questions selectively, and overall, they were of limited value. After speaking for an average of 90 minutes in a semi-structured fashion, it was difficult for me and for the interviewees to switch to this mode of questioning.

CHAPTER 4

FINDINGS AND INTERPRETATIONS

Introduction

An ecologist's reuse of data is the boundary between data gathered at different times in different ways. The stories of ecologists' experiences as secondary data users begin at varying points. Some ecologists are provided with data, and others must find them. What binds ecologists' experiences together is not the exact origin and progression of their journeys, but the knowledge, shaped by their fieldwork, that ecologists carry with them and that they employ to reuse data.

Roth and Bowen (2001b) observed that fieldwork experience has a formative function in evolving the formal (academic) and informal (anecdotal) knowledge of field ecologists (p. 479). They noted that the physical experience of working in the field "shapes the perceptual 'lens' brought to nature by ecologists giving them a unique understanding and forming the basis for membership in the discipline" (p. 460). My findings confirm Roth and Bowen's conclusions about the importance of ecologists' experiences in the field. Additionally, my results show that knowledge gained in the field transfers to ecologists' use of data they did not collect themselves. Ecologists exercise informal knowledge gained through fieldwork, along with formal knowledge of their discipline, to understand and judge data, two closely linked processes. The ability to understand data is the key to their reuse, and ecologists depend on the presence of

information that allows them to put their field-based knowledge into play to comprehend data. While ecologists recognize the importance of knowledge they acquire in the field, they rely primarily on shared notions about norms of scientific practice to guide their search for data to reuse and to frame their experiences because informal knowledge is not acknowledged publicly in the context of "real science" (Roth & Bowen, 2001b, p. 477).

Knowledge for Data Reuse: Interplay Between the Social and the Individual

The groundwork for ecologists' experiences as secondary data users is laid before the process even begins; ecologists' knowledge and experiences form the backdrop to their stories and serve as the base from which they make decisions regarding reuse. Hjørland and Albrechtsen (1995) differentiated between theoretical frameworks that emphasize knowledge as a social or cultural process and those that view knowledge as individual mental states (p. 409). They argued that information science should be seen as a social science, and they supported the domain-analytic approach.

> The domain-analytic paradigm in information science (IS) states that the best way to understand information in IS is to study the knowledge-domains as thought or discourse communities, which are parts of society's division of labor. Knowledge organization, structure, cooperation patterns, language and communication forms, information systems, and relevance criteria are reflections of the objects of the work of these communities and of their role in society. The individual person's psychology, knowledge, information needs, and subjective relevance criteria should be seen in this perspective (p. 400).

The authors recognized, however, that "there is an interplay between domain structures and individual knowledge, an interaction between the individual and the social level" (p. 409). My findings confirm the importance of a social science perspective in analyzing the secondary use of data by ecologists. However, my results also show that in order to

understand fully the data reuse experiences of ecologists it is necessary to consider both social and individual aspects of knowledge.

The study of nature is beset with uncertainties (Roth & Bowen, 2001a). For example, even with a field guide and specimens in hand it can be difficult to distinguish one tree from another because a field guide cannot show all the possible variations in leaves, bark, or structure that occur over time. Ecologists rely on knowledge and strategies that help them reduce uncertainty when conducting their own research **and** when they use data they did not collect themselves. The secondary use of data requires ecologists to call upon all aspects of their knowledge--domain and individual--to deal with uncertainty.

As members of a community of practice, ecologists share an interest in a domain of knowledge and a set of approaches that help them to 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, p. 47). Knowledge of their domain, which they acquire as part of their enculturation to the field, permeates ecologists' experiences and directs the choices they make throughout the data reuse process. It also serves as a standard that they draw from to reuse data; domain knowledge is the public form of knowing that ecologists rely on to span the distance between data gathered at different times, in different ways, and for a multitude of purposes. Porter (1995) noted that standards imply objectivity, which he defined as "knowledge independent of the people who make it" (p. ix). The aspect of domain knowledge that figures most prominently in ecologists' reuse of data is informal knowledge gained through fieldwork. Throughout this chapter, I

emphasize the insights ecologists acquire through fieldwork because it plays the key role in their secondary use of data. Formal disciplinary knowledge and standards of scientific practice are also important to ecologists' data reuse decisions.

Additionally, Porter observed that objectivity is distinguished by a lack of subjectivity. Subjectivity implies knowledge that is personal and local, and thus, it creates distance. Hjørland and Albrechtsen (1995) noted that one way to define individual knowledge is as cognitive processes "isolated from the social context and the developmental history, from which the cognitive processes are created" (p. 409). My results show that in the context of data reuse by ecologists, individual knowledge consists of unique and personal insights and connections that lead to trust and distrust of data that affect ecologists' decisions about what data to reuse. Trust and distrust spring from the same sources and are based on first-hand acquaintance with another's skills, on faith in another's reputation, and on perceptions of the skills or values of other scientists and of "the way things work." Individual knowledge is subjective in the sense that it is particular to each person; it consists of insights and perceptions that are not shared on a wide scale. Individual knowledge is not perceived by ecologists to have the objective status of domain knowledge. Although ecologists employ individual knowledge to reduce uncertainty, it is not openly discussed, and therefore, it does not become part of domain knowledge. Although the line between domain and individual knowledge is sometimes fuzzy, the general distinction is important because it has implications for the development of data sharing policies and resources.

Authors highlight personal connections and knowledge as an important component in the sharing and reuse of ecological data; my results clarify and delineate its

role. Specifically, ecologists employ individual knowledge to include or exclude data from consideration, to lessen concerns about data quality, and to improve their access to sources of information that help them to understand data. Finally, individual knowledge has the capacity to communicate messages about the values and skills of others that steer ecologists toward or away from sources of comprehension. While individual knowledge can help to reduce uncertainty, it plays a secondary role in data comprehension, and therefore, it is a subordinate factor in data reuse decisions.

Ecologists' data reuse experiences are determined largely by aspects of domain knowledge, particularly informal knowledge gained through fieldwork. However, their approaches and their decisions are also influenced by individual knowledge and criterion, such as tolerance for uncertainty. Therefore, two different ecologists will not necessarily make the **same** decisions about the **same** data since judgments about data quality are based on knowledge that consists of social and individual dimensions. Informal knowledge is difficult to build explicitly into formal data sharing systems, but ecologists recognize it as important. Actions based on individual knowledge, however, are hard to predict or imbed in data sharing systems because they are unique, idiosyncratic, and hidden. Ecologists' choices are influenced by the combined use of their domain and individual knowledge, by personal tolerance for uncertainty, and by the complexity of the data they reuse. The exact combination of these factors differs for each individual, and thus, not all ecologists make the same decisions or follow the same path.

This chapter is about distances--near and far--their causes, their ramifications, and their potential solutions. In the first part, I analyze ecologists' domain knowledge, and I show how they employ it to reuse data. In particular, I examine closely the field-based

insights that form the basis for ecologists' reuse of data. In addition, I discuss briefly the concept of individual knowledge. Second, I illustrate how ecologists employ all their knowledge to overcome some of the challenges associated with secondary data use. Ecologists carry with them knowledge to assist their choices about where to look for data, to aid their comprehension of data, to reduce or to eliminate concerns about data quality, and to integrate data from multiple sources. Ecologists' abilities to use their knowledge in an anticipatory way mask some of the real considerations that pervade the sharing of data among members of the same discipline and hide the frequently invisible relationships between the data that ecologists acquire for reuse and the methods they use to gather them. Ecologists select methods to locate and obtain data that work in concert to help them bound their collection of data, that increase their chances of obtaining data, and that reduce the risk of errors associated with data reuse. Ecologists' facility in drawing from their store of knowledge obscures the conscious and subconscious rationale for many of their choices. Acquiring data, understanding them, and assessing their quality can occur simultaneously and are often part of an iterative process. Additionally, the data reuse process is typified by adjustments and accommodations as opposed to simple decisions about whether or not to reuse data.

In the first portion of this chapter, I show that in many ways, ecologists' methods are effective at meeting inherent data sharing challenges. Domain knowledge is a powerful base for sharing data among members of the same community, even though it is unable to completely eliminate issues of trust and distance. However, increasing the amount of available data and scaling up the infrastructure for sharing ecological data require approaches that overcome impediments stemming from cultural, technical, and

social factors. In the second part of this chapter, I examine the limitations of ecologists' methods, and I analyze the role of data managers, one type of intermediary, in helping to address some of these limitations. Following Markus (2001), I define *intermediaries* as those who prepare data for reuse by eliciting, organizing, storing, sanitizing, and/or packaging data, and by performing various roles in dissemination and facilitation (p. 61). Ecologists and data managers have variant goals for their work and different standards, which create distances that must be overcome to improve mechanisms for sharing data.

Dual Roles

Ecological researchers who reuse data are users of existing data as well as generators of their own data. To understand ecologists' experiences, it is important to recognize both roles. To use data means to carry out a purpose or action by means of those data. However, data cannot be used until they are understood. Thus, understanding precedes and is vital to reuse. My results show that as users of data, ecologists are attentive to understanding the methods that other scientists or institutions employed to generate data in order to insure the quality of their own work. Ecologists strive to comprehend data they reuse at the same standard as data they collect in the field or laboratory themselves. Authors acknowledge the significance of the social issues of data sharing, such as ownership, rewards, and cultural norms, and they are important. The literature also recognizes that data are a commodity, but data are also a liability, especially when they are reused. The latter explains partly why personal interaction and networking are prevalent in fields such as ecology. Since data are the basic building blocks of scientific argument, researchers must understand them, or they risk

misinterpretation based on inappropriate use of data. Second, as generators of their own data, ecologists are aware that other scientists will examine the methods they use to collect and interpret data, and so they work hard to "make their measurements demonstrably rational and accountable" (Roth & Bowen, 1999, p. 744). Thus, as with any research they undertake, ecologists' experiences are influenced by their awareness of future peer scrutiny. This cognizance, combined with ecologists' individual knowledge and standards for research practice, affects the choices and adjustments they make throughout the stages of secondary data use.

The dual role that ecological researchers play as secondary data users leads two things to happen to data in the process of reuse. First of all, data are *reconstructed*, a word that my interviewees, ecologists and data managers included, used to describe the process of comprehending data collected by others. I define the term *reconstruction* more broadly to describe all the processes ecologists employ to mentally reassemble the original collection of the data they seek to reuse. My definition encompasses the stages involved in finding, obtaining, comprehending, and judging data. At first glance, the two initial steps may not appear to be associated with reconstruction, but the knowledge that ecologists carry with them and that enables them to understand and to assess data also helps them to locate data and provides them with conscious and subconscious strategies for acquiring them. For the ecologists in my study, reconstruction often entailed a mental visualization of the original data collection process. Ecologists, like scientists in many fields, uphold the ideal of replication of research results as a test of quality, but replication is frequently difficult or impossible to achieve. Even when it is possible to repeat observations, it is often not practical, nor is it culturally expected, and so it is not

typically done. In ecological experiments, replication is difficult due to factors such as heterogeneous experimental units, natural and human disturbances, and the difficulty or impossibility of locating candidate sites for replication (Michener, 2000, pp. 15 & 142). Thus, understanding the way an experiment was conducted or how observations were made is more important than repeating a study. Ecologists' mental reconstructions aid their understanding of data by helping them to determine the fitness of data for their purposes and by assisting them to assess measures of quality, especially potential points of data collection error. As I show in this chapter, insights gained from ecologists' own fieldwork plays a key role in their ability to reconstruct the data they reuse.

Secondly, when data are reused they are regenerated. I use the term *regeneration* to mean that data are collected again, this time not from the field or laboratory, but from one of the various places in which they might reside since their original collection. For example, the data could be located in a publication, a handwritten table, or in a publicly available database. Data are regenerated in the sense that they become part of a new study that involves its own data collection. Reconstruction is a key step in regeneration. In regeneration, it is an ecologist's prerogative to transform the data. Regeneration is characterized by accommodations and adjustments to the data available for reuse. The result of regeneration is a new data set upon which novel calculations are made, comparisons are based, or theories are examined.

An Introduction to Conventions, Terms, and Interviewees

In the remainder of this chapter, I show how ecologists harness all aspects of their knowledge to successfully reuse data, and I analyze the limitations of their methods and

the role that data managers can play in helping to scale up the infrastructure for ecological data sharing. I begin by providing background information about the data ecologists reused and the means by which they acquired them. Before proceeding, however, it is necessary to describe some stylistic choices I made in an effort to increase readability. The first of these choices relate to my definitions for terms and phrases that occur frequently in this chapter and the second concerns my presentation of interview excerpts.

Ecologists refer to the thirteen ecologists that I interviewed about their secondary use of data. Other individuals that I interviewed are also ecologists, but my intent in interviewing them was related primarily to another role they play in their professional lives, i.e. journal editor, science administrator, etc. My interviews with ecologists focused on papers they published in an issue of Ecology or Ecological Applications in 1999, 2000, or 2001. The term *case* refers to each of the instances of data reuse by one of the thirteen ecologists. *Data managers* describe the four individuals I interviewed for their professional expertise and skills in managing data. The first part of this chapter focuses on interviews with ecologists; the second half draws primarily on interviews with data managers. Table 1 includes pseudonyms for each ecologist along with other brief information, including each ecologist's affiliation at the time they reused the data reported in Ecology or Ecological Applications.

Besides the use of pseudonyms, I take precautions throughout this chapter to protect the identities of my participants. Insofar as possible, I retained the technical details of the data used and the ecological processes studied in order to maintain scientific believability. However, I changed geographic locales and ecological features, such as

river names, and I generalized species and topics studied. Further, in order to make it difficult to connect scientific details with published papers, I did not link pseudonyms to papers published in either of the two journals or to particular years of publication.

Ecologist	Gender	Years as an ecologist	Year Ph.D. received	Affiliation
Alan	M	29	1997	Government
Andrea	F	13	2001	Academic
Bill	M	25	1979	Academic
Cal	M	12	1994	Academic
Charles	M	7	2000	Academic
David	M	20	1995	Academic
Ellen	F	6	1998	Government
Katherine	F	11	2000	Academic
Michael	M	7	2000	Academic
Nancy	F	25	1980	Academic
Susan	F	11	2000	Academic
Stephen	M	35	1970	Academic
Tanya	F	6	expected 2002	Academic

Table 4.1. Ecologists' Backgrounds and Pseudonyms

Interview excerpts represent the data I provide as evidence for my conclusions and interpretations, and they are intended to embody respondents' points. My presentation of quotations follows suggestions made by Weiss (1994), who summarized the approaches taken by most social scientists that make excerpts easier to grasp without altering a respondent's meaning (pp. 191-200).

> They are likely to permit themselves to eliminate words, sentences and paragraphs--and also, most of the time, their own questions--in order to achieve a more compact statement. They will bring together in one place material dealing with the same issue that originally appeared in different sections of the interview transcript. They will standardize the slurrings of colloquial speech: "I was gonna" would be rendered as "I was going to." But never is a word changed, never is a word supplied (p. 194).

In presenting quotations, I eliminated most conversational spacers, such as "you know;" I corrected colloquial speech and grammatical errors; and I eliminated interviewer questions. My convention differs slightly from the advice offered by Weiss in dealing

with the collation of material on similar subjects. When I merge quotations on the same topic that appear in different sections of the transcript, I note this by the insertion of: *(segment cut)*. I chose this convention because it represents an analytical decision on my part that these interviews portions are related and that they support a particular idea.

Locating and Acquiring Data for Reuse

In this section, I describe briefly the methods ecologists used to locate and acquire data, the sources from which they obtained data, and the types and characteristics of the data they reused. In a later section, I analyze the rationale for ecologists' choices concerning the data they acquired for secondary use and the sources they relied on to obtain those data.

Defining data and describing what is shared becomes complicated quickly. As Brunt (2000) stated, "Where data end and metadata begin is often the subject of much discussion" (p. 37). Haila (1992) made a distinction between quantitative and qualitative ecological data and defined quantitative data as "systematically collected observations amenable to analysis and interpretation" (p. 233). I define data broadly to include all measurements and observations, along with the information relevant to the data that are independent of the data themselves. I chose this definition because data are incomprehensible without the information required to understand them. Examples of pertinent supporting information include methods used to obtain an observation or to conduct an analysis or 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 (Porter, 2000, p. 62).

The ecologists that I interviewed acquired a wide variety of data from a diverse array of sources. In addition, ecologists often collected multiple types of data as well as data from more than one source for use in the same study. In each case, I identified data of one or more types or data from one or more sources that were most critical to each research project. I refer to these as the *key data*. Key data were the focus of each interview, although I asked questions about all the data an ecologist acquired. Table 2 describes the key data reused by the ecologists I interviewed.

Ecologist	**Chief source(s) of data**	**Key data**	**Primary method(s) of locating data**
Alan	Bird-banding database Natural history museums Birding journals	Bird observations Bird weights	Letters to individual birders and to museums
Andrea	Peer-reviewed publications	Plant, soil, and water chemistry data	Literature search
Bill	Peer-reviewed publications	Animal population density (birds, insects, & mammals)	Literature search
Cal	Biological control database	Instances of biological control reported in the literature	Read or heard about the database (couldn't recall exactly which came first)
Charles	Natural history museums	Amphibian species observations	Requests made to museums
David	Historical stream survey	Observational stream data	Another scientist
Ellen	Forestry database	Observational forestry data	Another scientist
Katherine	Peer-reviewed publications	Plant experimental data	Literature search
Michael	Peer-reviewed publications	Lake zooplankton data	Literature search
Nancy	Peer-reviewed publications	Plant experimental data	Manual searches of particular journals for a specified time period
Susan	Two databases containing lake chemistry data	Water chemistry data	Another scientist
Stephen	Personal familiarity with research programs	Lake phytoplankton data	Personal connections and knowledge
Tanya	Tree-ring database Climate database	Tree-ring data Precipitation data	Another scientist (tree-ring data) "Common knowledge" (precipitation data)

Table 4.2. Key Data Reused by Ecologists

Ecologists accessed and received data in both electronic and print forms, and they created their own data sets from the data they collected. While many ecologists gathered similar types of data in small amounts from multiple sources, several ecologists used existing data sets. The National Research Council (NRC) (1995b) defined "small-volume data sets as those with volumes that are small in relation to the capacity of low-cost, widely available storage media and related hardware" (p. 17). According to this definition, the data sets ecologists created from the data they reused are small. However, several of the data sets from which these data were obtained are large according to the NRC in that archiving cost, longevity of media, and maintenance of the data holdings are dominant considerations.

Data sets are the focus of data sharing efforts and of metadata standards developed to enhance sharing.[1] With all the emphasis on documenting data sets, I assumed at the start of this study that entire data sets are what would be shared. What I found in my very first interview, however, is that the ecologists I spoke with often gathered a particular type of data from multiple sources, and the data they acquired were only a small portion of what the original collectors had gathered.[2]

Three of the ecologists conducted a meta-analysis, five ecologists acquired observational and/or analytical data from multiple sources, and four ecologists used existing observational data sets.[3] By definition, meta-analysis implies the use of multiple data sources. 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, p. 156; Smith, 1996). The data gathered for a meta-analysis are typically acquired from the published literature (NRC, 2002, p. 8). Eight

papers 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.[4] Over half of the ecologists (n=7) discussed other research they conducted that necessitated the secondary collection of data, and I have incorporated their observations about these experiences into my overall interpretations and conclusions.

Among my interviewees, public sources of data and information about data included museums, published literature, bibliographic databases, and databases available on CD-ROM, over the Internet, or through a public data center. Almost half (n=6) of the ecologists used the published literature as a main source of data.

The literature states that most data sharing among ecologists takes place on an ad hoc basis. Based on this, I assumed that there would be lots of personal interaction throughout the reuse process, including data collection. While personal networking played a role in many experiences, it was not the only way that ecologists in this study located data. They used bibliographic databases, the published literature, and the literature-cited sections of papers. They wrote letters and sent e-mail messages to unknown individuals and institutions they identified as potential data sources, and they used the Web. Less experienced ecologists sometimes relied on others to help them locate data.

At the end of each interview, I asked each ecologist how many years he/she had been an ecologist. I left it to each ecologist to choose the point of reference from which to calculate his/her years of experience. Based on this, as I described in the previous

chapter, 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. Table 1 describes each ecologist's experience level at the time of the interview, and his/her affiliation at the time of data reuse. Under this classification, five ecologists qualify as experienced (average = 26.8 years) and eight are 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. However, as I discuss in more detail in the section that follows, ecologists' identities are tied strongly to their own experiences in the field or laboratory. Therefore, educational level, particularly the number of years since an ecologist obtained a Ph.D., is not the most useful indicator of experience level. For example, Alan and David completed their doctoral degrees in the last five and seven years, respectively, but each has 20 or more years of professional work experience.

Domain Knowledge and Data Reuse

Fieldwork is important in shaping ecologists' formal and informal knowledge (Roth & Bowen, 2001b). The insights that ecologists gain through their own fieldwork extend to their use of data collected by others. Although ecologists do not have the same field experiences, they share a belief in the importance of fieldwork in helping them to form appropriate research questions and in providing them with a "sense" for data. Additionally, their field insights help ecologists to anticipate and overcome many challenges inherent to the secondary use of data, particularly the need to identify sources

of data and to understand data and to assess their quality. Data comprehension based on informal knowledge gained in the field is coupled with many judgments ecologists make about data quality, and so ecologists recognize and share some of the same criteria for assessing data. In this section, I analyze the ways in which ecologists employ their field-based insights to use data collected by others. To begin, I discuss briefly the nature of ecological field research and the enculturation of ecologists.

Becoming an Ecologist

Wolff-Michael Roth and G. Michael Bowen conducted a multi-year ethnographic study of field ecologists. They published several papers from their research that are particularly relevant to my findings (Roth & Bowen, 1999; Roth & Bowen, 2001a; & Roth & Bowen, 2001b). One of Roth and Bowen's research goals was to investigate the process by which field ecologists become members of the discipline. Thus, their research focused on undergraduate honors, masters, and Ph.D. students (Roth & Bowen, 2001a). Their work illustrates the nature of ecological research and the importance of fieldwork in developing disciplinary understanding. Roth and Bowen described the uncertainties inherent in studying nature and the means by which aspiring ecologists attempt to deal with these uncertainties.

Roth and Bowen followed one of their subjects, a doctoral student named Sam, as she collected, observed, and measured lizards (Roth & Bowen, 1999). Among the challenges Sam encountered was finding lizards, a somewhat haphazard process that required searches over large areas. Over time, Sam gained what she referred to as

"anecdotal" knowledge about their daily activity level and its correspondence to air temperature.

> I usually find about five a day. I sort of am getting this feeling that they are more active later in the day. They can't tolerate, I think preferred temperature is about 20, mid 20s or maybe high 20s. Probably mid. So in the real heat of the day I don't look for the animals 'cause they're buried down too deep and then I go out again in the 4 to 6 kind of range and lately I've noticed I've had better luck (Roth & Bowen, 1999, p. 720).

Sam faced such seemingly simple problems in addition to more complex ones during her time in the field. In the laboratory, she confronted challenges as she measured physical characteristics, such as length and weight and calculated lizard speed. Eventually, the informal knowledge she gained, as she weighed lizards, for instance, enabled her to advise her assistants when they reached a problematic juncture.

> For example, while weighing a lizard, Nikki (the high-school student) found that the scale measure kept changing up and down 'because the lizard is moving'. Sam suggested to 'take the lizard [in the sock] off the scale', put it back on to the scale', and 'do the reading as soon as possible' (Roth & Bowen, 1999, p. 725).

In the field and laboratory, Sam often expressed doubts about her abilities and her measurements, but in formal academic settings, her misgivings vanished into factual statements about "the statistical significance between sprint speed (dependent variable) and body length and back-leg length (independent variables)" (Roth & Bowen, 1999, p. 720). When Sam encountered difficult-to-make measurements, such as determining lizard color with some consistency, she replicated her own work in order to improve her precision, to reduce variability, and to account for her work to others. Roth and Bowen described the role of replication in ecological research.

> Our ecologists treat 'replication' as repetition of process to check reproducibility; repetitive measurement contributes to the authority of the data and to the credibility that the research is replicable across sites. ... The

> objectivity of the work of measuring and coding is provided for by arrangements that encourage the emergence of an accountable practice (Roth & Bowen, 1999, p. 751).

Besides the importance of replication, new ecologists learn through the enculturation process how to control, deal with, and 'doctor' uncertainty and how to translate highly abstract problems into practical scientific operations through the use of 'creative solutions' (Roth & Bowen, 2001a). Additionally, by means of informal conversation, what Roth and Bowen described as fieldwork narratives, "knowledge about nature, research methodology, and fieldwork behavior is circulated among old-timers and appropriated by newcomers..." in a way that "provides newcomers in ecology with greater opportunities for learning than the lectures they have attended (Roth & Bowen, 2001b, p. 475). In spite of the use of informal knowledge to find animals, to construct variables, and to make decisions, understandings from the field do not enter into ecologists' formal writings (Roth & Bowen, 2001b, p. 471). Ecologists make, and learn to make, a clear distinction between informal understandings and "scientific" ones. In the remainder of this chapter, I show how field-based knowledge and the distinction between anecdotal and scientific evidence extends to ecologists' secondary use of data.

Field-Based Insights and the Reuse of Data

Gathering one's own data helps with reuse. Ecologists' experiences as collectors of their own data in the field or laboratory plays a significant role in their secondary use of data.[5] Ecologists' experiences as data collectors provide them 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 deal with data complexities. Informal

knowledge gained through their own fieldwork also enables ecologists to visualize potential points of data collection error, which is an important part of assessing data quality.

Currently, ecologists are recognized generally, both publicly and privately, as scientists who collect their own data in the field or laboratory. This identity is an important aspect of reuse because the insights that ecologists gain as collectors of their own data are closely related to their experiences as secondary data users. 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 patterns or variations in nature from artifacts of data. 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. Insights gained in the field or laboratory lead to familiarity with particular types of data. The importance of these insights is reflected in the range of 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 mirror 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 as well as harmony--but they share the view that familiarity with data is integral to reuse.

In order to learn more about how the forestry data she was using were collected, Ellen spent time in the field working with the crews who gathered the data. She described how the ability to visualize the collection of the data aided her secondary use of them.

> How much can we rely on this deviation? How hard is 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.

Based on their specialized knowledge, ecologists attempt to reconstruct or 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.

The remarks from Bill, Ellen, and Charles 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." This does not mean, though, that ecologists must have collected the same type of data themselves in order to reuse data successfully. As Susan said, "For each time you use different data from somebody else you don't necessarily need to have gone out and collected it yourself." A lack of familiarity with a particular type of ecological data means that extra effort is required, however, in order to understand them. For example, most of Andrea's personal research experience was related to the analysis of plant tissues, so she understood the extraction chemicals mentioned and the meanings of the numbers and units presented in the papers from which she acquired data. As she collected data from papers on soil sampling, however, Andrea noted that she was "constantly having to refer back to some standard books that describe different sample analysis protocols." This process was frustrating, and it added to the time it took Andrea to acquire data.

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 gaining insight into the importance of reconstruction from veteran scientists, who they relied on to help them reuse data until they possessed this specialized knowledge themselves. This was most evident among ecologists who referred to work they did as Masters 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. *(segment cut)* 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 related aspects of their domain knowledge that they gained through the collection of their own data and that they relied on to reuse data. 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, something that data managers and other interviewees mentioned, too. 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. Alan summarized sentiments expressed by other interviewees.

> I think there are lots of different types of ecological data because there are lots of different research questions that people come up with. That is one reason. There are lots of different reasons to go collect data. And it depends on the question that one is trying to ask, or it depends on the style of the person collecting the 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. Ecologists perceive these factors as legitimate reasons for the use of different research methods, which explains partly why they do not place an overriding emphasis on

methodological standardization. As Andrea described, 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.

Cal stated, "As with any kind of data collection, there are always a lot of factors that go into how you choose your methods." 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. What is important is that a secondary data user is able to discern and to reconstruct the method used to generate data from different data sets, so that as Susan said, "If they are different at least you know why." The multitude of reasons for which ecological data are gathered hinders the retrieval of relevant data, however, because it is often not possible to distinguish different purposes from one another when searching for data to reuse.

Research purpose dictates the methods that are used to collect data, and this, in turn limits secondary use of those data. The ability to understand the limitations of particular types of ecological data is another important aspect of ecologists' knowledge that they learn through hands on experience in the field or laboratory. All data have limitations, and these limitations are pitfalls to reuse if they are not understood. Ecologists provided examples of data limitations easily, readily, and in great detail; the literature, too, is filled with numerous examples (e.g., Bowser, 1986; Michener, 1997; Van House et al., 1998). Michael's study, for example, required data on the abundance of zooplankton species in lakes. Ecologists collect zooplankton for different reasons. For instance, collections are made in order to identify the species that exist in a particular lake. Additionally, zooplankton are gathered in order to estimate the population numbers of different species present in a lake. The latter purpose requires a systematic sampling scheme in order to project population estimates, whereas the former requires only one member of each species in order to make a taxonomic identification. In his study, Michael was interested in measures of species abundance, and so only surveys based on systematic sampling schemes could provide the data he needed. Michael noted that if he found a study whose purpose was to collect zooplankton in order to identify the species that existed in a particular lake then he was not able to use those data because they could not be used to derive population estimates. Ecological data are limited in the types of questions they can appropriately address, and ecologists recognize these restrictions.

Ecologists employ their knowledge about the relationship between purpose, methods, and data limitations to make sophisticated decisions about appropriate reuse of data. An extensive quotation from Susan illustrates the depth and extent of the

specialized knowledge that ecologists possess about data that are familiar to them, and it shows how their insights are used to understand the limitations of particular data. When I asked Susan if the documentation available for U.S. Geological Survey (USGS) water data was sufficient for reuse, she explained to me that "what they provided was enough information for certain things," but that looking at the data in other contexts would require more information.

> Well, you can go on the USGS Web site and for the Black River, where I worked, download at this particular point on the river. They have data from approximately 1945 on what the flow of the river was–the cubic feet per second of water flow. So, if I wanted to know what the past flows were on the Black River that would make that data sufficient to answer that question, and it is useful, and it is relevant, and you are comfortable with it. If I wanted to look at the relative impact of precipitation–changes in weather versus when different dams were put in, how those things worked together to govern, then I might need more information. It sort of depends on what you want to do with the data. You can do a straightforward: "Okay this is a reporting of what happened on this river"; that is reasonable. If you want to look at how the dams influence that then you might need some more information.

When I queried Susan about the type of information needed to investigate the relationship between weather and dams, she described it in detail.

> ...timing or the number of dams upstream or when the sampling occurred and how the different dams were being run when the sampling occurred. So, maybe they did that in 1945–on September 1 or whatever, but–it's reasonable. Okay, we know what the flow was then, but do we know whether all the dams were... how they were being operated at that time? Were they holding water or releasing water at that time?

Susan's quotation illustrates two important points. First of all, research questions limit the reuse of data. Secondly, research questions influence the amount of documentation that is needed for secondary data users, and multiple potential uses of the same data render it difficult to document data for all purposes.

Visualizing Points of Potential Error

The ability to visualize data collection and to understand where errors can occur is a key aspect of ecologists' abilities to recognize data limitations. Thus, knowledge about what can go wrong is an important component of secondary data use. This is another reason why standard methods alone are inadequate substitutes for indicators of data quality. Standard methods provide clues about how data were obtained, but they do not tell a secondary user if the measurements or observations were gathered skillfully. Ecologists 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 (1995) related, "there is an element of unarticulated expertise built into every attempt to solve problems according to explicit rules…" (p. 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.

Knowledge based on an individual's sense of something is difficult to transfer. I attempted to learn from ecologists more about what made them trust and distrust data in order to uncover aspects of their judgments that might be made explicit. The explanations I received illustrate the difficulty of communicating tacit knowledge in formal knowledge systems. As Cambrosio and Keating (1988) observed, scientists recognize the local dimension of knowledge categorized as "art," and in that sense they

are able to verbalize it, but transferring this sense to another is more difficult (p. 258). To assess the data they reuse, ecologists combine their "sense" of data with disciplinary knowledge and with information available about the data.

The responses of ecologists who acquired observational data suggest that "adequate" description of methods suffice to convince ecologists of data quality when the data are "easy" to collect. Fricker (2002) described this as trust based on empirical knowledge "of whether the topic is one about which people are generally trustworthy" (p. 382). Ecologists indicated that some observational data are simpler to collect than others. Two factors that simplify the collection of observational data include the existence and stability of standard methods and low variability in nature. These elements also influence data complexity. David, for example, reused data from an historical stream survey, and he also collected current stream data from other agencies and scientists. He was able to integrate the data with little difficulty.

> The fortunate part is there is a fairly standard methodology for doing stream surveys. Everyone does sort of the basic thing the same way, and so that consistency in approach allowed me to derive some pretty basic information to do the comparison.

David also knew many of the people who collected the current stream data; it is difficult to know how easily he would have trusted stream data gathered by strangers. Cal related an experience in which he and two other researchers gathered data independently, but later they learned that each had used similar data collection methods, which made it possible for them to combine their data. Cal attributed some of the similarity in their methodological choices to the fact that they were working in the tropics, which limited the methods available to them. “Ecology has developed in temperate systems, and so there might be a lot of published methods and this and that. You don’t have access to

that in the tropics." However, Cal assigned most of the similarity to the fact that their variables were easy to measure.

> I think it is mostly because we are measuring really simple variables—So, we're measuring densities of caterpillars and some parasitization and… Things that really, with common sense, there's one good way to do it.

Alan described vegetation measurements gathered in another project he was involved with that included "a small amount of data that had to be collected by some pretty simple easy to apply methods," which made it "hard to do a bad job." It is difficult to know if it would be possible for ecologists to reach agreement on what is "simple," so that the information could be made explicit.

The more objective criteria that ecologists rely on to understand and judge data are based on information they glean about data collection methods. "Adequate description" of research methods was the most frequently mentioned information necessary to comprehend data. Based on this, it is not surprising that the most common hindrance to ecologists' abilities to understand data was a lack of information necessary to reconstruct data collection. When ecologists referred to metadata as a way to judge quality, they mentioned information such as sample size, the unit size in relation to the number of samples, or the number of replications of an experiment. Ecologists used this objective information in combination with knowledge gained from academic training and field experience, to assess the quality of data. Bill described some of the approaches he used to judge data.

> In terms of these studies, things that would convince me that it was better data would be if they had a larger sample size or visited all the study sites multiple times. It would be an assessment of sort of the effort involved. If they were one-time sampling at a series of study sites, I would be less confident that it represents the true pattern than if they had gone out on multiple visits and what I received was evidence of the multiple visits,

> then I could derive from those multiple visits an average, or average them across those sets of visits for each site. That would be more convincing and a more stable pattern.

This example also illustrates how formal notions about standards of scientific practice, such as replicate sampling, can guide ecologists' assessments. Documentation of research methods was important in order for ecologists to comprehend how data were collected and it was used frequently to help determine quality. However, an adequate description of methods did not guarantee quality because of the tacit knowledge ecologists sometimes applied to judge data.

In their interviews with me, ecologists emphasized the role that their field and laboratory experiences played in their secondary use of data. Other aspects of domain knowledge, such as that gained through formal education and familiarity with the professional literature, underlie ecologists' experiences and influence their choices as well. Porter (1995) noted that shared knowledge can help to alleviate distrust and overcome distance, but "the problem of trust can never be eliminated" (p. 214). For ecologists, data collected by others are judged on explicit and tacit dimensions of domain knowledge, which in many cases, is combined with individual knowledge to make decisions about which data to reuse. In this section, I focused on an examination of ecologists' domain knowledge because it forms the base for their reuse of data, and it provides the rationale for many of the choices they make throughout the reuse process, including decisions that incorporate personal insights. In the following section, I discuss the individual knowledge ecologists rely on to reuse data, and I show how it influences their decisions. Ecologists are alike in drawing on available individual knowledge, especially that based on first or second-hand acquaintance or on perceptions of the skills

or values of other scientists. However, it is also access to individual knowledge, along with personal tolerance for uncertainty, which explains the different approaches that ecologists take toward reuse. In the next section, I introduce briefly the individual dimensions that affect secondary use decisions. Following that, I demonstrate how ecologists interweave all aspects of their knowledge to reuse data.

Individual Knowledge

Secondary use of data is preceded by scientists' willingness to use data they did not collect themselves. 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. The ecologists that I interviewed had varying levels of tolerance for uncertainty, although one characteristic that distinguished them all was the willingness, as Michael observed, to step beyond their own data.

> There are definitely different comfort levels for people. Some people will forever be confined to studying their own system because they are unable to accept any degree of, you know, sort of taking other people's word--sort of dealing with data that they didn't actually see collected themselves.

Each ecologist I interviewed found ways to reach a personal level of comfort in using data collected by others. They were willing to accept some uncertainties, but not others. The means that ecologists used to arrive at their level of assurance revolved around understanding the data they gathered for reuse, which was based on domain knowledge, and on judgments of data quality that were based on domain and individual knowledge.

Individual knowledge consists of unique and personal insights and connections that lead to trust and distrust of data that affect ecologists' decisions about what data to reuse. Individual knowledge is not limited to direct experience with another data

collector; it can also be formed by the opinions of trusted others and by cues absorbed through the local milieu. Knowledge and perceptions of the skills and values of others and of "the way things work" enter into the reuse process from the start and affect choices ecologists make along the way. The acquisition of individual knowledge is unique and variable and formed by multiple experiences that follow different patterns. In addition, *individual* implies information and relationships that are private and not openly discussed, which may explain partly why such insights were rarely the first criterion that ecologists mentioned in discussing their assessments of data quality. Another reason for their reticence on this topic is that it conflicts with their emphasis on following standards of scientific practice.

Ecologists recognize the importance of knowledge they acquire in the field, but they rely mainly on shared notions about norms of scientific pursuit to guide their search for data and to frame their experiences because informal knowledge is not acknowledged publicly in the context of "real science" (Roth & Bowen, 2001, p. 477). Ecologists underscore objectivity, but my results show that ecologists' choices are influenced by domain knowledge, especially insights acquired in the field, by individual knowledge, by personal tolerance for uncertainty, and by the complexity of data they reuse. I analyze these factors in the next section, and I show how ecologists harness all the knowledge available to them to locate and use data they did not collect themselves.

Employing Knowledge to Understand and Assess Data

Ecologists often interweave their domain and individual knowledge, along with personal tolerances for uncertainty to make choices about where to look for data and to

embrace and exclude data for secondary use. The exact combination of these factors differs for each individual, and thus, not all ecologists make the same decisions or follow the same path. As I discussed previously, ecologists noted the importance of finding means to root their ideas in reality, and "book-learning" alone was insufficient to provide this insight. Cambrosio and Keating (1988) summarized the distinctions made between various types of knowledge--public, local, and tacit--by several sociologists of science, but they rejected these distinctions as inadequate to describe scientific work. They noted that tacit knowledge is described as non-verbal, inaccessible, and non-transmissible and "largely beyond the control and manipulation of scientists" (p. 246). They found, however, that scientists recognized the tacit and local dimensions of their knowledge. My interviewees, too, went beyond "ideal, algorithmic accounts of their work" in their conversations with me in order to describe a range of knowledge that they used to locate, comprehend, and judge data they did not collect themselves (Cambrosio & Keating, 1988, p. 258).

Ecologists' knowledge exists for them to use at any time, and thus, acquiring data, understanding them, and assessing their quality can occur simultaneously and are often part of an iterative process. Their intellectual adeptness obscures the rationale for ecologists' choices, masks some of the considerations that pervade the sharing of data among members of the same community, and makes it difficult to draw distinct lines between each stage of the reuse process. In this section, I show how ecologists use their knowledge to help them establish criteria for data and to understand data and judge their quality, two closely linked processes. In the section that follows, I analyze how ecologists employ their knowledge to make choices about where to look for data and to

develop strategies for acquiring them. Since their shared membership in a community of practice provides a rationale for many of ecologists' choices, I begin this section by discussing domain knowledge because it forms the most important component of secondary data use by ecologists.

Asking the Right Questions

In any research project, the acquisition of data is guided by a doable research problem and involves a number of decisions (Clarke & Fujimura, 1992). Markus (2001) noted that, "one characteristic separating experts from novices is that experts know what questions to ask" (p. 61). Data appeared simultaneously to drive questions and to be driven by them. In cases where ecologists were provided with existing data sets, their expertise aided them in forming research questions that were appropriate for the data. When data were not at hand, ecologists formed questions based on data that were familiar to them. In this instance, they often did not know at the outset what they would be able to obtain and make comparable, so sometimes their questions were altered based on the data they could ultimately find, access, and integrate. Regardless of the sequence, ecologists' expertise helped them to pose research questions that relied on data with which they were familiar, and this acquaintance assisted their comprehension, assessments, and acquisition of data in several ways: it provided ecologists with a sense that data were available; it helped them to understand and assess data; and it assisted them to develop methods to obtain data. These steps are part of the process of reconstruction, in which ecologists mentally reassemble the original collection of the data they seek to use in order to find, understand, and judge data.

First of all, their familiarity with certain areas of ecological research provided ecologists with a sense that the data they needed existed, a factor that helped to make their projects doable. At the outset, most ecologists thought they would be able to acquire data.

> Within my area, I knew that I could do this before I started it. So, it wasn't a question. The only question was how many lakes could I get within a reasonable amount of time. There was no question about: Could I get the information and would the analysis be feasible? There was no problem with that. I know that people have different ways of making species lists and measuring primary productivity but I didn't think it was... I just chose to ignore those differences, which I think is reasonable.

The above quotation from Stephen also shows how ecologists' knowledge helps them to anticipate factors that surround the reuse of particular data, such as the need to integrate data that were collected for a variety of purposes, using heterogeneous methods, and at different temporal or spatial scales.[6] Even less experienced ecologists, like Katherine, were confident at the outset about the existence of data. Katherine and her co-author conducted a meta-analysis that relied on the published literature as a source of data. When I asked her if they conducted an exploratory search of the literature to make certain that data needed to address their research question were available, Katherine explained to me why this step was unnecessary.

> We were all pretty up on the literature in terms of what the important questions were. … We were all very well aware of what the major potential issues would be in asking the question we did.

Familiarity with the literature, acquaintance with general research trends, and specific knowledge about who is working in what areas provide ecologists with insight into the types of data that are available for reuse. This knowledge also helps ecologists to judge initially whether they will be able to obtain data. For instance, ecologists who relied on

the published literature to acquire data were aware already of publications that contained data. As Michael said, when he began data collection, "I sort of knew about a bunch of papers like that, that sampled some numbers of lakes." In Alan's case, he knew that very little had been published about the bird species he studied, and so he was aware early on that his strategy for obtaining data would need to include sources beyond the published literature. Alan had some previous experience with museums and "knew that if they had specimens at all, they likely would have body weight measurements."

Dimensions of Data Acceptance

Secondly, since research questions were linked to areas of ecology with which they were familiar, ecologists carried with them some of the information they needed to understand, judge, and integrate the data they acquired. As Stephen explained it in talking about the data he gathered for reuse, "I chose it to be the kind of information that is readily available and that I am familiar with." The mix of domain and individual knowledge that ecologists employ to understand and assess data combined with personal research standards and tolerances for uncertainty make the process of data acceptance a complex one.

Edwards (2000) stated that, "data contamination occurs when a process or phenomenon other than the one of interest affects a variable or value" (p. 70). In the literature, data quality most often refers to aspects of data management, such as detection of data entry errors and consistency of coding. Nancy Van House (2002) observed that trust, particularly with regard to observational data, includes shared orientation and values, or what she described as "virtue." In judging data, Van House asserted that

secondary data users ask if the data collectors are "competent and virtuous?" For the ecologists in my study, competence was judged in relation to perceived skill and expertise in data collection that was based on domain and individual knowledge, sometimes used in combination, and virtue was assessed in terms of personal perceptions of the values of others. Below, I focus on the aspects of domain knowledge that contribute to ecologists' comprehension of data and to quality judgments.

Data Acceptance and Domain Knowledge

Ecologists employ their knowledge to make decisions about data to include or exclude from their study, which in turn, reduces their concerns about quality or relieves them entirely of the need to make those assessments. At times, ecologists' choices drew on a combination of their formal expertise and informal knowledge about particular types of data, such as the difficulty of collecting them, the variability of specific parameters in nature, and the recognition of valid measurement ranges. Michael, for example, 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 much 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.

In this case, 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. As I noted earlier, Fricker (2002) described this type of trust as based on knowledge of whether the topic is one about which people are generally trustworthy. In Susan's case, her 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 that 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 isn't 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.

The complexity of data, such as their variability in nature, influences their reuse.

Ecologists' expertise also helps them to identify valid and invalid data values. Ecologists revealed this aspect of their knowledge in their answers to questions about how they assessed data quality as they sorted through data to reuse. For example, as a doctoral student, Andrea analyzed wetland plants and soils for nitrogen and phosphorus, and she relied on this experience as she gathered data for reuse.

> I was familiar with the kinds of ranges that you should expect from, say, wetland plants or wetland soils. So, I did reject values that seemed extraordinarily high or extraordinarily low.

Ellen related an experience in which she gathered biomass equations from the literature. She said that she accepted all the data she found, but then she shared examples of situations that led her to reject data.

> There were some cases with the biomass paper, for example, where we are getting ten times forage biomass, or ten times woody biomass than forage. Okay, that is wrong. We know that can't be physically possible. But barring that, it was my inclination to say… if someone said this is woody NPP, that is what it is.[7]

The above quotations from Andrea and Ellen represent the responses I received from other ecologists who mentioned that they included and excluded data based on assessments of their perceived accuracy. Later, Ellen said that in all her data reuse projects, "If I honestly could not figure out what they had done, then I would not use that data point." Other ecologists reiterated this sentiment in ways that demonstrate the connection between data comprehension and quality judgments.

Ecologists performing meta-analyses are somewhat unique in having methodological guidance for the collection of data for secondary use.[8] Katherine and her co-author were stringent in following these guidelines, which relieved them of the need to judge data quality.

> Well, we tried not to judge quality. We weighted the studies based on their sample sizes and on the measure of variability. Variability can be due to real variability or can be due to experimental variability… *(segment cut)* Of course, variability goes down as your sample size increases and so on and so forth… *(segment cut)* Every single record that we calculated an effect size for was weighted both by its measure of variability and by its sample size. Because we... You can't... There is no way for us to judge the quality of those numbers, and if we tried to do that, we would be biasing the meta-analysis.

Cal, who discussed a meta-analysis that he conducted recently, noted that he accepted any data that made it through peer review. Andrea also used peer-reviewed publications to limit the sources from which she acquired data.

Peer Review and Data Quality

Peer review is intended to be a mechanism to objectively assess scientific competence. Ideally, therefore, peer review should reduce the effort required to make quality assessments. However, with the exception of the three less experienced ecologists I mention above, those who gathered data from the published literature did not rely on peer review to certify data quality. Peer review also emphasizes reproducibility of results, which is different from the secondary use of data to generate new knowledge.

In a recent paper, Chinn and Brewer (2001) offered a theory of how people evaluate data. They defined *data* to mean those that appeared in a published paper: "By *data* we refer to the details of the scientific study used to make observations to test the theory" (p. 331). Chinn and Brewer's definition reflects the fact that when a scientific paper appears in the peer-reviewed literature, we say that the data have been published. Yet, one rarely sees the data on which a paper is based; they are taken largely for granted by the scientific community. There are important differences, however, between the data that **underlie** a paper and the description of data as published. There are times when this distinction becomes critically important, such as in controversies over scientific results (Collins & Pinch, 1998; Service, 2002). In fact, Bill, a long-time advocate for data sharing, stated that his support stemmed from controversies he had been involved in that motivated his interest in seeing that scientists made their data available. Additionally, a

lack of terminology to clearly describe data in different forms creates confusion about the ability of peer review to spot poor quality data since, as Bill noted, "Most journal articles, the referees don't see the data; all they see is the manuscript." Nearly all the ecologists I interviewed cited examples of suspicious data reported in peer-reviewed publications. They recognized that one rarely sees the data on which a paper is based, especially in journals such as Science and Nature, which publish short papers that Bill described as "cartoons of papers." Bill also described examples from papers he read where figures and explanations about them in the text did not agree. Michael said that when he reviews papers he often counts up degrees of freedom in regression tables to see if they add up to the right number, and he often finds that they do not. Bill summarized some of the inadequacies of using peer review to judge data.

> I see a paper that comes out, and I think there is something strange about it--the conclusions they reach. Generally, I just go back and look at the methods. What the hell did they do? And if that's only cryptically described, and I am looking at the analytical techniques--what did they do? And that is cryptically described. So, sometimes I am figuring out what they did from reading captions on the figures.

Ecologists did not recommend abandoning peer review, but they recognized its limitations, particularly in terms of assuring data quality. Bill's quotation highlights, once again, the important link between ecologists' comprehension of data and their concerns about data quality. When the information provided in a publication is insufficient for an ecologist to understand how the data were collected, then it is difficult to make assessments about their quality. In such cases, quality issues are secondary since ecologists will not reuse data they do not understand. For all ecologists, understanding data was integral to their reuse, and they expressed frustration when the information needed to comprehend data were lacking in peer-reviewed publications. Less-

experienced ecologists, in particular, clung to the ideal of replication in scientific reporting. Ellen's comment typified what these ecologists viewed as an important part of what they learned in graduate school: "I was always taught that somebody should be able to replicate your methods using what you write in the methods section." When I asked Michael what his greatest obstacle was to finding the data that he needed, he answered my question in a way that connected repeatability, understanding, and data.

> There were a fair number of studies that may have had the right kind of data but didn't report what I needed to know. So, I mean it is surprising, you think to get a paper published you would have to... the criterion is supposed to be that someone else could go do the exact same thing. But for instance, there were a lot that didn't report how many samples they took from a lake. Some lakes they would sample three, four, five times, and other ones they would sample once, and you couldn't tell which were which.

Although they are less than perfect, overall, publications are a good source of information about data since ecologists were often able to reuse data from the literature without personal contact with the original data collectors.

Creative Integration, Data Acceptance, and Domain Knowledge

Secondary use of data is a process that involves many choices, adjustments, and accommodations. Once ecologists have data in hand and are ready to use them in a new way, they are faced with other choices. Ecologists use their domain knowledge to resolve the challenges that arise from the need to integrate data gathered at different times for multiple purposes. The ability or inability to accomplish this is another side of data acceptance. One of the most difficult aspects of combining data from multiple sources is making them comparable. Data complexity adds to this task.

Reductions in complexity help to make data more comparable. In order to make data comparable, inherent data complexity can be reduced by the way they are reused. For instance, Charles used data on a particular amphibian species that had been gathered from both systematic field surveys and museum records. Ecological surveys are organized efforts to locate species, whereas museum records can only be used as positive sightings. Charles described museum data as, "Someone found something. You have no negative records. And you have no idea of how much effort went into getting that one sighting." By using survey data differently, however, it was possible for Charles to make survey and museum data comparable.

> But, in order to compile different stuff together--different sets of data--one way that you can make them compatible is by kind of lowering the standard down to presence/absence. If someone sees something, you don't need to worry about whether they saw two of them or twenty of them and was that because of different methods. Just say, "Okay, something is there." So, that's all we have is a presence. That's one way of getting around that problem of lack of compatibility of methods.

Alan performed a similar transformation with the bird data he acquired. Some of the data he gathered were based on organized surveys and some were sightings made by bird watchers, but Alan used all the data to represent the presence of a bird at a particular point in time and space. Low complexity data, such as some museum records, are not without pitfalls to reuse. Charles observed that these snares might lead to intentional or unintentional misuse.

> In some ways it is just **very** simple, you know. Someone saw an animal on such and such a date at such and such a location. That's basically it. And you can explain that to six-year-old. The only tricky thing… and in some ways it is not that hard conceptually, but I see people making the mistake all the time... What does the absence of a record mean? And the absence of a record doesn't mean the absence of a species. It may just mean a lack of survey effort. And you see biological reports all the time that people consult the state biodiversity database and say, "Oh, we have

> no endangered species on this piece of property. It's okay; go ahead and turn it into a shopping mall."[9]

As Charles's experience shows, even when data complexity is reduced, it is important to understand the complexity of the data set from which the data were acquired in order use them appropriately. Additionally, it requires domain expertise, like Charles's, about the habitat requirements of certain species to make an informed guess about whether the species is likely to populate a locale for which no record exists.

In the process of being made comparable, data are sometimes transformed from their original form into something new. This transformation makes it difficult to follow data through the stages of reuse--from acquisition to final reporting--and even the original data collectors might not recognize their data when they "appear" in their new form. This transformation is part of the process of regeneration, in which data become part of a new study that involves its own data collection. Regeneration is characterized by accommodations and adjustments to the data available for reuse. There are several ways that the ecologists I interviewed transformed data in order to make them comparable. These methods included using one or more data values to calculate another, assigning different meaning to data, collapsing together categories assigned to observational data, using data as a proxy or to create derived indices, and extracting data from a graph or table when they were reported at enough resolution. Below, I provide several examples of data transformations from my interviews with ecologists.

In some cases data are transformed in the sense that they are used to create new data. In this way, existing data generate new data. For example, Bill examined the relationship between latitude and animal population density. Some of the papers that Bill acquired data from reported the latitude and longitude of the study site, but others did not.

For these latter cases, Bill used an atlas to estimate the spatial coordinates based on information presented in the papers. Bill also found studies that reported a range of latitudes, and from these he calculated mean latitude. Alan, besides collecting bird sightings from museums, bird banders, and individual birders, gathered information on a bird's sex and weight, when they were available. Later, using the bird weights that he had collected from others, Alan was able to calculate a bird's body fat. This measure was then used to further validate his model's results. So, a bird's weight was shared with Alan, but he used this measure to report body fat. In another example, Andrea talked about how she was able to convert information provided in a paper to obtain the data she needed.

> They would report data in a format that wasn't what I was looking for, but they gave me information that I was able to… For example, say they gave a table with the carbon to nitrogen ratio and then from another table I was able to figure out what the carbon concentration was. I would then be able to use those two pieces of data to then figure out what the nitrogen concentration was.

Andrea and several other ecologists also described the use of a pencil and a ruler to extract data from a table or graph that appeared in a published paper.

> In a small number of cases, I actually… Data would be reported with enough resolution on a figure as a graph or something, and I would actually extract data points from the graph. So, I would sit there with my pencil and ruler and draw lines and try to figure out what the numbers were.

One of the data managers also remarked on having seen ecologists use this method.

Data transformations frequently occur because more than one technique or method exists for making a particular measurement or because methods improve or change over time. In other instances, more than one method exists, and the ecologist makes a choice of which method to use based on the purpose of the study or on personal

preference. For example, pelagic primary productivity can be measured by the Carbon-14 method. According to Stephen, the Carbon-14 method is considered a more direct measure of productivity than proxy methods, such as nutrient loading, biomass, or soil fertility. Since Stephen preferred to measure primary productivity using the Carbon-14 method, he looked for the use of this method as he gathered data.

> One of the things that people were doing during that program, the IBP program, was measuring annual primary productivity using Carbon-14 techniques. And that was pretty much the first time it was used, and it was used very widely. And since then it has been used now and then. However, people would have been switching over to different ways of measuring primary productivity, like on an hourly scale and looking for a maximum rate is one way, or using chlorophyll as a surrogate for measuring the weight of photosynthesis. So, they just measure the amount of chlorophyll, and I could have used just like maximum chlorophyll concentration instead of the annual rate of carbon fixation. But since I was in graduate school in the 60's I think that the Carbon-14 technique is really neat. So, I just went with it. But I could have easily probably have gone with chlorophyll, but it had to be one or the other.

Stephen's training influenced his choice of method for measuring primary productivity. If the Carbon-14 method was not used, Stephen could convert other measures to make them comparable with the Carbon-14 technique, but for some data this was not possible. The differences in methods used to obtain measurements of primary productivity limited the number of lakes that Stephen was able to include in his study.

> There are a lot of lakes not on here. The reason for that is they didn't measure primary productivity using the Carbon-14 method or any other method. So, if I got a paper that had a productivity measurement using oxygen, then I could make a conversion. But if they told me how much chlorophyll was present in the lake, I didn't attempt to make the conversion. … I thought it was going to too much of an extrapolation.

The above example from Stephen's study illustrates that it is not always possible to find ways to make data comparable. Andrea, who gathered data from studies that reported

information on tissue nutrient data for wetland plants, described several limitations that surfaced as she collected data.

> The other thing, too, is that when I started collecting some of these data, I started to realize that you can't just go through old publications, put together a bunch of data and run some stats on it because in most cases those data have been collected in different ways using different methods. In our case--for the tissue nutrient data for wetland plants--there are a number of ways that you can collect plant tissues. You can collect them at different times of the year. You can collect different parts of the plant. And then there are many ways you can process samples and many ways you can analyze for nitrogen and phosphorus. If two samples have been collected in different ways and analyzed in different ways, then often those numbers aren't comparable.

Andrea's description also shows the iterative process of data gathering and the ways in which the process of understanding data is linked with their collection. Colleagues alerted Andrea to the fact that the older method for extracting plant tissue nitrogen underestimated total nitrogen, but they were not able to tell her how to make old and new measures comparable. Thus, she was unable to use data measured using the older method. Ecologists do not know always at the outset what they will find and be able to reuse successfully, so they often gather data that do not get used.

Individual Dimensions of Data Acceptance

In talking about their secondary use experiences, ecologists emphasized their domain knowledge, especially their informal knowledge shaped by fieldwork and their attention to standards of scientific practice. The place their individual knowledge played in data reuse emerged secondarily, and at times, reluctantly and subtly. This reticence is attributable to a clash between individual knowledge, which is unacceptable to share on a wide scale unless it is tied to a scientific rationale, and objective norms of science.

Except for Stephen, ecologists acknowledged less reliance on personal knowledge of other data collectors to find data and assess their quality than the literature suggests. When it is available, individual knowledge lessens the effort ecologists expend on judging data. Additionally, personal connections can provide access to information necessary to understand data. Individual perceptions and knowledge also influenced the length to which ecologists would go to pursue data comprehension. Since judgments about an unknown data collector's skill are often difficult to assess based on available information, it is not surprising that personal insights into the skills or values of other scientists, whether positive or negative, enter into data reuse decisions. Individual knowledge plays a secondary role in data comprehension, however, and therefore, it is a subordinate driver in data reuse decisions.

Stephen was unique among the ecologists I interviewed in relying almost exclusively on personal networking to identify and obtain data. By taking this approach, he eliminated many quality concerns by confining his collection of data to research programs that were familiar to him. Stephen's individual knowledge went hand-in-hand with his concerns about data quality. Interestingly, though, his public presentation of data collection methods highlighted a more objective tack.

> So, well studied lakes… something I didn't address in the paper. I said something about… it had to have been sampled like more than once over more than one year or something like that. But also I think another assumption is that it was being studied by some recognized professional aquatic ecologist. So, that was a hidden assumption. *(segment cut)* I tended to take lakes that were the focus of long-term research programs run by one or two people, who have a good reputation.

Later in the interview, Stephen noted that at the outset he excluded data from large, federal, limnological databases because "I know something about the quality of some of

the older lists and some of the newer lists, and I know that the newer lists are lower quality because I have been involved in checking some of those determinations." Stephen made skill judgments before he began to collect data, and in this way, he sought to avoid the need to make quality judgments.

Other ecologists were less apt to note the role of personal insights. After some discussion about how he judged 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. Some ecologists only hinted that individual insights about others entered into their experiences. For example, I asked Nancy if she knew the authors of the published papers that she extracted data from, and she replied, "I probably know most of them, actually." Yet, she did not connect this directly with her quality judgments, although she later told me, "I am nervous of other people's data." Andrea, the third author of a paper published in Ecology, gathered the data that were reused. As I mentioned earlier, Andrea limited her collection of data to the peer-reviewed literature, and this served as her first round of quality control. In addition, she looked for papers that had clearly outlined methods for both sample collection and tissue or soil analysis, and so information available to understand data was her second criteria for quality. Personal aspects did come into play, however.

> Those were the main two... the main two sort of criteria that I used. You know, it is kind of funny--working with somebody like Elizabeth, she has

> a very... She has been working in this field for awhile, and she has a really good sense of who does good work.

Andrea acknowledged that her advisor, the paper's first author, personally knew the authors or knew of their work for "eighty-five percent of the papers." Knowing people and recognizing who does good work does not lead to automatic acceptance of data, but it lessens concerns about data quality.

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's reasons for choosing not to use the data from the large database demonstrate the complex arrangements of knowledge that make up many assessments of data. He employed personal insights along with knowledge about the difficulty of identifying this species. Alan explained to me that he kept all the data he collected, but then recalled that he had eliminated data from the source he refers to in the quotation above.

> I think they do probably as good a job as one can for this nation-wide type effort. I think if that data were available on the Internet, for example, where you have with it the documentation from the Bird Center showing what they do in the way of training and providing materials and so forth to people that do surveys to insure quality, that would give you some assurance that it was a pretty good database. And it is pretty good. I just didn't think it was quite as good as the other data that I had, so I just went with the best basically. I think I would have still made that same decision had all the data I used been available on the Internet with good documentation. *(segment cut)* I use my own judgment, too. I mean it is

> kind of my feeling about knowing how these--how people operate in the field. It's hard to put that kind of stuff as documentation on a database.

Alan's experience also illustrates how the acceptance of data can be affected by the amount of data available from one source. As Charles indicated, unless there is a systematic bias, more data is better since a larger sample size reduces error.

> My general strategy is... hopefully doing analyses that are dependent on lots and lots of data. Any one data point has very little influence on kind of the overall results, and so if there's a sprinkling bad data, hopefully it doesn't make a difference.

If data quality is questionable, however, a lot of data from one source can bias results. For Ellen, who used data from one large database, first-hand experience with the data collectors in combination with the thousands of available data points cinched her confidence in the data.

> I guess there are two things that help me to trust the quality. One is knowing the people who collect the data. Getting to know them and talking with them about the data. I am supervising these field crews--you would not believe the amount of pride they took in collecting this data and doing it accurately. They are so concerned; it amazed me. *(segment cut)* The second thing that really increases my confidence in the quality of the data is just the tremendous number of data points.

Together, these examples help to illustrate the combined domain and individual knowledge that shape ecologists' decisions, and they show the unshakable influence of positive or negative personal knowledge of other data collectors on data acceptance.

Individual connections that provide access to others' knowledge of data is also used to obtain information necessary to understand data. This type of "metadata" presents a different set of challenges from more impersonal sources. For example, in spite of her experience working in the field with the data collectors, Ellen's understanding of the forestry data was complicated by the fact that the data had been gathered for a

purpose much different from the one for which she was reusing them. Even though Ellen's boss was committed to the reuse of the data, and Ellen was co-located with the group responsible for the data, she stepped carefully to gain access to information about the data.

> Probably the biggest challenge with that was trying to navigate really... it is all about people and personalities, but I think... I mean, actually, that is very, very big for me in this particular project, so I think the biggest issue really is for me--or has been--getting the information that I need about how the data were collected, or put together, or analyzed--without, at the same time, burning bridges, if you will, with the people who did all of that work.

For other ecologists, the process was not as delicate, but it required attention to cultural norms of acknowledgment and reciprocity. When I asked Alan about his experiences in sharing his data with other scientists, he mentioned that one of the ecologists he acquired data from for his bird study later asked him for data.

> One of the people who gave me data was a woman in Mexico. She gave me quite a big chunk of data, actually. And then see... then it turns out that she was one of the ones that asked me later, just a couple of years ago, a year ago, for some of my data. Of course, I felt I had to send all the data I had because she had helped me out.

The Mexican scientist learned about weather data that Alan compiled from her interaction with him when he borrowed data from her. Alan's example illustrates the way a scientist's personal experiences are called upon when they seek data to reuse. As he said, "She helped me first, and then later she needed some help and I was able to provide it." Additionally, Alan and other ecologists were careful to acknowledge data sources in their published papers, and they noted that they sent reprints of their papers to those who supplied them with data. This behavior did not extend to asking data contributors to review drafts, however.

Knowledge Combinations

Finally, as I indicated above, assessments of data quality sometimes draw on a combination of ecologists' domain and individual knowledge. It is here that the line between informal knowledge and personal insight sometimes blurs and where the virtue that Van House (2002) described comes into play. One way that ecologists I interviewed spoke about virtue was in terms of what Stephen described as "commitment to the organisms." Stephen 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. Probably mostly graduate students, but 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 and others defined commitment in terms of consistency and qualifications of personnel, including years of experience and dedication to the work at hand. They made assessments about others' commitment based on a mix of domain and individual knowledge. One of Alan's misgivings about the large database of bird observations was based on the fact that the data collectors change frequently.

> The person who does them changes from year to year and typically a new person starting is a green person who doesn't know the birds real well yet. A person does these things for a couple of years and gets good at it, but then you get burned out, and they don't do it anymore. So, you find a new person to go do them. That person has got to come up the learning curve.

Several ecologists made reference to the contemporary practice of using seasonal and temporary employees to collect data. The skills and experience, and less frequently, the

dedication, of these personnel were mentioned as sources of doubt about data. David connected his positive judgments about the historical stream survey data he reused with the consistency of the scientists who collected the data. He had their field notes and published reports to draw from in determining their longevity with the project.

> Most of these people that worked on the survey were pretty seasoned biologists. A lot of them had advanced degrees. They were not sort of people 'off the street.' … Many of them stayed with the survey through the bulk of it. So, it wasn't sort of like today, where we hire seasonals in the summer, and they work for us for three months and then we never see them again. These were people that were professional biologists and that was their job. So, they thought through from the data collection to the analysis and the report writing.

The documentation related to the historical stream survey provided information that allowed David to mentally reassemble much of the scientists' work. The opportunity to confirm his reconstruction with a member of the survey team, who he described as "an eminent limnologist," strengthened his assurance in his understanding of the data and their quality.

> It really fit what we had pieced together so... given his stature and his credibility as a scientist and his clearly vivid recollection you were able to sort of walk away and say, "Okay, this is how they did it, and our collecting additional information in the present is going to work." That was... I think... We were very confident in what we were doing at that point but having sort of independent confirmation from someone who worked on the project was really powerful.

The above quotation illustrates the length that ecologists will go 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.

Employing Knowledge to Find and Obtain Data

In collecting data for reuse, ecologists are attentive to standards of scientific practice and to future public scrutiny. This is where the formal side of community membership-related notions of scientific standards comes into play. Ecologists' research questions helped them to identify specific criteria for the collection of data, which they used to decide what data to look for and to devise methods to obtain the data they needed. Ecologists apply the same principles they follow to gather their own data in the field or laboratory, and this helps to explain the approaches they choose. Ecologists select methods to gather data for reuse that work in concert to help them bound their collection of data, that increase their chances of obtaining data, and that reduce the risk of errors associated with the secondary use of data.

Ecologists' specific criteria and requests for data served a couple of purposes. For one, they helped ecologists obtain the data they needed in a form they could understand and reuse more easily. Second, specific requests increased the potential that the data would be shared because the request stated the purpose for which the data would be used. Some ecologists, such as Alan, attributed their success in obtaining data to their explicit requests.

> In our case, it was almost a perfect match because we had used a model to predict what birds should be doing. Then we knew precisely the kind of data we needed. So, it was real easy to go out and say, "We need this type of data--precisely. If you have it, we would love to have it. If you don't have that, we don't need it." Then it is real easy. If you get data, it is going to be good data, and if you don't you don't.

Pre-defined specifications for data were especially important to ecologists who conducted meta-analyses. Guidelines for meta-analysis are attentive to issues of bias in the collection of data, and they dictate that collection criteria be established before data are

gathered in order to reduce potential bias. The ecologists in my study interpreted these recommendations at various levels of stringency, but all were heedful of them. Like Alan, Katherine attributed her and her co-author's success in obtaining data to the clarity and specificity of their request.

> We were very, very specific in our requests. We would send them an email saying, "We read your paper; we are doing this meta-analysis; we were hoping you could provide us with…" And we would specifically list: "One this, two that, three this. From this page." And so they had a very specific reference and a very specific request.

Data criteria also contributed to bounded and unbiased methods for collecting data, which is another important aspect ecologists consider in gathering data to reuse.

"Bounded" Sources of Data

Ecologists' collection of data for reuse mirrors the standards that guide the gathering of their own data in the field or laboratory. Thus, ecologists need some assurance that their sampling scheme is scientific, which requires them to find means to identify and draw data for reuse from some representative "population." In other words, ecologists look for strategies that place bounds around their collection of data and that provide believable rationales for their choices. In some cases, the use of an existing data set, which provides its own bounds, satisfies this requirement. Since ecology has few comprehensive databases available for secondary data use, ecologists often have to find other means to place boundaries around their acquisition of data. The ecologists I interviewed accomplished this objective in one or more ways.

First of all, ecologists utilized the published literature to provide a frame around a segment of the world of data. Bibliographic databases are a recognizable tool to access

an enclosed portion of this literature, particularly journal papers. They are not, however, the most effective tools for identifying relevant publications. For one, ecologists' criteria for data are very specific, and these requirements are not captured in abstracts or in other information indexed by bibliographic databases. Bill's experience was common among those ecologists who searched bibliographic databases to locate data in publications.

> Abstracts don't necessarily tell you what data are available. They might reach some sort of derived conclusion from data, and to figure out if it was the kind of data you needed you had to go look at the paper.

The information presented in bibliographic databases is often insufficient to determine a particular study's purpose. This factor makes it difficult to determine if a publication will contain the necessary data. For example, earlier I related that Michael's study required zooplankton data that could be used to address questions of population abundance, and data from taxonomic studies were unable to meet this condition.

> At the beginning, I used a bunch of computer searches. I'd do like Web of Science or something and looked for "zooplankton survey" or whatever. Then… but then from there it was sort of hard to tell from that sort of thing whether they are going to be useful. You generally had to go get the paper itself.

The multitude of reasons for which ecological data are gathered hinders the retrieval of relevant data because it is often not possible to distinguish different purposes from one another until an ecologist looks closely at the published paper. As Katherine described, "There are a certain number of experiments that will not fit your criteria because they have a different goal." Differences in research purpose create complications throughout the reuse process.

Another common frustration with bibliographic databases is that they do not adequately cover the literature because they provide access to limited years of

information. So, ecologists had to find ways to get beyond what Stephen referred to as "the digital curtain." Thus, they sometimes searched selected years of particular journal titles manually, alone or in combination with database searches, to frame a population of data. Literature cited in relevant papers added to ecologists' confidence that their methods retrieved data from the existing pool. Ellen described a study she conducted recently that relied on the published literature as a source of secondary data. Her comments represent a popular approach taken by ecologists who utilized this method to gather data.

> We started with a literature search--basically using Agricola. Then every time I got a paper, I would comb the references in that paper for other ones that related to biomass. So, it's this iterative process and eventually all the papers started citing one another. Then you know you are done.

Some ecologists added literature or unpublished data that they obtained from colleagues or from their personal files, but the published literature provided the main, publicly-presented frame for data acquisition. Additionally, the literature clued ecologists into the existence of unpublished data held by scientists they did not know personally; some ecologists pursued these leads to obtain more data.

Individual tolerance for uncertainty, practical issues of time versus effort, and the nature of a particular area of ecology also influenced ecologists' choices about how extensively to review the literature. In regard to looking for additional data, Katherine stated, "We could have searched beyond that, and I am sure we would have found a handful more articles to review, but the number of usable articles per unit search time would have been really, really small." Over the years, scientific questions change and technologies improve, and knowledge of these shifts can direct ecologists' efforts to

locate data to reuse. Michael noted that ecological surveys were more common in the past, and so older literature was a valuable source of data.

> It was sort of more of a thing back thirty years ago or something you'd go out and sample a bunch of lakes and report what you found, which now that is not quite the way--really the way ecology is done.

Katherine followed up her earlier comment by adding, "but I have to say, especially in a field like ecology things change pretty quickly, and this particular topic, most of the work on it has been done in the last twenty years." Once again, we see that ecologists' domain knowledge provides important information for decision making that reduces uncertainty and helps them to achieve standards of scientific practice.

Since the guidelines for meta-analysis, in particular, are sensitive to bias in the collection of data, the published literature was a common method to set boundaries around data gathering. Cal collected data from three journals to use in a meta-analysis he conducted recently. He was concerned about biasing his results based on the journals he chose, especially since he works in the tropics, an area of ecology that is not covered evenly by ecological journals. So, Cal chose a well-known but relatively new journal on tropical ecology to anchor his sampling design, which in turn, established an objective time frame for his collection of data.

> I wanted a balanced design. … It was convenient that it was only twelve years old. It was convenient because it set this fixed time limit. Something that gave me a time limit that I didn't subjectively choose, which is another good thing to have is as much objectivity as possible. So, I had three journals that were picked for getting the balanced design. And then I had time limits based on the age of one of the journals. And I just went with that.

Nathan, a data manager, noted, "Doing a meta-analysis based on just published literature is a poor substitute for having a complete archive of all the data that have been collected."

Ecologists were aware that their techniques had limitations. If they could, ecologists addressed these deficiencies. For example, meta-analysis methods are attentive to biases in data collection that arise based on how the data were collected, such as from selected journals or from what a scientist has on hand. The methods also provide means to address limitations based on data that are unobtainable, which is referred to as "the file drawer problem" (Cooper and Hedges, 1994). Meta-analysis methods provide statistical techniques to address potential data collection biases. Nancy noted that other biases are more difficult to deal with.

> There's the publication bias. Before that, there's even a study bias. You choose to do a study where you think you're going to find competition, or you think you're going to find facilitation. Those are tough ones, and partly you just need to incorporate that into your interpretations.

Often, it was not possible for ecologists to address methodological limitations in data collection, but as Cal noted, this is the case with all research methods. Ecologists are accustomed to imperfect research environments because not all variation in nature is controllable. Relying on the published literature to gain access to published and unpublished data provided one way for ecologists to bound their data collection. As I noted earlier, peer-reviewed publications are also a good source of information about data, although they do not guarantee data quality.

Geography was another frame that ecologists utilized to bound their collection of data for reuse. Sometimes, these boundaries were defined by an existing data set to which ecologists had access; this was the case for David, Susan, Ellen, and Charles. Alan used reports in bird journals, a bird-banding database, and museum records to gather as many bird sightings as possible from a particular geographic area. When I asked Alan if these sources covered the data that were available, he said, "Yes. Of course, one never

knows that for sure, but I feel like we got most things that were out there." The key for Alan and for other ecologists was to choose an approach that met research standards and that could be defended publicly. For the most part, ecologists did not relate this directly, but I concluded it from the ways in which they discussed their methods and choices.

Individual Dimensions of Gathering Data for Reuse

Efforts to bound data collection may explain partly the lack of emphasis that ecologists gave to personal connections as a source of data. Nancy, for example, noted that she preferred to take a systematic approach versus "whatever I have in my files," which explained her discomfort with a sentence one of her students included in a manuscript.

> She's got a sentence in here that I feel very uncomfortable with. Ok. "To obtain sources, we searched Biosis between 1995 and November 1999 using the keywords... In addition, we added references from our files." I said, "Lora, you're going to get creamed on that one." So, she added the sentence: "Although not the most systematic approach, this increased the time period from which references were drawn, and it increased the number of ecological relative to agricultural studies." So, those are both important things to do, and I agree in this case that we should use them. Ah, but I just... that sounds so... "We added references from our files..." is sort of like, "Well, we happened to have it around."

Less commonly, ecologists relied on their individual knowledge to help frame their acquisition of data. Stephen's approach of limiting his collection of data to research programs well known to him was the most notable in this regard. When it comes to issues of trust, distrust is often emphasized because it is used to discount or eliminate, whereas trust is an unspoken confidence. Distrust can lead to an outright negative judgment, but it can also manifest itself as a more subtle "lack of confidence" based on an individual's perception of the way things work. When I asked Charles about the

difference between locating data for reuse versus finding other information for his research, he distinguished public sources from private ones.

> They are different. They are really different. The published stuff is all somewhere in a library somewhere. There are people out there who want you to have it. While data that has been collected by individuals, getting that is much more having to do with personal relationships--trust and people's willingness to share and all of that--which is a whole different set of issues.

Some ecologists' preconceived notions about the difficulty of obtaining data from people they did not know kept them from asking for data. For example, Nancy, an experienced ecologist, followed advice from others in deciding not to contact authors of published papers for additional data who told her, "You get very little response." In addition, as an ecologist with lots of her own data, Nancy was sympathetic to the time involved to fill data requests.

> I think we all say, "Oh, well, it shouldn't be any problem!" I've gotten requests for data, and it's a paper I wrote 10 years ago, and they aren't even in compatible format with the computer I have now, and who knows where it is, and it's going to take me a whole day to dig it out and put it in… and I just don't do it.

Even Andrea, who felt that in the academic environment, "You are likely to get an answer if someone knows you, and you are not likely to get an answer if you contact somebody out of the blue," noted that the authors of a paper may not have actually collected the data themselves, which could make it hard for them to answer specific questions about how the data were gathered. Catherine acknowledged that she and her coauthor had no idea what kind of responses they would get to their data requests from other scientists. Their experience was positive, however, and Catherine noted, "They were all wonderful. They were all happy to oblige." Ecologists' choices reflect the

variety of personal perceptions that drive their decisions. These impressions continually accrue and change and are unique to individuals.

Comments from ecologists indicate that they absorb cues from their teachers and mentors about the values of other scientists to form notions about the "way things work." They factor these cues into decisions about where and how to look for data to reuse and to form general opinions about the trustworthiness of data collected by others. Ecologists with recent graduate school experience indicated that data sharing was not a formal part of their academic training, but they formed opinions about the data sharing values of other scientists from the local culture and from scientific ideals. Ecologists described their local environments as open to sharing, but they also learned, as David said, "it was a 'be careful' thing," and as Susan described it, "not all places work this way." Cal learned that even when scientists share data, they sometimes "hide" them.

> I know that people might want to have the data available, but I think they are also afraid of… They still want to use the data themselves. So, they are afraid of just putting like a spreadsheet on there that somebody could just easily do some analysis on right away. So, they want general information out there, but I don't think they want people analyzing their data right away. *(segment cut)* So, I think there is a little bit of intent for the database to not be easy to use for analysis.

As less-experienced ecologists gained more experience and had data of their own to share, they gained a personal perspective on these sentiments.

Data Gathering Methods: Further Rationales

Besides addressing the need for scientific schemes for sampling data, ecologists' data acquisition methods appeared to increase their chances of obtaining data and to decrease their concerns about errors associated with the secondary use of data. Although

many ecologists collected data from multiple sources because the data they needed did not reside in one database, the collection of small amounts of data from more than one source provided a couple of advantages. The acquisition of low volumes of data from multiple sources reduced error, and this method demanded only a small portion of data from each sharer. Thus, some typical data sharing concerns, such as scientists' worries that their data sets will be reanalyzed in order to disprove their conclusions, were addressed upfront. Data requests that included the reasons for seeking the data also anticipated such concerns. The amount of data requested and the proposed use of data may have influenced their sharing by others. For instance, Alan attributed his success in obtaining data to the specific nature of his data requests and to the fact that he was looking for individual data points that were not worth much individually, but that were valuable collectively. Michael and Charles, on the other hand, were unable to obtain large amounts of data from several sources. The reasons given in these instances included an outright refusal to share, a restriction on distribution stemming from a country's policies on the release of data outside its borders, and a scientist's inability to share because he could not find the data. Michael was suspicious of the reason given in the latter situation.

> There was one guy who had data on five hundred lakes in the southern United States, and he couldn't find the data, or so he claimed. To me if you are going to sample five hundred lakes you hold onto that data; it's an awful lot of work to go to.

Another explanation for a lack of sharing in instances such as these is based on the amount of data requested.

The source of the data played a role in data acquisition choices, too. Data related to publications are considered public, and this made them easier to request; it may also

have provided subtle cultural incentives for sharing. The norms of science dictate that the data associated with a publication should be available upon request. Even though this norm is flaunted occasionally, according to the ecologists that I interviewed, it is a recognizable scientific ethic.

Other Factors that Influence Data Sharing and Secondary Data Use

The literature says that ecological data have a high level of ownership, and this is seen as a significant barrier to their sharing. Not surprisingly, the ecologists I interviewed were strongly in favor of data sharing. In looking purposefully for ecologists who successfully obtained and reused data, it may appear that I have downplayed the more intractable social and cultural hindrances to data sharing that are discussed elsewhere. However, the ecologists I interviewed confronted and recognized many of the same obstacles discussed in the literature, and these factors emerged when I asked them about their experiences in acquiring data for their specific projects, about their experiences in sharing their data with others, and about their general opinions on the topic. Nearly all ecologists encountered or knew of situations in which other scientists were unwilling to share their published or unpublished data. They acknowledged the lack of reward for sharing data, and they recognized issues of data ownership and its relationship to scientific advancement. As Ellen said, "It is sort of human self-preservation, I think, to not just necessarily be driven solely by, 'This is the right thing to do.' So, it is a little bit ugly. But it is there." Social and cultural factors do play a role in data sharing, and they can make the reuse process more difficult, but they do not change the overall approach that ecologists take toward the secondary use of data. Ecologists

choose methods that meet standards, exclude subjectivity, and satisfy personal tolerances for uncertainty. Ecologists are adept at drawing from their knowledge to meet data reuse challenges, which obscures their approaches and rationales.

Ecologists' mental facility for data reuse, when viewed in combination with social and cultural factors, points to a tough struggle for quick changes to data sharing approaches in ecology. My interviews show that less-experienced ecologists form attitudes about data sharing from their mentors. Even when these are positive, they are limited by wider distrust, a lack of reward, and by the nature of the practice of ecology itself, including the technology, standards, and practices that accompany work in this field. Additionally, since less-experienced ecologists cling to norms of scientific behavior, they may appear to signify a change in the wind, but this is more likely due to a lack of first-hand knowledge about other scientists or to close insight into "the way things work." Certainly, the practice of ecology will continue to evolve, but as Nancy said about her students, "They'll be better at wanting to use other people's data that's for sure!" She also wondered, though, "Will they be better about their own metadata?" Experienced ecologists, especially those who are educators, recognize that significant change lies in the enculturation of the next generation. As Ed said, " I think what it's going to take is more and more people will get the obligation or responsibility, and that's something we have to train in our students."

Once the sources of an individual's knowledge are identified then the possibility exists to transfer portions of that knowledge to others explicitly. Of course, the task is not simple, and the challenges form one focus of knowledge management research. Domain knowledge fosters the sharing of data within the same community, but it is

unable to totally dissolve distance and distrust because of tacit judgments and individual knowledge. These knowledge aspects can be verbalized, but they are difficult to build explicitly into formal data sharing systems. Data sharing challenges also arise from the fact that data are a liability, especially when they are reused. This explains partly why personal interaction and networking are prevalent in fields such as ecology. Since data are the basic building blocks of scientific argument, researchers must understand them, or they risk misuse. The externalization of knowledge is also hampered because individuals are unwilling sometimes to let go of their knowledge on a wide scale because they distrust others who might capitalize on or use representations of that knowledge. In the case of data sharing, data owners are suspicious of others' capabilities to use the data, of others' motives, and of the unknown. Distrust is a mutual feeling since secondary users of data are often wary of data collected by those they do not know. Personal networking is a way for individuals to control access to their knowledge and for those who seek that knowledge to establish trust with the provider.

Data Managers: A Different Set of Standards

Throughout this chapter, I have shown how ecologists employ their knowledge to anticipate and to overcome the challenges associated with the secondary use of data. In many ways, ecologists' methods are effective at dealing with the collection and integration of data gathered at different times, in different ways, and for multiple purposes. Although membership in a community of practice is unable to totally dissolve issues of trust and distance, domain knowledge is a powerful base for sharing data among members of the same epistemic culture. However, increasing the amount of available

data and scaling up the infrastructure for sharing ecological data requires approaches that overcome limitations stemming from cultural, technical, and social factors. Chief among these challenges are issues related to data integration and to the lack of resources or incentives for ecologists to document data. Intermediaries are one means to address some of the factors that limit the availability and integration of data. Markus (2001) defined an intermediary as an individual "who prepares knowledge for reuse by eliciting it, indexing it, summarizing it, sanitizing, packaging it, and who performs various roles in dissemination and facilitation" (p. 61). In this section, I draw from my interviews with data managers to explore their role as intermediaries.

Data managers and ecologists have variant goals for their work, and they adhere to different standards. Below, I contrast the ways in which these differences affect how each community views and treats data. My analysis illustrates how standards can contribute to distance as well as help to overcome it. If ecology is to expand its data-sharing infrastructure it must better utilize intermediaries, and in order to do this, the distance between the standards of ecologists and data managers must be narrowed. If more data are to be available for sharing, then each side must work to collapse the distances between them. As Porter (1995) observed, "There can be no consensus in a world of specialists, all attempting to follow strictly the rules of their own discipline--all in this sense forms of local knowledge" (p. 215).

M. Lynne Markus (2001) proposed three major roles in the knowledge reuse process: knowledge producer, knowledge intermediary, and knowledge reuser. The important work that data managers handle as intermediaries is evident when some data sharing challenges are viewed at a larger scale. All groups of interviewees concurred that

the most intractable hindrances to data sharing are cultural as opposed to technical. Markus indicated that intermediaries have a role to play in addressing a variety of challenges. According to Markus, producing information that meets the needs of reusers requires a great deal of effort, and there is little incentive and few resources for knowledge generators to do so. In these cases, Markus argued, reward for knowledge producers must increase and some of the work of packaging and disseminating must be shifted to intermediaries. In order to understand the role that data managers play as intermediaries and to contrast their views with ecologists, I introduce the data managers I interviewed, and I analyze my conversations with them in terms of their standards.

Three of the four individuals I interviewed--Mark, Sandra, and William--are currently in positions that focus on the management of data. They work primarily with scientific data, but they also manage bibliographic and textual databases and administrative systems, such as accounting and personnel databases. Nathan, the fourth interviewee, previously worked in a data management position. He has since taken a larger role within his organization where he continues to be involved in ecological data issues. Two of the four data managers have Masters degrees in biology, but all are knowledgeable about the science, which includes time spent in the field collecting data. None of the data managers were associated with the ecologists I interviewed. I made this decision in order to protect the identity of the members of each group from one other.

The Purpose of Work

Individual ecologists collect their own data and reuse data gathered by others in order to address specific research questions. Thus, for the most part, their concerns in

regard to data management are short-term and informal. Data managers, on the other hand, deal with data sets, especially data stored in computers in the form of relational databases, and they are concerned with issues of longer-term storage, retrieval, and preservation. "Management of the data stresses their accessibility and integration" (Baker, Benson, Henshaw, Blodgett, Porter, & Stafford, 2000, p. 964). Thus, data managers are concerned with structure, format, relationships, and processes. The broad areas defined by Baker et al. (2000) reflect the responsibilities of the data managers I interviewed.

> The term "data manager" developed to describe an individual dealing with specific data sets. A data manager may prepare, calibrate, document, and assure the quality of raw data. Additionally, the data manager may develop techniques and formats for exchanging data with a central site and for eventual distribution or preservation. Current research often requires data sets to be integrated from multiple projects and sources into intercomparable groups of data sets. In fact, the term "information manager" may better describe the individual dealing with the broader aspects of data (p. 965).

Because data managers work largely with information stored in a computer, they must follow specific technical standards. Data managers frequently used the word *mechanisms* to describe the nature of the work they do and to refer to the processes associated with planning for, entering, and maintaining data in a relational database. Sandra described her role in managing the data associated with a multi-investigator ecosystem study. Sandra's experience with this project was frustrating because a formal data management system was not put in place until the third-year of a five-year project.

> I came in, like I said, in the third of a five-year, so we spent a year trying to set up the system and the mechanisms to automate it--because that was the goal--because just formatting data can be a true task in and of itself. So, we were trying to format, provide information, put out maps and then set up these mechanisms. And so, what we did is try to capture as much as we could from the existing data that was collected in three years and put it

> together, maybe reformat it--like retrack some of the sample locations and readdress those locations, to make them accurate. And then get people online for the new mechanism. And we went through it and tried to approve a chain of custody... set up a chain of custody system and code various things that were sampled so everything could relate… like substrate had a certain code... Try to make it more efficient. But it also gave them more work because they weren't capturing everything. All the information, we tried adding more metadata, to have longevity in the data.

Integrating the data into a common system was hindered by a lack of planning, by the variety of approaches scientists in the project were using to manage their data, and by an absence of incentives to participate in a centralized data management system. For example, location codes were not standardized, and the information necessary to relate data from different scientists was not captured. Sandra likened this to trying to erect a house without a plan: "It's like you're building a house, but you already built it. You don't build a house without architecture plans and that's what the design of a database is." Part of the difficulty stemmed from the fact that scientists were building a different structure, and its foundation was centered on answers to individual research questions. Data managers recognize the gap that exists between them and ecologists in relation to the purpose of their work.

In general, the standards data managers follow are different from those that concern ecologists who manage data for their own use, but they serve an important purpose, particularly at larger scales. Since ecologists are not taught to manage them they have, as Michael described it, "all kinds of crazy ways of storing their data." At small scales, these idiosyncrasies are addressed somewhat easily. At larger scales, however, they are significant obstacles to the integration of data from multiple sources. William described problems he encountered in working data in his organization: "Some of the structures are just awful and have cost literally months, if not years, of time on

some of the long-term reference stand data that we had to restructure…." Mark noted, "The processes can really be impaired by the fact that the design isn't correct, and so it's important to get the design right."

Data managers recognize the importance of planning ahead for data integration, and they note the advantages of involving them at the start of a project. Advance planning includes agreement on methodological standards and on data structures. William noted that it is also important to document multiple and changing purposes. "If you don't document those sorts of things as you go along, you will really miss out on some of the whole perspective, why certain things were done in certain ways and why it was collected." Like ecologists, data managers recognize that the multitude of purposes for which ecological data are collected complicates and influences their secondary use.

Ecologists are interested in obtaining specific data to address particular questions, and therefore, they typically want raw or summarized data, equivalent to spreadsheet-like reports from a database. Data managers work primarily with relational databases. Relational databases provide many advantages, such as efficient query engines, sophisticated reporting capabilities, and the ability to more easily revise database structures to add new relationships. Data managers talk about relational databases in terms of normalization, syntax, schema, and entity relationships. The complex, but powerful capabilities of relational databases can make it difficult to "see" the data, a factor that conflicts with ecologists' abilities to obtain the data in the form they want them. William related an interaction with a scientist that illustrates the difference between the way data are structured in a relational database and how they are reported. A

scientist in William's organization wanted to integrate climate data from ten sites across the country in order to compare a number of different parameters.

> So, he brought this to the data managers, and of course, he was trying to dictate exactly how we would structure the database and everything, and so we kind of said, "Well, the structure--the way we structure it, and the way we report it," which is really what he was interested in, "are separate issues." That was a really good thing to get past. We realized, "Okay, we could do this in anyway we want." Structure it anyway the data managers thought it was most appropriate--on more of a normalized relational form--and we wouldn't have to worry. Then we would generate the type of spreadsheet like Scott Brown wanted.

Until they resolved their distances, the ecologist perceived data manager standards as an obstacle to his goal. In reality, however, it was William's ability to structure heterogeneously formatted data into one database that could be reported in many ways that helped the ecologist to achieve his goal and that also made the data more usable for other potential users.

Quality Concerns

Ecologists' concerns about the quality of the data they reuse revolve around understanding how the data were collected, including judgments about whether they were gathered skillfully. Data managers are responsible chiefly for aspects of data quality that relate to their structure and storage in a database. This includes issues such as the consistency and validity of codes and generic computer testing to scan for completeness and duplication. Some of these quality issues are related to information, or metadata, that aids comprehension of the data. As Mark said, "Physical format is certainly an issue. But then, understanding the database--and that's where metadata comes in. How it was collected, and what are the units of measurements, what do these codes mean?" Mark

distinguished between a data manager's and an ecologist's responsibilities for data quality.

> I guess the primary place where I have to do with quality has to do with the consistency of coding and the linkages between the different parts of the database. … As far as how [fish] lengths are measured on a boat and so forth, that's kind of out of my--a data manager's view.

Or, as William summarized it, "The responsibility in the field…has to belong with the scientist." The data managers I interviewed recognized all aspects of data quality since they had spent time in the field themselves or had worked with ecological data for a long time.

Shared Perceptions

Ecologists and data managers recognize many of the same cultural, social, and technical challenges to data sharing. Social and cultural issues include a reward structure that provides few incentives to share data and that encourages short-term thinking. Ecologists and data managers also agree that it is difficult to document data sets so they can be used for multiple purposes, and they concur that strategies to preserve data must be driven by scientific questions. William summed up the importance of scientific drivers to data sharing.

> You have to have the science... I mean while it's the information manager's idea pretty much to do this, if you don't have the scientist really wanting that information or wanting to do that sort of comparative work, the whole thing is really not going to fly. You really have to have that science background pushing that sort of project, or it just doesn't... It just goes nowhere.

In addition, ecologists and data managers agree that no one understands the data better than the scientist who gathered them, and that it is that scientist who must document

those aspects of the data. Nathan described the limitations of those outside the data collection process to fully document a particular data set.

> They can put value-added metadata with data sets. But they can't replace the information that was associated with the collection of the data. There is no way. So, they can take existing metadata: titles, abstracts, information about the data set. They can look at the data itself, and they can present some sort of an analysis or categorization of that that is useful in like resource discovery or in other ways. In other words, some value-added product. But when it comes down to it, the only people that know what happened on March 23 in the field, if it is not written down on the data set or metadata, are the people who did it. And that is the way it is. The principal investigator, the research technicians that were in the field, the data managers that were involved in quality assurance analysis--those are the people that have to provide the documentation of the process. It is not something that you can reconstruct.

Ecologists have few incentives to document their data for other users. Publication is the most formal way in which most ecologists document their work, but it is not always sufficient to enable data reuse. Markus (2001) noted that a great deal of effort is required to produce documentation that meets users' needs and that knowledge producers are frequently expected to produce high quality information without the incentives or the resources to do so.

Ecological Data for Whom?

There are several factors to consider in scaling up the sharing of ecological data. First, is to identify whom data resources are meant to serve. Markus (2001) noted that the effectiveness of knowledge repositories is contingent upon meeting the needs of knowledge reusers. She identified four types of knowledge reusers based on their distance from those who produced the knowledge "where distance is measured in terms of shared knowledge" (p. 63). She theorized that the closer reusers are to the knowledge

producers the more they can understand the contextual information in the documentation and "can successfully reuse the raw, unprocessed records that are created as a by-product of knowledge work" (p. 68). In fact, Markus noted that the information could be difficult to reuse without its context, a conclusion supported by my interviews with ecologists. To reuse ecological data, ecologists rely heavily on information that allows them to put their field-based knowledge into play. Most important is the information that enables ecologists to reconstruct the original collection of the data they use secondarily. Since the knowledge necessary to document data gathering resides with the data collector, then it makes sense to encourage this individual to document those aspects of the data in a way that will allow other members of the same community to reconstruct the process. As Markus's theory postulated, however, knowledge generators primarily document for themselves. Records kept for a knowledge producer's own use are informal and biased toward short-term needs, and they tend to rely on their memories, not always successfully, for records of things with longer-term value.

> Since knowledge workers often have difficulty anticipating distant future needs for information, their records tend to be biased toward short-term needs. … However, they tend not to keep records of things with longer-term value (such as details of why issues were resolved as they were, which are useful when a new team member is brought on board, or when the system is upgraded). For such long-term matters, they rely on their own memories. Unfortunately, they often forget, and the organization loses access to knowledge when team members depart (p. 73).

Conversely, Markus proposed that the more dissimilar the reusers are from the knowledge creator in terms of shared knowledge, "the more difficulty they may have in defining the search question, locating and selecting experts and expertise, and reusing even carefully packaged knowledge (p. 70). Markus stated that such users require information "that has been carefully decontextualized: Otherwise they will drown in

unnecessary, unhelpful or conflicting data" (p. 70). Markus's theories in regard to the amount of documentation required for various levels of users fit with other observations in the literature and with comments made by ecologists and by data managers, such as William:

> No data set is particularly easy to get the entire piece of information across. If you are doing it for someone in your organization, you only have to really document the more finely detailed things. If you are documenting it for somebody in another organization, but who is familiar with the work, then I think you step up a level in terms of the purpose that you are collecting and your designs that have been used. Where the local people may know those sorts of things, you need more and more metadata to be able to get it across to that group. But if you are really preserving it for the long-term so that it is totally out of the hands of the original collectors, then I think you are looking at really following one of these metadata standards.

At this time, ecologists primarily document data for their own use or for other members of their community with whom they share knowledge. As Markus noted:

> The usual expectation is that knowledge producers will author repository documents for use by others (whether community-of-practice members or novices). But this expectation contains two problems. First, the records knowledge producers make purposely for their own use are not likely to meet the needs of others. Second, the records knowledge producers make for others may not meet their own needs, and therefore, they may not have adequate incentives to produce quality documents that meet the needs of others.

In making decisions about which data to document, the question, perhaps, is not what data sets are most usable within ecology, but which are most usable outside of it. If more documentation is required for ecological data reusers that are farther away from the knowledge generator, then those with the least shared knowledge require the most documentation, such as those found in complex metadata standards. Intermediaries, such as data managers, are necessary if data are to be available for such dissimilar users. Ecologists cannot be expected to meet the needs of these distant “others.”

This chapter analyzed distances within one scientific community and between ecologists and data managers. Data sharing systems intended to serve all potential secondary users must successfully span all distances. Yet, traversing these gaps requires different and multiple approaches. Distances also arise for varying reasons; some of these can be addressed more easily than others. In the next chapter, I discuss the implications and limitations of these findings, and the areas they point to for additional research.

Notes to Chapter 4

[1]I use the phrase *data set* to refer broadly to data in both print and electronic form. I use the term *database* when I refer to data stored in a computer.

[2]When another scientist was the primary method used to locate data, the process of finding data was not an issue; the ecologist was led to data by someone else.

[3]Bill, Katherine, and Nancy conducted meta-analyses.

[4]All ecologists interviewed were the first author of the published paper, except for Andrea, who was the third author and the chief person who gathered the data that were reused.

[5]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.

[6]One of the data managers I interviewed had extensive experience working with ecologists collaborating on projects that required the integration of data from multiple sources. It was his experience that, at the start of a project, these groups did not know what data they needed or wanted, nor did they understand the issues involved in combining data from more than one source. He also noted, however, that awareness of these issues varied among groups and individuals. The ecologists I interviewed appeared more certain about their data needs, and they recognized the challenges of integrating data. The differences in the data manager's experiences and my interviews could be related to three factors. For one, collaborating groups might reflect different dynamics and expectations. Second, this variance could reflect a limitation of the interview method. Ecologists spoke with me about a past experience, and so they may have left out information about these early stages. Since this issue was not integral to my study, I did not focus my questions on it, and so they had little reason to relate it. Lastly, the ecologists I interviewed may not be representative of other ecologists in this regard. I discuss this limitation further in the following chapter.

[7]NPP stands for net primary production.

[8]One resource that ecologists mentioned is, The handbook of research synthesis; see Cooper and Hedges in the bibliography. Undoubtedly, there are other sources. For example, Cal told me he learned how to do a meta-analysis by reading about it and Katherine said she acquired a set of papers on the topic.

[9]The specific name of the database that Charles mentioned was generalized in the quote in order to further protect his identity.

CHAPTER 5

CONCLUSIONS

This concluding chapter serves several purposes. I begin by reviewing the major findings from my study. Next, I discuss what is learned about data sharing and reuse based on these results, and I relate it to the conceptual foundations provided by Theodore Porter and Bruno Latour. Third, I suggest several areas to which my findings are applicable, and I offer some specific recommendations for implementing them. Fourth, I analyze several limitations of my study, and last, I examine areas in which additional research is needed.

Overview of the Major Findings

My study was unique in examining successful data sharing experiences within a particular scientific community. By taking this approach, I was able to test assumptions about ecological data and to show how ecologists overcome the challenges of reusing data gathered at different times, in different ways, and for multiple purposes.

My main research question was: What are the experiences of ecologists who use shared data? I defined ecologists' experiences by the following subquestions: How do ecologists locate data and assess their quality?; What are the characteristics of the data they receive?; What information do they need to use the data?; and What challenges do they face throughout the process? The literature review pointed to informal modes of

data sharing characterized by personal interaction, which exist because of the complex nature of ecological data and because ecology is typified by small-scale, single-investigator studies and high ownership of data. My results show that while personal interaction and cultural factors play a role in nearly all experiences, neither changes the overall approach that ecologists take throughout the process. Ecologists choose methods to gather data for reuse and to make decisions about data acceptance that meet community and individual standards and that can be defended publicly. Ecologists' decisions regarding what data to reuse are influenced by a combination of domain knowledge, personal tolerance for uncertainty, and individual knowledge.

Shared Knowledge and Practices

As members of a community of practice, ecologists share an interest in a domain of knowledge and a set of approaches that help them to 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, p. 47). Knowledge of their domain, which ecologists acquire as part of their enculturation to the field, directs their choices and serves as a standard for ecologists because it is well established and familiar; it conforms to the prevailing norms of scientific practice; and it names "a set of strategies for dealing with distance and distrust" (Porter, 1995, p. ix). The aspects of community membership that figure most strongly in ecologists' reuse of data are informal knowledge gained through fieldwork, formal disciplinary knowledge, and standards of scientific practice. In contrast, *individual knowledge* is particular to individuals; it consists of insights and

perceptions that are not shared on a wide scale. In the context of Theodore Porter's ideas, domain knowledge is objective and distance-spanning, whereas individual knowledge is subjective and distance-creating. Together, these concepts of knowledge help to explain ecologists' experiences as secondary data users.

Ecological researchers who reuse data are users of existing data as well as generators of their own data. It is important to recognize both roles in order to understand ecologists' experiences and to account for the seeming dichotomy between the knowledge ecologists employ to understand data and the approaches they take to gather data for reuse and to publicly frame their experiences. The first is driven by informal knowledge and disciplinary expertise and the latter is propelled by formal norms of scientific practice.

Informal Knowledge

First, as users of existing data, ecologists are attentive to understanding the data they reuse at the same standard as data they collect themselves. The knowledge ecologists gain from gathering their own data helps them comprehend the data they reuse. As Roth and Bowen (2001b) noted, and as my findings confirm, fieldwork performs an important function in shaping ecologists' formal and informal knowledge. My results show that the informal knowledge ecologists acquire as collectors of their own data in the field or laboratory plays the most important role in their secondary use of data because it helps them to understand data. The ability to comprehend data is the key to their reuse, and ecologists rely heavily on knowledge from their own fieldwork experiences in order to "reconstruct" data they did not collect themselves. I define the term *reconstruction*

broadly to describe all the processes ecologists employ to mentally reassemble the original collection of the data they seek to reuse. Field experiences, along with disciplinary knowledge, enable ecologists to recognize the relationship between research purpose, methods, and data, which they call upon to help them determine whether existing data will meet their needs and purposes. Ecologists depend on information about data that enables them to put their field-based knowledge into play. Therefore, it is not surprising that the greatest hindrance to data reuse is ecologists' inability to comprehend data.

In addition, data comprehension is related closely to many assessments ecologists make about data quality, and so ecologists recognize and share some of the same criteria for judging data. Fieldwork familiarizes ecologists with the vagaries of particular types of data and provides them with the ability to visualize potential points of data collection error. They employ this knowledge to assess data quality based on factors such as the variability of specific parameters in nature, valid and invalid data ranges, and the level of difficulty of collecting particular data. Furthermore, ecologists describe their experiences in the field and laboratory as giving them a "sense" for data, which they draw from to understand and assess them. This tacit dimension of knowledge, which ecologists recognize and can describe, is difficult to externalize. This helps to explain why standard methods alone are inadequate substitutes for indicators of data quality. Standard methods provide clues about how data were obtained, which aid understanding, but they do not tell a secondary user if the measurements or observations were gathered skillfully.

Formal Knowledge and Norms of Scientific Practice

Second, as generators of data, ecologists are heedful of future public scrutiny, and so they work hard to follow norms of scientific practice in their gathering of data for reuse and in their reporting of results based on secondary data use. Standards for scientific practice, which emphasize objectivity and exclude subjectivity, are part of domain knowledge and influence strongly the approaches that ecologists follow throughout the process and the ways in which they describe their choices. Ecologists recognize the importance of knowledge they gain in the field, but they rely primarily on formal notions of scientific practice to frame their experiences and to direct their search for data to reuse because informal knowledge is not acknowledged publicly in the context of "real science" (Roth & Bowen, 2001b, p. 477). Ecologists' collection of data for reuse mirrors the standards that guide the gathering of their own data in the field or laboratory. The emphasis on objective norms leads ecologists to seek strategies that *bound* their collection of data, in order to draw data to reuse from some representative sample. The ecologists I interviewed relied primarily on the literature, geography, and existing databases to frame their collection of data for reuse. Ecologists also choose methods that increase their access to data, that reduce the potential for error, and that provide believable rationales for their choices.

The domain knowledge that ecologists carry with them also assists their choices about where to look for data and helps them to integrate data from multiple sources. Ecologists' expertise influences their approaches throughout the process and helps them to pose research questions. Their research questions rely on data with which ecologists

are familiar, and in turn, this acquaintance assists their comprehension, assessments, and acquisition of data.

Individual Knowledge

Shared knowledge and practice is a powerful base for sharing data within the same community, but it is unable to totally dissolve issues of distance and distrust. There are two main reasons for this. First of all, as I noted above, judgments about the tacit skills of other data collectors are an important aspect of ecologists' data reuse decisions. These assessments are based on knowledge acquired through fieldwork or on individual knowledge. Field-related knowledge is hard to externalize. Adding to the difficulty is the fact that formal reports of scientific research remove the informal aspect of ecologists' work. Thus, publications, which serve as a popular source of data, can downplay or eliminate information that aids data comprehension. The second reason that domain knowledge alone is unable to eliminate distance and distrust is because ecologists' decisions are influenced by individual knowledge, such as first-hand acquaintance with another's skills, perceptions of the values of other data collectors, and views "of the way things work." Individual knowledge consists of unique and personalized insights and connections, and ecologists employ it to lessen concerns about data quality, to improve their access to sources of information about data, and to include or exclude data from consideration for reuse. Individual knowledge, however, plays a secondary role in data comprehension. Personal insights and connections can provide access to information that aids comprehension of data, but they are not a substitute for disciplinary knowledge or fieldwork experience. In summary, while individual knowledge and connections

sometimes play an important role in ecologists' choices about where look for data and about what data to reuse, it is a subordinate driver of reuse decisions.

Since assessments about an unknown data collector's tacit skill can be difficult to make based on available information, it is not surprising that individual knowledge of the skills or values of other scientists, whether positive or negative, enters into data reuse decisions. These judgments, however, are private and are not acceptable to share on a wide scale unless they are couched in a scientific rationale. Choices based on individual knowledge can be at odds with objective, scientific standards, which is one reason that two different ecologists will not necessarily make the same decisions about the same data. This tension also explains why ecologists rarely mention individual knowledge as a significant factor in decision-making. Personal tolerance for uncertainty also explains partly why ecologists make different choices, but it is a more accepted rationale because it can be tied to scientific standards. In summary, ecologists choose approaches to gather data for reuse and to make decisions about data acceptance that meet community and individual standards and that can be defended publicly.

The Limits of Knowledge

Ecologists' concerns about the use of data that they did not collect themselves are put to rest by a combination of factors, the most important of which is an ability to understand the data. Ecologists' knowledge exists for them to use at any time, and thus, acquiring data, understanding them, and assessing their quality can occur simultaneously and are often part of an iterative process. Their intellectual adeptness obscures the rationale for ecologists' choices, masks some of the considerations that pervade the

sharing of data among members of the same culture, and makes it difficult to draw distinct lines between each stage of the reuse process. Despite their ability to overcome many challenges, certain factors limit the sharing and reuse of data. These include an inability to understand data, a lack of trust in another's skills or values, and the challenge of locating and integrating data collected at different times and scales and for many different purposes. Data managers, as one form of intermediary, can help to address some of these limitations. Following Markus (2001), I define *intermediaries* as those who prepare data for reuse by eliciting, organizing, storing, sanitizing, and/or packaging data, and by performing various roles in dissemination and facilitation (p. 61). Ecologists and data managers have variant goals for their work and different standards, which create distances that must be overcome in order to improve mechanisms for sharing data.

Discussion of the Major Findings

By looking closely at data reuse within one scientific community, my study revealed factors that both assist and complicate or hinder data sharing among members of the same community. The conceptual foundations of Theodore Porter and Bruno Latour provide a useful frame in which to view data sharing generally, and in particular, the results of this study. Porter's ideas of distance and standards and Latour's notions of reduction and amplification help to explain ecologists' choices throughout the process. The use of this theoretical lens to analyze data sharing within a community of practice led to several interesting findings, and it helped to explain factors that keep communities, such as ecologists and data managers, apart. The implications of my results extend beyond ecology and add substantially to our knowledge about data sharing.

Homogeneity and Heterogeneity in a Community of Practice

Research on the social aspects of digital libraries has emphasized the concept of communities of practice to describe and to differentiate between groups of people, such as disciplines and professions (e.g., Bishop et al., 2000; Van House, 2002; Van House et al., 1998). Although Wenger (1998) was careful to state that it is not intended to convey complete homogeneity, the community of practice framework has been employed primarily to highlight the sameness of communities in order to contrast one with another. Since my study investigated the secondary use of data among ecologists, one community of practice, understanding my results relies on remembering that participation in social communities can "involve all kinds of relations, conflictual as well as harmonious, intimate as well as political, competitive as well as cooperative" (Wenger, 1998, p. 56). As Wenger summarized, "Indeed, what makes engagement in practice possible and productive is as much a matter of diversity as it is a matter of homogeneity" (p. 75). Ultimately, participation forms individuals, but it also shapes communities. In the case of ecologists, shared membership in a community of practice is not always sufficient to overcome barriers that arise. Diverse scientific viewpoints, such as epistemological frameworks, create differences that can be impossible to resolve. Additionally, social and cultural issues, such as a lack of reward, concerns about misuse, and the establishment of priority hinder data sharing. My conclusions augment findings from previous research that acknowledged the significance of the social issues of data sharing, such as ownership, rewards, and cultural norms. The multi-faceted nature of a community, when viewed within the theoretical framework of Porter and Latour, leads to

several key findings regarding standards and their ability to span distances within and among communities.

Standards as Distance Spanners

My results show that we need to broaden our definition of what count as standards. At the same time, we must recognize that the ability of standards to traverse distance is not equal and is not stable. Formal standards travel farthest, but within a community of practice, information that enables members to put their informal knowledge into play is critical. My interviews with ecological researchers demonstrate that the distance spanning capabilities of "more standard" standards, such as research methods, are affected by informal knowledge--at least when it comes to the secondary use of data with which an ecologist is familiar.

Berg (1997) argued that, in medicine, protocols contribute to the loss of information that is difficult to explicate or quantify, and thus, standards reinforce the notion that information that "can be made explicit is more important, more scientific, more of value than that which cannot be (or is not) made explicit" (p. 1085). My findings show that informal knowledge, which is often sacrificed for the sake of standardization, is a prime component of data reuse. For ecologists, informal knowledge gained through fieldwork is the key to their reuse of data. The challenge in drawing on ecologists' informal knowledge as a standard is that it originates from their field experiences, which makes it difficult to anticipate, and thus, to imbed in systems design. Additionally, since ecologists bring their informal knowledge with them, it is hard to estimate how far it can travel. In other words, it is difficult to predict how widely applicable an ecologist's

informal knowledge, gained through the gathering of particular types of data, is to the reuse of other types of ecological data. However, my results and those of other researchers show clearly that tacit knowledge is a key component of scientific practice, that scientists are aware of its importance, and that although it can be difficult to explicate, it is not impossible to unravel (Berg, 1997; Cambrosio & Keating, 1988; Collins, 2001).

Informal knowledge is not generally viewed as a standard in systems design, at least outside of knowledge management circles, because it can be difficult to communicate within a culture, and it is even more challenging to transfer to those outside the community from which the knowledge was generated. Further, informal knowledge is not seen as "scientific" even among scientists. Informal knowledge is the key for data sharing among ecologists, however, and systems for sharing ecological data among ecologists must include information that helps ecologists to reconstruct the original collection of those data.

It is easy to recognize how standardized research methods, measurements, and disciplinary knowledge span distances within and outside communities. They comprise practices and knowledge that are amenable to written explication, even if, as in the case of research methods, tacit skills and social interaction are necessary in order to replicate another's results (cf., Collins, 1992 [1985]; Collins, 2001). The power of these types of standards to span distance comes from leaving out information. As Latour (1999) observed, in order to transform local knowledge into public knowledge standards involve a loss of information, which he referred to as *reduction*. Reduction allows for much greater standardization and compatibility, and this makes it possible to move from local

to public, or in Latour's vocabulary, to *amplify*. For ecologists, standard methods provide clues about how data were obtained, which aid understanding, but they do not tell a secondary user if the measurements or observations were gathered skillfully. Latour (1999) stated, "To know is not simply to explore, but rather is to be able to make your way back over your own footsteps, following the path you have just marked out" (p. 74). Latour is referring to scientists and their own work, but ecologists who reuse data also want to be able to traverse the journey from reduction to amplification. The ecologists that I interviewed tried to unpack the processes that led to data generation, and if they could not do so, they were unlikely to use the data.

All this is not to say that formal standards are not valuable. The absence of standards and the use of different methods do hamper data sharing, as the literature suggests and as my interviews show. For one, without a means to amplify informal knowledge, each secondary data user is forced to expend the mental energy to "travel back" in order to reconstruct the data; this limits the "scaling up" of ecological data sharing. Another hindrance to the sharing of ecological data is the idiosyncratic methods that ecologists employ to organize their data, which make it difficult to integrate data at a large scale. Greater amplification is impossible without means to simplify the integration of data. However, these complications are difficult, if not impossible at times, to eliminate, and it would not always be desirable to do so. As I discussed previously, there are many considerations that go into methodological choices, and these elements work together to complicate the implementation of common research methods. In addition, at times, standards exist and are followed, and it is other factors that impede data integration. Furthermore, the use of different methods across studies does not always

hamper the integration of data. The knowledge and practices that stem from community membership help ecologists to deal with data reuse challenges, such as a lack of standardized research methods, that others view as significant obstacles. However, since ecologists reuse data that are familiar to them, it is difficult to project how far standards based on informal knowledge can travel. As I discuss in a later section, this is a topic that demands further research.

Standards as Contributors to Distance

Standards, when shared, can help to span distance. When they are not shared, however, they can create space within and among communities. Some of these issues can be dealt with, but others are more intractable. Data sharing systems intended to serve all potential secondary users must successfully span all these distances. Yet, traversing these gaps requires different and multiple approaches.

Ecologists share the domain knowledge that helps them to reuse data, but their approaches and their decisions are also influenced by individual knowledge and personal criteria, such as tolerance for uncertainty. Individual knowledge has the dual capacity to reduce or to increase distance. Trust in the skills or values of another can lessen concerns about data quality, and personal connections can increase access to data and to information about them. In this way, personal insights and relationships span distance. Individual knowledge, once possessed, is difficult to disregard, particularly if it leads to a negative assessment of another individual's data collecting skill or values. Thus, it can also create distances and distrust that are difficult to overcome. One way in which to close the gaps caused by individual knowledge is to make the private public; however,

this is only acceptable if it can be related to a scientific rationale. Outright personal attacks hold little weight; they must at least be shrouded in scientific clothing.

Individual knowledge is often related to first- or second-hand acquaintance with another scientist's tacit skill, and in this way, it is connected to domain knowledge. Scientists recognize the importance of the tacit expertise they acquire from their own data gathering, and so it can be communicated within a publicly acceptable context. For the reasons I discussed previously, though, this is often difficult to do. Scaling up the sharing of ecological data among ecologists is dependent on finding ways to accomplish this, however. Additionally, informal knowledge is not considered "scientific," and therefore, scientists remove it from formal communications, such as journal papers (Collins, 2001; Roth & Bowen, 2001b). One of the most important aspects of peer review is that it places information in a forum from which it can be found, ignored, accepted, and/or disagreed with. Formal data sharing systems serve a similar function because they place data in a public context where they can be reviewed, tested, and discussed. In other words, data benefit from reuse. However, much ecological data has little potential for reuse, and thus, the private and tacit will often remain so unless other means of "public discussion" are found.

Peer review is most often touted as a quality filter. My findings show that its most important role is the forum it provides for discussion; this is a feature that scientists have begun to recognize (Gura, 2002). Berg (1997) stated, "Through explicating that which was implicit, through making public what was private, patterns of practice become open for scrutiny and contestation" (p. 1086). He also noted that such discussions have a spillover effect because they can help to enhance understandings *between* disciplines. In

order to take advantage of informal knowledge for intra- and cross-disciplinary benefits, scientists must consider what aspects of field-based insight would profit from explication. For example, Collins (2001) suggested that replication of experimental results would be easier if journals allowed and encouraged authors to present information regarding the difficulty of an experimental skill or procedure.

My findings also show that the same standards are not equally effective for all communities. For example, data managers and ecologists have variant goals for their work, and they adhere to different standards. Thus, two important and inseparable questions to ask in the design of data sharing systems are: 1) Whose standards? and 2) Who are the users?

Implications for Sharing Data

The results of my study suggest several factors related to mechanisms for sharing data. First of all, metadata standards are created with the intention of assisting all possible users and uses of data, which according to my results, is an impossible task. Data gathered for a particular purpose may be used in multiple ways, but those reuses are limited. In addition, it is difficult to anticipate all potential uses; some secondary use requires documentation beyond that which exists in a metadata standard. Secondly, complex metadata standards may best serve those who are farthest from the data, and there is little incentive, and perhaps limited value, in documenting most data sets at this level. In a field such as ecology, where few intermediaries exist, scientists are expected to take on the burden on documenting their data, yet there are few incentives to do so. The lack of reward is recognized widely, and it was reinforced by the results of my study.

The important role of intermediaries in expanding access to ecological data deserves recognition. Yet, there will always be much more data generated than there are intermediaries to help maintain, store, preserve, package, and disseminate data, and so ecologists must often serve in this role. Ecologists and data managers agree that no one understands the data better than the individual who gathered them, and therefore, it is this person who must document data collection. The best way to help individuals do this is to develop tools that assist them to provide the knowledge they are best able to provide and not to ask them to document their data for all conceivable users and uses.

There are several important factors to consider as ecology looks to increase access to data beyond the bounded sources that are currently available. In the previous chapter, I noted that almost half the ecologists I interviewed relied on the published literature as their main source to locate and obtain data for reuse. Publications served primarily as a way for ecologists to place bounds around their collection of data and to locate "public" data. These are not the primary uses for which publications are intended or for which journals are used and read by scientists (King & Tenopir, 2001). However, this use has implications if information that appeared traditionally in journal papers is relegated to electronic archives or is dropped altogether. For one, we do not know whether the information needed to understand and judge data is provided in most electronic archives because we know little about their use. Additionally, it is important to consider the costs in efficiency if certain information and data traditionally reported in publications, especially journal papers, is moved to electronic archives. Finally, it is also important to consider the ability of readers to judge research reported in publications if information is dropped from them or shifted to other locations. As I noted previously, peer review

places research results in a public forum, and if data and information are removed from this arena it will affect scientists' ability to scrutinize, argue, ignore, or reach consensus on each others' findings.

Another factor to consider in increasing the availability of ecological data is to contemplate the relevance of the distinction between published and unpublished data. My results suggest that it may not be as important as we think. Ecologists recognize that peer review does not guarantee data quality. Publications are used more as a way to locate data and to bound their collection than as a means to ensure quality. Once found, these "public" data have another advantage in that they come with information about them, even if it is not always sufficient for their reuse. Ecologists will use unpublished data if they can find a means to frame their data collection, and if they can obtain information to understand them. When small amounts of data are gathered from multiple sources, concerns about the quality of individual data points are lessened; this also diminishes the distinction between published and unpublished data.

One way to analyze data sharing systems is to view them as formal or informal. In terms of data, *formal* implies public resources that are staffed and funded and that consider issues, such as access, quality, long-term storage and preservation, and user needs. Additionally, a formal system often contains data that are processed, standardized, and quality controlled. On the other hand, *informal* modes of data sharing are private and are characterized by personal interaction. As one of the data managers told me, informal methods are precise, but they are low on recall and inefficient. The challenge in scaling up informal systems is to find ways to increase recall and efficiency without trading off too much in precision.

Application of Research Findings

The results of my study are applicable to existing problems, and they suggest recommendations for data documentation requirements, scientific administration and policy making, education, and publication norms and practices.

As I noted previously, all interviewees concurred that there is great benefit to be realized from the reuse of ecological data. If this is the case, then two goals emerge: 1) to increase the amount of ecological data available for sharing, and 2) to scale up the infrastructure for sharing ecological data. These two objectives are not mutually exclusive, although the first will primarily benefit members of a community of practice, while the second has the potential to meet the needs of users both within and outside the ecological community. Increasing the amount of ecological data available requires mechanisms and policies that make it easier for ecologists to share their data. Tools are needed that make it simpler for ecologists to document their data and that focus on gathering the information that a data collector is best suited to provide. Scientists should not be asked to document their data for all conceivable uses and users. Policies that mandate or encourage data sharing must also address issues of control and rewards. Since ecologists' concerns in regard to data management are short-term and informal, it is challenging to find incentives that outweigh this orientation. This situation makes it more critical that policies are realistic in terms of the information that they require ecologists to provide about their data. Lastly, as Markus (2001) observed, it is difficult for knowledge creators to imagine "distant" other users of their knowledge. For ecologists, this leads to concerns about misuse, which are legitimate when viewed in light of Markus's theory of

knowledge reuse. Van House noted that fears about misuse stem from data owners' concerns about a secondary user's lack of competence and the purposes for which they intend to use the data. She stated, "Sharing information requires that users and providers trust one another" (Van House, 2002, p. 100). Van House referred primarily to information exchange between domains, but trust is also an issue among members of a community of practice. Requirements for scientists to share data must take these concerns into account.

Scaling up the infrastructure for sharing data depends on finding more efficient ways to make data comparable, so they can be more easily integrated, and on providing the decontextualized information that non-ecologists may need to reuse ecological data. Meeting these goals depends largely on intermediaries to prepare data for reuse by eliciting, organizing, storing, sanitizing, and/or packaging data, and by performing various roles in dissemination and facilitation. Intermediaries are an important infrastructure component of knowledge sharing that ecology currently lacks on a wide scale.

Increasing the amount of ecological data available and scaling up the infrastructure for ecological data sharing also depend on education. Most ecologists manage their own data, and they do it with little or no formal training. There are two aspects of data management for ecologists that serve currently as major impediments to data sharing: 1) poor data management skills, and 2) diverse ways of organizing and storing data. The first problem is easier to address than the second. The ability of scientists to manage data can be improved quickly if educators and mentors spend time discussing and teaching these skills. Instead of emphasizing replication, it might be more

valuable for educators to train scientists to describe their methods so that others who use the data can reconstruct the data collection process. This is a subtle, but important distinction. 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 acknowledge and to include field-based insights in their documentation of data, if not in formal communications, would go a long way to improve the longevity of their data. Education in data management will also help to decrease the distance between the standards of scientists and intermediaries, such as data managers; this may lead to solutions that address diversity in data organization and storage.

Almost half of the ecologists I interviewed relied on the literature as their primary source for locating data to reuse and others used it secondarily. As I noted previously, this is not the main use for which publications are intended. My results indicate, however, that editors, publishers, and scientists themselves should consider this use as they make policies or change the information that is presented in publications. Additionally, with the exception of three less experienced ecologists, interviewees did not rely on peer review as a means to judge the quality of data. The literature is, however, an important source of data because it is public and bounded, and it generally serves as a good source of metadata. Since journals are already a familiar and bounded source of data, they may have a key role to play in creating new sources. My interviews show that current methods of bounding are based on geography, the literature, existing databases, and personal knowledge. My findings also point to the value of building formal databases based on important scientific questions.

Limitations of the Study

The use of qualitative research methods for data collection and analysis revealed insights into data sharing and reuse that were not likely to emerge through quantitative research methods. There are, however, several limitations to my study. The main limitations of my investigation are methodological and include a lack of generalizability and a potential to downplay the challenges of data sharing and to underemphasize the importance of formal standards to data sharing. Limitations also stem from the scarcity of theory regarding knowledge reuse.

First of all, the number of participants in my study was small. This is typical of qualitative methodologies, but it limits the generalizability of my findings and makes it difficult to assess the extent to which my respondents are representative of other ecologists. A related limitation concerns my purposive selection of interviewees. My study focused on "successful" data reuse experiences, and therefore, it may exaggerate ecologists' abilities to overcome the challenges described in the literature. Additionally, the ecologists I interviewed were biased in favor of data sharing, and thus they may represent conservative views regarding the misuse of data, and they may reflect overly optimistic perceptions of the benefits of data sharing. For these reasons, it is difficult to estimate the extent to which my interviewees are representative of other like-minded ecologists or of ecologists in general. However, these limitations enabled me to examine some of the assumptions made in the literature about the difficulty of sharing ecological data and to examine how secondary users overcome them. The ability of the ecologists I interviewed to integrate data gathered in different ways and at varying times and scales

may also have a tendency to downplay the importance of formal standards to the sharing of ecological data. My interviews with data managers helped to balance this view.

The focus of my study was on ecological researchers as secondary data users. This approach was beneficial, but it necessarily reduced or eliminated the emphasis on other key data-sharing participants, especially the providers of data. I attempted to address this limitation by asking the ecologists I interviewed about their own experiences in sharing data and by querying data managers about insights they gained in working with data owners and in assisting data users. Answers to these questions revealed the role of social issues, such as ownership and rewards, which other authors note as significant obstacles to data sharing.

In addition to being researchers, my respondents were field ecologists. Even those who performed laboratory analyses possessed substantial field 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 ecologists without substantial field experience might be different. The field of science studies has demonstrated that tacit knowledge and social exchange are important components of many fields, including "harder" sciences such as physics and molecular biology. So, while the data reuse experiences of field, experimental, and theoretical scientists may differ, evidence also suggests strongly that informal knowledge plays an important role in many, if not all, domains (e.g., Collins 1992[1985]; Knorr Cetina, 1999).

My study investigated the reuse of existing data to address questions at a larger spatial or temporal scale. I did not examine the reanalysis of data in order to repudiate or

replicate findings, for educational purposes, or for many other reasons applicable to secondary data use (Fienberg et al., 1985). The potential reuses of scientific data, if viewed through a prism, would refract into multiple possibilities, and we have much to learn about their overlaps and their distinctness. This study looked at data sharing within one community of practice, and it analyzed reuse of ecological data among ecological researchers with fieldwork experience. Thus, my findings are limited in terms of revealing if the same processes would work for other types of ecologists, for ecologists who use data outside their community, and for non-ecologists who reuse ecological data. In fact, we should expect that different uses by diverse users would vary. However, as I noted above, in spite of various nuances unique to different fields, we should also anticipate that members of a community of practice rely on informal knowledge to replicate experiments and to share data and information. This commonality among different communities provides a hook that might be leveraged for sharing across fields.

Finally, limitations stem from the paucity of theory regarding the reuse of data or knowledge. As Shapin (1995) summarized, studies of the sociology of scientific knowledge have been "concerned to show in concrete detail the ways in which the making, maintaining, and modification of scientific knowledge is a local and mundane affair" (p. 304). Recently, Karin Knorr Cetina (1999) shifted the focus from the creation of knowledge to the "construction of the machineries of knowledge construction" (p. 3). She employed the phrase *epistemic cultures* to refer to "cultures that create and warrant knowledge," which she viewed as differing among the sciences (p. 1). In a study of American and Italian physics laboratories, Harry Collins (1998) observed that the meaning of data depended on what he called, "evidential culture" and on institutional

setting. The open evidential culture of the Italian laboratory, characterized by high statistical risk and low interpretative risk, contrasted with the closed evidential culture of the American group. The physicists' different cultures affected the demarcation of "interesting data from uninteresting noise" (Collins, 1998, p. 335). The research of Collins and Knorr Cetina, along with Markus's (2001) exploration of a theory of knowledge reuse, begin to move toward comparisons and revelations that might help to illuminate the various factors at work in reuse and how they relate to institutional forces, scientific traditions, and knowledge creation. At this time, however, theory on these topics is limited.

Future Research

Several topics emerge as rich and important subjects for further investigation based on the findings of this study. Together, these areas point to an agenda for future research focused on the following areas: 1) further investigation into the factors that enable data understanding within a community of practice and the limits of that knowledge; 2) examination of the requirements of other users of ecological data; and 3) comparative studies.

Building on a Community of Practice

Research is needed in order to learn more about the key factors that positively influence data sharing and reuse within a community. Results from my study show that ecologists depend on information that enables them to reconstruct the original collection of the data they reuse. My definition of reconstruction encompasses the stages involved

in finding, obtaining, comprehending, and judging data. While this is a useful finding, we require a more detailed understanding of the information and mechanisms that are at work. Attacks on several interconnected fronts are necessary in order to examine the roles and the characteristics of the factors that influence data sharing and reuse. Findings from my study point to several elements beyond the cultural and social factors discussed in the literature; these include: 1) the nature of the data, such as their complexity (e.g. variability in nature, existence and stability of standard methods) and ease to collect; 2) available documentation about the data, particularly information about data collection methods; 3) the purpose for which ecological data are collected; and 4) the familiarity with particular data that the secondary user brings to their reuse. The first topic, the nature of data, is ripe for comparative study, and I will discuss research needs on this subject a bit further on. Below, I discuss the second, third, and fourth areas, which in many respects, are related.

First of all, close investigation is needed into the written information that enables ecologists to understand data and to form the quality assessments that stem from community knowledge, particularly informal knowledge gained in the field. Informal knowledge is often removed from formal representations of ecologists' work. Yet, ecologists who use the literature as a data source are often able to reuse data based on the information provided. At other times, the information provided is insufficient. We need to better comprehend how and when written documentation provides the vital information necessary to understand and assess data. This is a particularly important, but thorny area for future research because documentation must convey concrete information about how data were collected and analyzed, but in many cases, it must also trigger the

informal knowledge that secondary users bring with them and that helps them to assess another's tacit skills. One way to reach this goal is to work with ecologists and other scientists to help them think more consciously about the information that provides them with clues to data quality.

Secondly, all aspects of data sharing and reuse are complicated by the diversity of purposes for which ecological data are collected. This variety makes it difficult to locate data relevant to a secondary user's needs. In addition, the purpose for which data are collected influences the data that are reported and the information made available about them. Depending on the purpose for reuse, the reported data and metadata may be inadequate for a secondary user, even when they meet community standards for reporting. Finally, the multiple reasons for which ecological data are gathered leads to semantic diversity, which complicates the location of data amenable to integration. Bowker (2000a) discussed this challenge at length in regard to data for biodversity studies.

Thirdly, an ecological researcher's informal knowledge plays the most important role in data reuse. Above, I discussed the need for research into the documentation cues that set this knowledge in motion. Another challenge for future research is to better distinguish quality assessments based on individual knowledge from those related to informal knowledge. Informal knowledge and individual knowledge can look similar, and therefore the line between them is sometimes fuzzy. For example, it is difficult to discern the knowledge at play when ecologists state that they "know of others and their work" and that they rely on this knowledge as a means to lessen concerns about data quality. Personal insights are hard to imbed in data sharing systems; informal knowledge

is also difficult to capture, but it holds greater promise for amplification. We require a better understanding of the distinction in order to identify knowledge that can and cannot be imbedded in data sharing systems. Finally, a further topic of study is to examine how far ecologists' formal and informal knowledge extends when using data they did not collect themselves. Since the ecologists I interviewed reused data with which they are familiar, it is difficult to assess how far their expertise can travel.

Boundary objects, such as standard methods, forms, theories, and symbols, have been examined and discussed as means of translation between communities, but it is possible they could fill a similar function within a community. It would be useful to investigate entities that serve as boundary objects within a community of practice, and to compare and contrast them with those that provide a bridge between communities. Research in this area could help identify boundary objects that work in both directions--to connect people within a particular group and to cross lines between communities.

Needs of Non-Ecologists

A second area that calls for future research is investigation into the needs of secondary users of ecological data who are outside the field of ecology. My interviewees concurred that there is great potential for the reuse of ecological data, but it remains to be learned for what type(s) of uses and users the data are most valuable. Data used secondarily by other types of users for different purposes would most certainly vary from those of ecological researchers who reuse data to address new questions. If Markus (2001) is correct, the further data get from the community that generated them the more important and necessary become the work of intermediaries. However, intermediary

work is expensive, and intermediaries are rare, so investments in this area must be guided by knowledge of the range and needs of other secondary users of ecological data. Metadata standards for ecological data should be evaluated against the needs of data users both within and outside the field of ecology.

Comparative Studies

Finally, comparative investigations are needed in two particular areas. First, there is a need to analyze, describe, and contrast scientific data and communities in terms of the similarities and uniqueness of issues related to the sharing of data from different scientific disciplines. Writers have examined the characteristics of data from different fields in terms of size, complexity, and infrastructure for sharing. Authors have also analyzed cultural and social aspects that hamper or encourage data sharing. Despite the fact that much of this writing is prescriptive, it has made important contributions. What needs to occur next, however, is to investigate more closely why sharing and reuse appears to succeed in some disciplines and not in others. My study has shed some light on this question, which could serve as a starting point for further research. For one, the results from my study indicate that the perceived ease of sharing data in fields such as genetics and climatology may be deceptive. The success of sharing data in other fields may be due to aspects beyond the characteristics of their data. Comments from several of my interviewees, along with other authors, suggest that data sharing is enhanced when data are reused. In this case, "reuse" refers to formal programs for data analysis and dissemination and for the provision of customer service. It also refers to "discussions" about the quality and reuse potential of particular data that appear in the literature, in

metadata, and in communications from reusers. These discussions make public valuable information about the data, such as problems in data collection in a particular year. The burden of assessing individual data points can be lifted in favor of reducing the effort to one of judging the reliability of one or more databases.

A particularly interesting contrast with ecology would be climate science, another important field of environmental study. Today, climate scientists rely heavily on complex computer models and large global data sets to study the atmosphere. Edwards (2001) noted that the "huge size of global data sets makes it impossible to process or understand them in any detail without the aid of computers" (p. 35). However, prior to the introduction of the computer into climate science, the field depended, like ecology, on manual records tied to particular locations. Although climate scientists are divided between those who prioritize theory and those who give precedence to observation, as a whole, climate scientists have had to relinquish the kind of close contact with data that ecologists still demand (Edwards, 1999). A better understanding of the processes that enabled climate scientists to move from a local to global paradigm could help provide clues to the cultural, scientific, and social adjustments necessary to increase the scale of data sharing in ecology.

Second, just as ecology can benefit by learning from other fields, so, too, can other disciplines learn from ecology. Many fields that are characterized by formal infrastructures for data sharing also include data with characteristics similar to those found in ecology. For example, oceanography consists of large, shared databases and small, single-investigator data sets (Hesse et al., 1993). Climate scientists, too, continue to work with smaller data sets in order to study the relationship between small- and large-

scale processes (Edwards, 1999). Further research on ecological data and similar data from other fields would, therefore, benefit many sciences.

Conclusion

Fieldwork serves a central function in the development of ecologists' formal and informal knowledge, which carries over to their reuse of data. As ecology attempts to scale up its data-sharing infrastructure, this core is pulled in different directions. The secondary use of data on a larger scale naturally alters the reliance on informal knowledge, which must now make room for the trappings of the "formal." The formal includes an emphasis on standardization, peer review, and quality control, and its strengths emanate from the forum it provides for discussion, the opportunity it offers the individual to capitalize on collective wisdom, and the presence of intermediaries who bear much of the labor of cleaning, describing, storing, packaging, disseminating, and preserving data. These benefits can reduce the mental and physical energy that scientists must expend to reconstruct, integrate, and judge data. However, a formal system offers only some of the information that scientists require in order to reuse data, and there is a danger in thinking that informal knowledge is easily replaced and is no longer necessary or important. The secondary use of shared data is hard work. Data are political, mutable, slippery, interpretable in multiple ways, and immersed in context, and thus, all types of knowledge are needed to reuse them. Ecology teaches us that there are multiple sides to issues of trust, standards, understanding, and judgments about data quality. To be effective vehicles of data sharing, digital libraries and data repositories must capture all dimensions of knowledge, and we must find ways to document the implicit knowledge

that scientists recognize and can articulate. Much of the informal knowledge that ecologists speak of is amenable to explication and common across fields of study. For example, data limitations, knowledge about the difficulty of collecting particular data, and the variability of certain parameters is common amongst scientific fields and can provide a base for sharing data within and outside a community. We must also recognize and accept that some forms of knowledge and practice are personal or hard to explicate, and they will remain so. It is the prerogative of scientists to make decisions, whether they are based on individual and/or domain knowledge, and to attempt to convince others that their choices meet standards of scientific practice.

Many scientific fields are distinguished by more than one form of data sharing--the informal coexists with the formal, and this situation will continue. The findings from my study make visible knowledge and practices that have implications for both realms and that are applicable beyond the field of ecology.

APPENDIX A

Ecologist Recruitment Letter

Date

Ecologist name and address

Dear Dr. X:

I am a doctoral candidate in the University of Michigan's School of Information. The goal of my dissertation research is to investigate the secondary use of data by ecologists. In my study, s*econdary use* is defined as the use of data collected previously for one purpose to study a new and different problem. Through interviews, I hope to learn about the experiences of ecologists who located, accessed, and used existing data to investigate a new research question. The ultimate aim of my study is to help evaluate and improve mechanisms for sharing data and to inform data sharing policies. The purpose of this letter is to seek your participation in my research.

I am writing to you because of your article entitled, "XXXXX," which was recently published in Ecology/Ecological Applications. In this paper, you and your co-authors acknowledged the use of data from numerous sources to assess changes in ecological processes. As we know, research that attempts to understand ecological processes at various scales and across studies has become an important area of study, and it frequently depends on access to multiple data sets.

If you agree to participate in my study, I will schedule an interview with you, which I anticipate will last 1-2 hours. The interview will be held at a time that is convenient for you; and with your permission, it will be audiotaped. All information you provide will remain confidential. I will ask you questions about your experiences in locating, accessing, and using data. Since often more than one person carries out tasks in a research study, it is possible that there are individuals who worked with you, such as co-authors, graduate students, or research assistants who may be able to provide additional details about certain aspects of the project. In that case, I would contact these individuals and seek permission to talk with them.

Thank you for considering my request. I will contact you soon to talk further and, I hope, gain your participation in my study. In the meantime, please feel free to contact me if you have any questions.

Sincerely,

Ann Zimmerman
asz@umich.edu
Phone: 734-214-7210

APPENDIX B

Data Manager Recruitment Letter

Date

Data manager's name and address

Dear Mr/Ms:

I am a doctoral candidate in the University of Michigan's School of Information. The goal of my dissertation research is to investigate the sharing and reuse of ecological data. The ultimate aim of my study is to help evaluate and improve mechanisms for sharing data and to inform data sharing policies. To date, most of my research has consisted of interviews with ecologists who have located, accessed, and used existing data to investigate a new research question. I have talked with ecologists about how they found data, what they needed to know to understand and use them, and the challenges they confronted throughout the process.

I am writing to you because data managers have a unique and valuable perspective to offer in terms of what is needed to successfully document, organize, store, and disseminate data, and I hope you might be willing share your experiences in these areas with me.

If you agree to participate in my study, I will schedule an interview with you, which I anticipate will last 1-2 hours. The interview will be held at a time that is convenient for you; and with your permission, it will be audiotaped. All information you provide will remain confidential. I will ask you questions about your role at the XXXX; the types of data you manage and the challenges they present in terms of documentation, storage, and dissemination; your experiences in working with data collectors and data users; and your thoughts about data sharing.

Thank you for considering my request. I will contact you soon to talk further and, I hope, gain your participation in my study. In the meantime, please feel free to contact me if you have any questions.

Sincerely,

Ann Zimmerman
asz@umich.edu
Phone: 734-214-7210

APPENDIX C

Consent Form

CONSENT FORM

School of Information
University of Michigan

Use of Shared Scientific Data: Experiences of Ecologists

1. *What is the purpose of the study?* The purpose of this research project is to investigate the secondary use of data by ecologists. In this study, s*econdary use* is defined as the use of data collected for one purpose to study a new problem. I hope to learn how ecologists locate, access, and use existing data to investigate new research questions. The ultimate aim of the study is to inform data sharing policies and to help evaluate and improve mechanisms for sharing data.

2. *What will be involved in participating?* I will schedule 1-2 interviews with you, depending on what seems most useful. Interviews will be conducted face-to-face or over the telephone. Telephone interviews will last approximately one hour; face-to-face interviews will last 1-2 hours. I will tape the interviews and make transcriptions from the tape.

Please sign below if you are willing to have the interview(s) recorded on audiotape. You may still participate in this study if you are not willing to have the interview recorded.

I am willing to have the interview(s) recorded on audiotape.

Signed: _____ Date: _____

3. *Who will know what I say?* All information collected will remain confidential except as may be required by federal, state, or local law. I am a doctoral candidate in the School of Information at the University Michigan (UM). My committee is comprised of: Margaret Hedstrom (Chair), Paul N. Edwards, and Jeffrey K. MacKie Mason, professors in the UM School of Information, and Brian Athey a faculty member in the Cell and Development Biology Department at the UM.

4. *What risks and benefits are associated with my participation?* I do not foresee any risks to you other than a possible breach of confidentiality. To protect against that risk, tapes and transcripts will be kept in a secure location. Access to data will be limited to me and to members of my doctoral committee. Your name will not appear in the transcripts. In any publication or public statement based on the study, all names or other potentially identifying information will be omitted or changed. Two years after the end of the study,

the tapes will be destroyed. Data, including transcripts, will be destroyed 3 years after the completion of the study.

Sometimes people find participating in an interview to be beneficial insofar as it gives them a chance to talk about things that matter to them and to contribute to research directed toward this interest.

5. *What are my rights as a respondent?* Your participation is voluntary. You may ask questions, both before agreeing to be involved and during the course of the study, and they will be answered fully. You may refuse to participate before the study begins, discontinue at any time, or skip any questions that make you feel uncomfortable.

6. *What will result from the study?* I will make the findings known through my doctoral dissertation and professional presentations. I anticipate publishing a portion of my dissertation as article-length reports that will appear in scholarly journals or as book chapters.

7. *If I want more information, whom may I contact about the study?* This study has been approved by the UM Institutional Review Board (IRB) Behavioral Sciences Committee. Questions about the study's approval or your rights as a research subject may be directed to Kate M. Keever, IRB Administrator, 1040 Fleming Administration Building, Ann Arbor, MI 48109-1340; phone (734) 936-0933; fax (734) 647-9084; email: keever@umich.edu. Ann Zimmerman, doctoral candidate and the study's investigator, can be reached at (734) 214-7210; email: asz@umich.edu.

_____ _____

Ann Zimmerman, Investigator Respondent,
Date

APPENDIX D

Ecologist Interview Guide

Thank you for your willingness to talk with me about your data reuse experience(s). As I said in my letter, I'd like to talk with you about your recent Ecology/Ecological Applications paper in terms of how you located the data, the challenges you confronted, and the information you needed to use them. If you've had other similar experiences, I'd be interested in discussing those as well. I will also ask you about your experiences in sharing your own data and your general feelings about the topic.

Do you have any questions before we start? Please feel free to ask questions at anytime.

Background

Can you give me some background on this work? How it came about? How you got interested in this topic?

When you started the project, did you know you'd need to collect other data?

Have you made secondary use of data in other work?

Locating data

I'd like to start by asking you about the process of locating the data you used in your Ecology/Ecological Applications paper.

Did you have a feeling for what data you might find before you started looking? How much of this work were you already aware of? Knowing people that were doing it?

Could you walk me through the steps in locating the data?

Could you describe the process of extracting the data you needed from the articles? How easy was it to find what you needed within a particular paper?

Did you know what data points you would be looking for when you started? Did you create fields in the database ahead of time? Did these fields change as you gathered data?

Is the database continuing to be used? If so, when you started, did you think that the database would continue to be useful? Is the database being added to?

Could you talk about the unpublished data that was acknowledged in the article? How did you find out about and obtain this? Did you do this, or did one of the other authors? What form did you receive the data in?

Did you contact anyone who refused to share data?

What was the biggest obstacle/challenge to locating the data you needed? Could you describe a specific example of that?

What could have made locating relevant data easier for you?

Understanding the data

Was the metadata sufficient for you to use the data? What was the metadata?

How did you judge the quality of the data?

What role did standards play in your use of the data?

Did you contact any of the authors directly to get further information about the data? If yes, can you give me an example of the type of assistance you received?

Have you gotten any feedback from the authors of the data sources since you published the paper? If so, can you describe one or two of these?

How was using this data different from using your own? How was it the same?

Do you think it's necessary to have worked in the field in order to understand the data?

Can you describe your greatest challenge in using the data. Can you provide an example of that?

What information that you didn't have might have the data easier for you to use?

Publishing in *Ecological Archives*

Who created the metadata? Whose idea was it? What prompted it?

Have you heard from others who have used or were interested in using the database? If so, can you describe one of those interactions?

Based on your experiences in reusing data, do you think the metadata required by *Ecological Archives* is the type of information needed for data reuse?

Sharing Data

Have you shared your own data in the past?

If so, what kind of help did you provide to these users? Directly or indirectly?

Is data sharing a topic that was addressed when you were in graduate school?

Quantitative Questions

Approximately how many years have you been an ecologist?

On a scale of 1 to 10, with 1 being very easy and 10 being very difficult, how would you rate your experience in locating data to use for your work?

Working with data that I collect, is: easier, the same, or more difficult than working with the data I gathered from other experimental studies?

Finding data is -- easier, similar, or more difficult -- than finding published literature related to a topic that I am conducting research on.

Final question

Are there are any questions I should have asked, but didn't? Any comments you'd like to make?

Next steps

If I think of additional questions as I review our interview, would it be ok if I contacted you for additional information?

Would you be willing to look at and comment on a brief summary of my interpretation of our conversation?

APPENDIX E

Data Manager Interview Guide

Thank you for your willingness to talk with me. As I said in my letter, I'd like to talk with you about your role at the XXXX; the types of data you manage and the challenges they present in terms of documentation, storage, and dissemination; your experiences in working with data collectors and data users; and your thoughts about data sharing.

Do you have any questions before we start? Please feel free to ask questions at anytime.

Background

Could you tell me a little bit about your background and your role?

What types of data have you worked with?

Do you think a data manager needs to understand the discipline he/she is serving in order to manage those data?

Data

Do you see differences in terms of what's required in order to manage different types of data? For example, are some data more difficult to document than others?

Do you think metadata standards, such as the FGDC standard, reflect what secondary users need to know in order to reuse data? Do you think data can be documented adequately so they can reused in many ways?

What kinds of uses do you see others make of your organization's data?

In what forms do scientists want to receive data? Raw? Summarized?

Do you think there is a large potential for misuse of ecological data? Have you seen examples of this?

Quality

What data quality issues are you concerned about?

Final question

Are there are any questions I should have asked, but didn't? Any comments you'd like to make?

Next steps

If I think of additional questions as I review our interview, would it be ok if I contacted you for additional information?

Would you be willing to look at and comment on a brief summary of my interpretation of our conversation?

APPENDIX F

Transcription Guidelines

- Each tape contains a label with the Respondent (ex. 004) and the date of the interview. Please include this information at the top of the transcript.
- Font: Times New Roman, size 12
- Double-space text
- Number each page (bottom center)
- Number each line (start line numbering over with each page)
- Precede Respondent comments with: **R:**
- Precede Interviewer comments with: **I:**
- You do not need to transcribe all the "ums" and "ahs." However, you also do not need to correct grammatical errors, etc. in people's speech. For example, if the Respondent did not finish one sentence before starting the next, you can use ellipses (ex. …) or 2 hyphens (ex. --) to indicate where they left off before starting a new sentence. See my transcript sample for additional examples.
- You will notice in my transcript sample that if the Respondent or I audibly laughed or sighed, I noted that. This helps me to remember the tone of the different parts of the conversation. If something was inaudible, I also made a note of that; it reminds that something was said that the tape did not pick up.

APPENDIX G

Codebook for Ecologist Interviews

**The interview guide included direct questions on these topics that were asked of all ecologists (see Appendix D). One aspect of the use of these codes is to identify and index the answers to these questions.

Data: Descriptions of the data that ecologists reused, including the characteristics of the data, such as type of data (ex. analytical, experimental, or observational, size of the original data set from which the data were drawn and/or the size of the resulting data set created by an ecologist, and references to whether the the data were "easy" or "difficult" to understand and reuse. This code is also used to index portions of each paper published in Ecology or Ecological Applications that describe the data used in the research that is being reported.

****Data - Form:** The physical form in which data were received, such as hand-written tables, electronic spreadsheets. Or, the method by which data were accessed, such as downloaded from a computerized database.

Motivation: Any of the reason(s) ecologists give for reusing data collected by others, including encouragement from other individuals, and to answer scientific questions.

****Finding:** Any of the methods ecologists employed to locate data for reuse. Also use this code to identify instances in which ecologists were provided with data and did not have to seek them out. This code is also used to index portions of each paper published in Ecology or Ecological Applications that describe how the data were collected.

****Finding - Challenges:** Ecologists' answers to my question about the challenges they encountered in locating data for reuse. Also, any references to or statements related to factors that made it difficult to locate data.

****Finding - Improvements:** Ecologists' answers to questions about what would have made it easier for them to locate data for reuse.

****Understanding:** The information that ecologists needed in order to comprehend the data they reused, including references to the sources of this information, descriptions of the specific information required, and factors that limit ecologists' abilities to comprehend data.

****Quality:** References to positive or negative assessments of data and direct statements ecologists made about how they judged the quality of the data they reused.

****Standards:** Answers to my question about the role of standards to an ecologist's reuse of data. In addition, this code includes reference to any kind of standard, i.e. personal, methodological, or data management.

****Data sharing:** General feelings about data sharing, experiences in sharing their own data, or their experiences in trying to get data from others.

Questions: Answers to the "quantitative" questions that I posed at the end of most, but not all, interviews.

People: People who influenced data reuse; people who saw it as important and encouraged or facilitated it.

Contradictions: Things said by the same person that do not appear to agree with each other.

Publishing data: What does it mean for data to be "published"? These are instances in which people refer to this.

Purpose: References to the relationships between the purpose of a research study and affects on the data collected and how they can be reused.

Comparing data: Ecological data compared with data from other fields. For example, statements about other types of data with which ecologists are familiar; thoughts about whether ecological data are more complex than data from other fields; and why and how data differ, especially in terms of reuse.

E-Archives: Five of the thirteen ecologists made supplemental information associated with their paper available through Ecological Archives. In those instances, I asked ecologists several questions about the material they deposited in Ecological Archives, and this code is employed to identify responses to those questions.

APPENDIX H

Codebook for Data Manager Interviews

**The interview guide included direct questions on these topics that were asked of all data managers (see Appendix E). One aspect of the use of these codes is to identify and index the answers to these questions.

****Data:** Descriptions of the types of data that data managers have managed and the similiarities or differences among data of different types. Also use this code to identify references to the forms (i.e. raw, summarized) that various types of users want.

****Documentation:** Answer to question about metadata standards and whether they contain the type of information that secondary users require. Also use this code to index any reference to the documentation of data; to the type of information required to understand data; and to responsibilities for data documentation.

****Quality:** Use this code to index aspects of quality that data managers are concerned with and responsible for and to identify aspects of quality that data managers note that others, especially scientists, are responsible for.

****Standards:** Answers to my question about the role of standards to the secondary use of data. In addition, this code includes reference to any kind of standard, i.e. personal, methodological, or data management.

Uses: The types of uses data managers have observed of their organization's data and information about the users of their organization's data.

BIBLIOGRAPHY

Adams, K. (2002). The Semantic Web: Differentiating between taxonomies and ontologies. Online 26(4), 20-23.

Alder, K. (2002). The measure of all things: The seven-year odyssey and hidden error that transformed the world. New York: The Free Press.

Arksey, H., & Knight, P. (1999). Interviewing for social scientists: An introductory resource with examples. London: Sage Publications.

Averch, H. A. (1985). A strategic analysis of science and technology policy. Baltimore, MD: Johns Hopkins University Press.

Baker, K. S., Benson, B. J., Henshaw, D. L., Blodgett, D., Porter, J. H., & Stafford, S. G. (2000). Evolution of a multisite network information system: The LTER information management paradigm. BioScience 50(11), 963-978.

Baskin, Y. (1997). Center seeks synthesis to make ecology more useful. Science 275(5298), 310-311.

Ben-Ari, E. T. (1998). Ecologists find synergy in Santa Barbara. BioScience 48(12), 992-993.

Bénel, A., Egyed-Zsigmond, E., Prié, Y., Calabretto, S., Mille, A. Iacovella, A., & Pinon, J-M. (2001). Truth in the digital library: From ontological to hermeneutical systems. In P. Constantopoulos & I. T. Sølvberg (Eds.), Research and advanced technology for digital libraries: 5th European conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001: Proceedings (pp. 366-377). Berlin: Springer-Verlag.

Berg, M. (1997). Problems and promises of the protocol. Social Science & Medicine 44(8), 1081-1088.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. Scientific American 284(5), 35-43.

Bertot, J. C., & McClure, C. R. (1996). The Clinton Administration and the National Information Infrastructure (NII). In P. Hernon, C. R. McClure, & H. C. Relyea (Eds.), Federal information policies in the 1990s: Views and perspectives (pp. 19-44). Norwood, NJ: Ablex.

Bikson, T. K., Quint, B. E., & Johnson, L. L. (1984). Scientific and technical information transfer: Issues and options. Santa Monica, CA: Rand Corporation.

Bishop, A. P. (1999). Document structure and digital libraries: How researchers mobilize information in journal articles. Information Processing and Management 35(3), 255-279.

Bishop, A. P., Neumann, L. J., Star, S. L., Merkel, C., Ignacio, E., & Sandusky, R. J. (2000). Digital libraries: Situating use in changing information infrastructure. Journal of the American Society for Information Science 51(4), 394-413.

Bishop, A. P., & Star, S. L. (1996). Social informatics of digital library use and infrastructure. Annual Review of Information Science and Technology 31, 301-401.

Blumenthal, D., Campbell, E. G., Anderson, M. S., Causino, N., & Louis, K. S. (1997). Witholding research results in academic life science: Evidence from a national survey of faculty. JAMA 277(15), 1224-1228.

Borgman, C. L. (1990). Editor's introduction. In C. L. Borgman (Ed.), Scholarly communication and bibliometrics (pp. 10-27). Newbury Park, CA: Sage Publications.

Boruch, R. F. 1985. Definitions, products, distinctions in data sharing. In S. E. Fienberg, M. E. Martin, & M. L. Straf (Eds.), Sharing research data (pp. 89-122). Washington, DC: National Academy Press.

Bowker, G. C. (2000a). Biodiversity datadiversity. Social Studies of Science 30(5), 643-683.

Bowker, G. C. (2000b). Work and information practices in the sciences of biodiversity. In A. El Abbadi, M. L. Brodie, S. Chakravarthy, U. Dayal, N. Kamel, G. Schlageter, & K-Y. Whang (Eds.), Marking the millennium: 26th international conference on very large databases, Cairo, Egypt, 10-14 September 2000 (pp. 693-696). Orlando, FL: Morgan Kaufmann.

Bowler, P. J. (1993). Science and the environment: New agendas for the history of science? In M. Shortland (Ed.), Science and nature: Essays in the history of the environmental sciences (pp. 1-21). British Society for the History of Science.

Bowser, C. J. (1986). Historic data sets: Lessons from the past, lessons for the future. In W. K. Michener (Ed.), Research data management in the ecological sciences (pp. 155-179). Columbia: University of South Carolina Press.

Boyatzis, R. E. (1998). Transforming qualitative information: Thematic analysis and code development. Thousand Oaks, CA: Sage Publications.

Brunt, J. W. (2000). Data management principles, implementation and administration. In W. K. Michener & J. W. Brunt (Eds.), Ecological data: Design, management and processing (pp. 25-47). Oxford: Blackwell Science.

Buckland, M. and Plaunt, C. (1998). Selecting libraries, selecting documents, selecting data. In Proceedings of the International Symposium on Research, Development, and Practice in Digital Libraries (ISDL 97). Retrieved February 1, 2003, from http://www.sims.Berkeley.edu/~buckland/isdl97/isdl97.html

Burnett, K., Ng, K. B., & Park, S. (1999). A comparison of two traditions of metadata development. Journal of the American Society for Information Science 50(13), 1209-1217.

Cambrosio, A., & Keating, P. (1988). "Going monoclonal": Art, science, and magic in the day-to-day use of hybridoma technology. Social Problems 35(3), 244-260.

Cameron, G. (1998). Electronic databases and the scientific record. In I. Butterworth (Ed.), The impact of electronic publishing on the academic community (pp. 154-159). London: Portland Press.

Campbell, E. G., Clarridge, B. R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A., & Blumenthal, D. (2002). Data withholding in academic genetics: Evidence from a national survey. JAMA 287(4), 473-480.

Campbell, E. G., Weissman, J. S., 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.

Carter, G. C. (1980). Numerical database retrieval in the U.S. and abroad. Journal of Chemical Information and Computer Sciences 20(3), 146-52.

Ceci, S. (1988). Scientists' attitudes toward data sharing. Science, Technology, and Human Values 13(1-2), 45-52.

Ceci, S. J., & Walker, E. (1983). Private archives and public needs. American Psychologist 38(4), 414-423.

Chen, C., & Hernon, P. (Eds.). (1984). Numeric databases. Norwood, NJ: Ablex.

Chinn, C. A., & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data. Cognition and Instruction 19(3), 323-393.

Clarke, A. E., & Fujimura, J. H. (1992). What tools? Which jobs? Why right? In A. E. Clarke and J. H. Fujimura (Eds.), The right tools for the job: At work in twentieth-century life sciences (pp. 3-44). Princeton, NJ: Princeton University Press.

Cohn, J. P. (2001). Scientists concerned about proposed data quality guidelines. BioScience 51(11), 922.

Cole, S. (1992). Making science: Between nature and society. Cambridge, MA: Harvard University Press.

Collins, H. M. 1992 [1985]. Changing order: Replication and induction in scientific practice. Chicago, IL: University of Chicago Press.

Collins, H. M. (1998). The meaning of data: Open and closed evidential cultures in the search for gravitational waves. American Journal of Sociology 104(2), 293-338.

Collins, H. M. (2001). Tacit knowledge, trust and the Q of sapphire. Social Studies of Science 31(1), 71-85.

Collins, H. M., & Pinch, T. (1998). The sex life of the whiptail lizard. In The golem: What you should know about science, 2nd ed. (pp. 109-119). Cambridge: 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. Retrieved February 1, 2003, from http://esa.sdsc.edu/FLED/FLED.html

Cooper, H. & Hedges, L. V. (Eds). (1994). The handbook of research synthesis. New York: Russell Sage Foundation.

Cortez, E. M. (1999). Use of metadata vocabularies in data retrieval. Journal of the American Society for Information Science 50(13), 1218-1223.

Crawford, S. Y. (1996). Scientific communication and the growth of big science. In S. Y. Crawford, J. M. Hurd, & A. C. Weller (Eds.), From print to electronic: The transformation of scientific communication (pp. 1-8). Medford, NJ: Information Today.

Creswell, J. W. (1994). Research design: Qualitative & quantitative approaches. Thousand Oaks, CA: Sage Publications.

Dawes, S. S. (1991). A theory of interagency information sharing. Ph.D. Thesis, Rockefeller College of Public Affairs and Policy, State University of New York at Albany.

De Cagna, J. (2001). Tending the garden of knowledge: A look at communities of practice with Etienne Wenger. Information Outlook 5(7), 6-8, 10,12.

Dionne, J. (1984). Why librarians need to know about numeric databases. In C. Chen, & P. Hernon (Eds.), Numeric databases (pp. 237-246). Norwood, NJ: Ablex.

Dodd, S. A. (1979). Bibliographic references for numeric social science date files--Suggested guidelines. Journal of the American Society for Information Science 30(2), 77-82.

Dodd, S. A. (1982). Cataloging machine-readable data files: An interpretive manual. Chicago: American Library Association.

Domaratz, M. (1996). Finding and accessing spatial data in the National Spatial Data Infrastructure. In L. C. Smith & M. Gluck (Eds.), Geographic information systems and libraries: Patrons, maps, and spatial information (pp. 31-40). University of Illinois at Urbana-Champaign: Graduate School of Library and Information Science.

Edwards, D. (2000). Data quality assurance. In W. K. Michener & J. W. Brunt (Eds.), Ecological data: Design, management and processing (pp. 70-91). Oxford: Blackwell Science.

Edwards, J. L., Lane, M. A., & Nielsen, E. S. (2000). Interoperability of biodiversity databases: Biodiversity information on every desktop. Science 289(5488), 2312-2314.

Edwards, P. N. (1999). Global climate science, uncertainty and politics: Data-laden models, model-filtered data. Science as Culture 8(4), 437-472.

Edwards, P. N. (2001). Representing the global atmosphere: Computer models, data, and knowledge about climate change. In C. A. Miller and P. N. Edwards (Eds.), Changing the atmosphere: Expert knowledge and environmental governance (pp. 31-65). Cambridge, MA: MIT Press.

Elliott, C. A. (Ed.). (1983). Understanding progress as process: Documentation of the history of post-war science and technology in the United States: Final Report of the Joint Committee on Archives of Science and Technology. Chicago, IL: Society of American Archivists.

Ercegovac, Z. (1999). Introduction: Special issue: Integrating multiple overlapping metadata standards. Journal of the American Society for Information Science 50(13), 1165-1168.

Executive Office of the President. (1994). Coordinating geographic data acquisition and access: The National Spatial Data Infrastructure (Executive Order 12906). Federal Register 59(71), 17671-17674.

Federal Geographic Data Committee. (1994). Content standards for digital geospatial metadata. Washington, DC: Federal Geographic Data Committee.

Fienberg, S. E., Martin, M. E., & Straf, M. L. (Eds.). (1985). Sharing research data. Washington, DC: National Academy Press.

Finholt, T. A., & Olson, G. M. (1997). From laboratories to collaboratories: A new organizational form for scientific collaboration. Psychological Science 8(1), 28-36.

Fletcher, P. D., & J. C. Bertot. (1999). Introduction: Special Issue: The National Information Infrastructure. Journal of the American Society for Information Science 50(4), 295-298.

Ford, J., & Martinez, D. (Eds). (2000). Invited feature: Traditional ecological knowledge, ecosystem science, and environmental management. Ecological Applications 10(5), 1249-1340.

Frank, S. (1997). Cataloging digital geographic data in the information infrastructure. Encyclopedia of Library and Information Science 59, 20-54.

Fraser, B., & Gluck, M. (1999). Usability of geospatial metadata *or* space-time matters. Bulletin of the American Society for Information Science 25(6), 24-28.

Fricker, E. (2002). Trusting others in the sciences: *A priori* or empirical warrant? Studies in History and Philosophy of Science 33(2), 373-383.

Funtowicz, S. O., & J. R. Ravetz. (1993). The emergence of post-normal science. In R. von Schomberg (Ed.), Science, politics, and morality: Scientific uncertainty and decision making (pp. 85-123). Dordrecht, The Netherlands: Kluwer.

Gasaway, L. (2002). What's happened to copyright? Information Outlook 6(5), 16-17, 19-21.

Garvey, W. D. (1979). Communication: The essence of science: Facilitating the exchange among librarians, scientists, engineers, and students. Oxford: Pergamon Press.

Geda, C. L. (1979). Social science data archives. American Archivist 42(2), 158-166.

Gershon, D. (2000). Laying a firm foundation for interdisciplinary research endeavours. Nature 406(6791), 107-108.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). The new production of knowledge: The dynamics of science and research in contemporary societies. London: Sage Publications.

Glaze, W. H. (1996). A new "home" for ES&T. Environmental Science & Technology 30(9), 373A.

Godin, B. (1998). Writing performative history: The new *New Atlantis*? Social Studies of Science 28(3), 465-483.

Goldberg, A. I. (1997). Vulnerability and disclosure in science: The interplay between agency and structure. Science Communication 19(2), 99-123.

Goodchild, M. F. (1995). Sharing imperfect data. In H. J. Onsrud and G. Rushton (Eds.), Sharing geographic information (pp. 413-425). New Brunswick, NJ: Center for Urban Policy Research.

Gould, C. C., & Pearce, K. (1991). Information needs in the sciences: An assessment. MountainView, CA: Research Libraries Group.

Gray, A. S., & Dodd, S. A. (1984). The roles of libraries and information centers in providing access to numeric databases. In C. Chen, & P. Hernon (Eds.), Numeric databases (pp. 247-262). Norwood, NJ: Ablex.

Griffith, B. C. (1990). Understanding science: Studies of communication and information. In C. L. Borgman (Ed.), Scholarly communication and bibliometrics (pp. 31-45). Newbury Park, CA: Sage Publications.

Gubiotti, R., Pestel, H., & Kovacs, G. (1984). Numeric data information analysis centers at Battelle. In C. Chen, & P. Hernon (Eds.), Numeric databases (pp. 71-104). Norwood, NJ: Ablex.

Gura, T. (2002). Peer review, unmasked. Nature 416(6878), 258-260.

Haas, J. K., Samuels, H. W., & Simmons, B. T. (1985). Appraising the records of modern science and technology: A guide. Cambridge, MA: Massachusetts Institute of Technology.

Haeuber, R., & Ringold, P. (1998). Invited feature: Ecology, the social sciences, and environmental policy. Ecological Applications 8(2), 330-331.

Haila, Y. (1992). Measuring nature: Quantitative data in field biology. In A. E. Clarke and J. H. Fujimura (Eds.), The right tools for the job: At work in twentieth-century life sciences (pp. 233-253). Princeton, NJ: Princeton University Press.

Harvey, F., Kuhn, W., Pundt, H., Bishr, Y., & Riedemann, C. (1999). Semantic interoperability: A central issue for sharing geographic information. The Annals of Regional Science 33(2), 213-232.

Heim, K. M. (1987). Social scientific information needs for numeric data: The evolution of the international data archive infrastructure. Collection Management 9(1), 1-53.

Hernon, P. (Ed). (1995). Special issue: Geographic information systems (GISs) and academic libraries. Journal of Academic Librarianship 21(4), 231-296.

Hernon, P., & McClure, C. R. (1993). Electronic U.S. government information: Policy issues and directions. Annual Review of Information Science and Technology 28, 45-110.

Hesse, B. W., Sproull, L. S., Kiesler, S. B., & Walsh, J. P. (1993). Returns to science: Computer networks in oceanography. Communications of the ACM 36(8), 90-101.

Hilgartner, S. (1995). Biomolecular databases: New communication regimes for biology? Science Communication 17(2), 240-263.

Hilgartner, S. (1997). Access to data and intellectual property: Scientific exchange in genome research. In National Research Council, 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 (pp. 28-39). Washington, DC: National Academy Press.

Hilgartner, S., & Brandt-Rauf, S. I. (1994). Data access, ownership, and control: Toward empirical studies of access practices. Knowledge: Creation, Diffusion, Utilization 15(4), 355-372.

Hill, L. L., Carver, L., Larsgaard, M., Dolin, R., Smith, T. R., Frew, J., & Rae, M-A. (2000). Alexandria Digital Library: User evaluation studies and system design. Journal of the American Society for Information Science 51(3), 246-259.

Hilts, P. J. (1999, July 31). A law opening research data sets off debate. The New York Times, sec. A, p. 1., col. 4.

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain-analysis. Journal of the American Society for Information Science 46(6), 400-425.

Howard, K. (2000). The bioinformatics gold rush. Scientific American 283(1), 58-63.

Hurd, J. M. (1996). Models of scientific communications systems. In S. Y. Crawford, J. M. Hurd., & A. C. Weller (Eds.), From print to electronic: The transformation of scholarly communication (pp. 9-33). Medford, NJ: Information Today.

Hurd, J. M. (2000). The transformation of scientific communication: A model for 2020. Journal of the American Society for Information Science 51(14), 1279-1283.

Hurd, J. M, Weller, A. C., & Crawford, S. Y. (1996). The changing scientific and technical communications system. In S. Y. Crawford, J. M. Hurd., & A. C. Weller (Eds.), From print to electronic: The transformation of scholarly communication (pp. 97-114). Medford, NJ: Information Today.

Information Infrastructure Task Force. (1993). The National Information Infrastructure: Agenda for Action. Washington, DC: Information Infrastructure Task Force.

Ingersoll, R. C., Seastedt, T. R., & Hartman, M. (1997). A model information management system for ecological research. BioScience 47(5), 310-316.

Kaplan, N. R., & Nelson, M. L. (2000). Determining the publication impact of a digital library. Journal of the American Society for Information Science 51(4), 324-339.

Kareiva, P., & Anderson, M. (1988). Spatial aspects of species interactions: The wedding of models and experiments. In A. Hastings (Ed.), Community ecology (pp. 35-50). Berlin: Springer-Verlag.

King, D. W., & Tenopir, C. (1999). Using and reading scholarly literature. Annual Review of Information Science and Technology 34, 423-477.

Klein, J. T. (1996). Interdisciplinary needs: The current context. Library Trends 45(2), 134-154.

Kling, R. (1999). What is social informatics and why does it matter? D-Lib Magazine 5(1). Retrieved February 1, 2003, from http://www.dlib.org:80/dlib/january99/kling/01kling.html

Kling, R., & McKim, G. (1999). Scholarly communication and the continuum of electronic publishing. Journal of the American Society for Information Science 50(10), 890-906.

Kling, R., & McKim, G. (2000). Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. Journal of the American Society for Information Science 51(14), 1306-1320.

Kling, R., Rosenbaum, H., and Hert, C. (1998). Social informatics in information science: An introduction. Journal of the American Society for Information Science 49(12), 1047-1052.

Knorr Cetina, K. (1999). Epistemic cultures: How the sciences make knowledge. Cambridge, MA: Harvard University Press.

Kuhn, T. S. 1970 [1962]. The structure of scientific revolutions, 2nd ed. Chicago: University of Chicago Press.

Kuklick, H., & Kohler, R. E. (1996). Science in the field: Introduction. OSIRIS, Second series 11,1-16.

Kvale, S. (1996). InterViews: An introduction to qualitative research interviewing. Thousand Oaks, CA: Sage Publications.

Kwa, C. (1993). Radiation ecology, systems ecology and the management of the environment. In M. Shortland (Ed.), Science and nature: Essays in the history of the environmental sciences (pp. 213-249). British Society for the History of Science.

Larsgaard, M. L. (1996). Cataloging planetospatial data in digital form: Old wine, new bottles--new wine, old bottles. In L. C. Smith & M. Gluck (Eds.), Geographic information systems and libraries: Patrons, maps, and spatial information (pp.17-30). University of Illinois at Urbana-Champaign: Graduate School of Library and Information Science.

Latour, B. (1999). Circulating reference: Sampling the soil in the Amazon forest. In Pandora's hope: Essays on the reality of science studies (pp. 24-79). Cambridge, MA: Harvard University Press.

Latour, B., & Woolgar, S. (1979). Laboratory life: The social construction of scientific facts. Beverly Hills, CA: Sage Publications.

Lave, J. (1988). Cognition in practice: Minds, mathematics and culture in everyday life. Cambridge: Cambridge University Press.

Levin, S. A. (1992). The problem of pattern and scale in ecology. Ecology 73(6), 1943-1967.

Lindquist, M. G. (1998). Digital library work: Meeting user needs. In I. Butterworth (Ed.), The impact of electronic publishing on the academic community (pp. 105-110). London: Portland Press.

Loewen, C. (1991-92). From human neglect to planetary survival: New approaches to the appraisal of environmental records. Archivaria 33, 87-103.

Long, S. M. (1995). Documenting federal scientific and technical information (STI): A discussion of appraisal criteria and applications for the National Archives and Records Administration. Journal of Government Information 22(4), 311-319.

Lopez, X. (1998). The dissemination of spatial data: A North-American--European comparative study on the impact of government information policy. Greenwich, CT: Ablex.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. Journal of the Washington Academy of Science 16(12), 1199-1235.

Louis, K. S., Jones, L. M., & Campbell, E. G. (2002). Sharing in science. American Scientist 90(4), 304-307.

Lubchenco, J., Olson, A. M., Brubaker, L. B., Carpenter, S. R., Holland, M. M., Hubbell, et al. (1991). The Sustainable Biosphere Initiative: An ecological research agenda. Ecology 72(2), 371-412.

Luedke, J. A. Jr., Kovacs, G. J., & Fried, J. B. (1977). Numeric data bases and systems. Annual Review of Information Science and Technology 12, 119-181.

Lutz, M. (Ed.). (1995). Special issue: Making GIS a part of library service. Information Technology and Libraries 14(2), 77-122.

Lyman, P. (1999). Digital documents and the future of the academic community. In R. Ekman & R. E. Quandt (Eds.), Technology and scholarly communication (pp. 366-379). Berkeley: University of California Press.

Lynch, C. A. (1993). The transformation of scholarly communication and the role of the library in the age of networked information. Serials Librarian 22(3/4), 5-20.

Lynch, M. & Woolgar, S. (1990). Introduction: Sociological orientations to representational practice in science. In M. Lynch & S. Woolgar (Eds.), Representation in scientific practice (pp. 1-18). Cambridge, MA: MIT Press.

Macilwain, C. (2000). NSF puts big money into complex ecology. Nature 407(6806), 823.

Maier, D., Landis, E., Cushing, J., Frondorf, A., Silberschatz, A., Frame, M., & Schnase, J. L. (Eds.). (2000). Research directions in biodiversity and ecosystem informatics. Report of an NSF, USGS, NASA workshop on biodiversity and ecosystem informatics held at NASA Goddard Space Flight Center, June 22-23, 2000.

Mangan, E. U. (1995). The making of a standard. Information Technology and Libraries 14(2), 99-110.

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.

Marshall, E. (2000). Epidemiologists wary of opening up their data. Science 290(5489), 28-29.

Marshall, E. (2002). Clear-cut publication rules prove elusive. Science 295(5560), 1625.

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.

McCain, K. W. (1995). Mandating sharing: Journal policies in the natural sciences. Science Communication 16(4), 403-31.

McClure, C. R. (1989). Increasing access to U.S. scientific and technical information: Policy implications. In C. R. McClure & P. Hernon (Eds.), U.S. scientific and technical information (STI) policies: Views and perspectives (pp. 319-354). Norwood, NJ: Ablex.

McClure, C. R., Moen, W. E., & Bertot, J. C. (1999). Descriptive assessment of information policy initiatives: The Government Information Locator Service (GILS) as an example. Journal of the American Society for Information Science 50(4), 314-330.

McGinley, L. (1999, March 1). Scientists challenge provision opening access to data. The Wall Street Journal, sec. A, p. 24, col. 1.

McGrath, R. E., Futrelle, J., Plante, R., & Guillaume, D. (1999). Digital library technology for locating and accessing scientific data. In E. A. Fox & N. C. Rowe (Eds.), Digital Libraries '99: Proceedings of the fourth ACM conference on digital libraries (pp. 188-194). New York: ACM Press.

McLaughlin, R. L., Carl, L .M., Middel, T., Ross, M., Noakes, D. L. G., Hayes, D. B., & Bayliss, J. R. (2001). Potentials and pitfalls of integrating data from diverse sources: Lessons from a historical database for Great Lakes stream fishes. Fisheries 26(7), 14-23.

Michener, W. K., & Brunt, J. W. (Eds.). (2000). Ecological data: Design, management and processing. Oxford: Blackwell Science.

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. (1997). Nongeospatial metadata for the ecological sciences. Ecological Applications 7(1), 330-342.

Miles, M. B., & Huberman, A. M. (1994). Qualitative data analysis: An expanded sourcebook, 2nd ed. Thousand Oaks, CA: Sage Publications.

Milstead, J., & Feldman, S. (1999a). Metadata: Cataloging by any other name. Online 23(1), 24-31.

Milstead, J., & S. Feldman. (1999b). Metadata projects and standards. Online 23(1), 32-40.

Mish, F. C. (Ed). (1993). Merriam-Webster's Collegiate Dictionary, 10th ed. Springfield, MA: Merriam-Webster.

Moen, W. E. (1996). The Government Information Locator Service: Discovering, identifying, and accessing spatial data. In L. C. Smith & M. Gluck (Eds.), Geographic information systems and libraries: Patrons, maps, and spatial information (pp.41-67). University of Illinois at Urbana-Champaign: Graduate School of Library and Information Science.

Murdock, J. W. (1980). Numerical data indexing. Journal of Chemical Information and Computer Sciences 20(3), 132-36.

Murphy, E. (1990). Agencies, journals set some rules. Science 248(4958), 954.

National Academy of Public Administration. (1992). The archives of the future: Archival strategies for the treatment of electronic databases: A study of major automated databases maintained by agencies of the U.S. Government. Washington, DC: National Academy of Public Administration.

National Research Council. (1995a). Finding the forest in the trees: The challenge of combining diverse environmental data: Selected case studies. Washington, DC: National Academy Press.

National Research Council. (1995b). Preserving scientific data on our physical universe: A new strategy for archiving the nation's scientific information resources. 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. (1999). A question of balance: Private rights and the public interest in scientific and technical databases. Washington, DC: National Academy Press.

National Research Council. (2000). Ecological indicators for the nation. Washington, DC: National Academy Press.

National Research Council. (2002). Access to research data in the 21st century: An ongoing dialogue among interested parties: Report of a workshop. Washington, DC: National Academy Press.

National Science Board, Task Force on the Environment. (2000). Environmental science and engineering for the 21st century: The role of the National Science Foundation. Washington, DC: National Science Foundation.

Nelkin, D. (1984). Science as intellectual property: Who controls research? New York: Macmillan.

Neuhold, E. J. (1998). Access to scientific data repositories: Introduction. In I. Butterworth (Ed.), The impact of electronic publishing on the academic community (pp. 153-154). London: Portland Press.

Ng, K. B., Park, S., & Burnett, K. (1997). Control or management: A comparison of two approaches for establishing metadata schemes in the digital environment. In C. Schwartz & M. Rorvig (Eds.), ASIS '97: Proceedings of the 60th ASIS annual meeting: Digital collections: Implications for users, funders, developers and maintainers (pp. 337-346). Medford, NJ: Information Today.

Nunberg, G. (1993). The places of books in the age of electronic reproduction. Representations 42, 13-37.

Office of Management and Budget. (1996). OMB Circular A-130 (Revised). Transmittal Memorandum No. 3. Washington, DC: Office of Management and Budget. Retrieved February 1, 2003 from, http://www.whitehouse.gov/omb/circulars/a130/a130.html

Office of Management and Budget. (2002). Guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by federal agencies; republication. Federal Register 67(36), 8452-8460.

O'Neill, E. T., Lavoie, B. F., & McClain, P. D. (1998). Web characterization project: An analysis of metadata usage on the Web. Annual Review of OCLC Research. Retrieved February 1, 2003 from, http://www.oclc.org/research/publications/arr/1998/oneill_etal/metadata.htm

Palmer, C. L. (1996). Information work at the boundaries of science: Linking library services to research practices. Library Trends 45(2), 165-191.

Palmer, C. L. (2001). Work at the boundaries of science: Information and the interdisciplinary research process. Boston: Kluwer.

PCAST Panel on Biodiversity and Ecosystems. (1998). Teaming with life: Investing in science to understand and use America's living capital. Washington, DC: President's Committee of Advisors on Science and Technology, Executive Office of the President.

Pierce, S. J. (1990). Disciplinary work and interdisciplinary areas: Sociology and bibliometrics. In C. L. Borgman (Ed.), Scholarly communication and bibliometrics (pp. 46-58). Newbury Park, CA: Sage Publications.

Pierce, S. J. (1999). Boundary crossing in research literatures as a means of interdisciplinary information transfer. Journal of the American Society for Information Science 50(3), 271-279.

Poland, J. A. (1994). Informal communication among scientists and engineers. Encyclopedia of Library and Information Science 53, 171-181.

Porter, J. H. (2000). Scientific databases. In W. K. Michener & J. W. Brunt (Eds.), Ecological data: Design, management and processing (pp. 48-69). Oxford: Blackwell Science.

Porter, J. H., & Callahan, J. T. (1994). Circumventing a dilemma: Historical approaches to data sharing in ecological research. In W. K. Michener, J. W. Brunt, & S. G. Stafford (Eds.), Environmental information management and analysis: Ecosystem to global scales (pp. 193-202). London: Taylor & Francis.

Porter, T. M. (1995). Trust in numbers: The pursuit of objectivity in science and public life. Princeton, NJ: Princeton University Press.

Porter, T. M. (1999). Quantification and the accounting ideal in science. In M. Biagioli (Ed.), The science studies reader (pp. 394-406). New York: Routledge.
(Reprinted from: Social Studies of Science 22 (1992): 633-651)

Powell, R. R. (1999). Recent trends in research: A methodological essay. Library & Information Science Research 21(1), 91-119.

Pundt, H., & Bishr, Y. (2002). Domain ontologies for data sharing--an example from environmental monitoring using field GIS. Computers & Geosciences 28(1), 95-102.

Redfearn, J. (1999). OECD to set up global facility on biodiversity. Science 285(5424), 22-23.

Reichhardt, T. (2002). Concern mounts as US agencies face challenges to data quality. Nature 416(6878), 249-250.

Reichman, J. H., & Uhlir, P. F. (2001). Promoting public good uses of scientific data: A contracturally reconstructed commons for science and innovation. Paper prepared for Conference on the Public Domain, Duke Law School, November 9-11, 2001. Retrieved February 1, 2003 from, http://www.law.duke.edu/pd/papers/ReichmanandUhlir.pdf

Rinderknecht, D. (1991). Nonbibliographic databases: Determining level of service. RQ 30(4), 528-533.

Roberts, M. J., Thomas, S. R., & Dowling, M. J. (1984). Mapping scientific disputes that affect public policymaking. Science, Technology, and Human Values 9(1), 112-122.

Rockwell, R. C. (2001). Policies and politics in social science data archiving. A paper presented to the Seminar on Archives, Documentation, and the Institutions of Social Memory, Advanced Study Center and International Institute, University of Michigan, February 14, 2001.

Roosendaal, H. E., & Geurts, P. A. TH. M. (1999). Scientific communication and its relevance to research policy. Scientometrics 44(3), 507-519.

Roth, W-M., & Bowen, G. M. (1999). Digitizing lizards: The topology of 'vision' in ecological fieldwork. Social Studies of Science 29(5), 719-764.

Roth, W-M., & Bowen, G. M. (2001a). 'Creative solutions' and 'fibbing results': Enculturation in field ecology. Social Studies of Science 31(4), 533-556.

Roth, W-M., & Bowen, G. M. (2001b). Of disciplined minds and disciplined bodies: On becoming an ecologist. Qualitative Sociology 24(4), 459-481.

Rothenberg, J. (1995). Ensuring the longevity of digital documents. Scientific American 272(1) 42-47.

Rubin, H. J., & Rubin, I. S. (1995). Qualitative interviewing: The art of hearing data. Thousand Oaks, CA: Sage Publications.

Schiff, L. R., Van House, N. A., & Butler, M. H. (1997). Understanding complex information environments: A social analysis of watershed planning. In R. B. Allen & E. M. Rasmussen (Eds.), Proceedings of the second ACM international conference on digital libraries: ACM digital libraries '97, Philadelphia, PA, July 23-26, 1997 (pp.161-168). New York: ACM Press. Retrieved February 1, 2003 from, http://sims.berkeley.edu/~vanhouse/waters.html

Schnase, J. L. (2000). Research directions in biodiversity informatics. In A. El Abbadi, M. L. Brodie, S. Chakravarthy, U. Dayal, N. Kamel. G. Schlageter, & K-Y. Whang (Eds.), Marking the millennium: 26th international conference on very

large databases, Cairo, Egypt, 10-14 September 2000 (pp. 697-700). Orlando, FL: Morgan Kaufmann.

Sepic, R., & Kase, K. (2002). The National Biological Information Infrastructure as an e-government tool. Government Information Quarterly 19(4), 407-424.

Service, R. F. (2000). Chemists toy with the preprint future. Science 289(5484), 1445-1446.

Service, R. F. (2002). Physicists question safeguards, ponder their next moves. Science 296(5573), 1584-1585.

Shapin, S. (1995). Here and everywhere: Sociology of scientific knowledge. Annual Review of Sociology 21, 289-321.

Shelby, R. (2000). Accountability and transparency: Public access to federally funded research data. Harvard Journal on Legislation 37(2), 369-389.

Siang, S. (2002). NIH seeks comment on proposed data sharing policy. Journal of the National Cancer Institute 94(8), 555.

Sieber, J. E. (1988). Data sharing: Defining problems and seeking solutions. Law and Human Behavior 12(2), 199-206.

Sieber, J. E. (Ed.). (1991). Sharing social science data: Advantages and challenges. Newbury Park, CA: Sage Publications.

Sieber, J. E., & Trumbo, B. E. (1995). (Not) giving credit where credit is due: Citation of data sets. Science and Engineering Ethics 1(1), 11-20.

Slobodkin, L. B. (1988). Intellectual problems of applied ecology. BioScience 38(5), 337-342.

Smith, J. T. Jr. (1996). Meta-analysis: The librarian as a member of an interdisciplinary research team. Library Trends 45(2): 265-279.

Smith, L. C. & Gluck, M. (Eds.). (1996). Geographic information systems and libraries: Patrons, maps, and spatial information. University of Illinois at Urbana-Champaign: Graduate School of Library and Information Science.

Sprehe, J. T. (1994). Federal information policy in the Clinton administration's first year. Bulletin of the American Society for Information Science 20(4), 20-25.

Sprehe, J. T. (1999). Government information: From inaccessibility to your desktop and back again. Journal of the American Society for Information Science 50(4), 340-345.

Stanley, B., & Stanley, M. (1988). Data sharing: The primary researcher's perspective. Law and Human Behavior 12(2), 173-180.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-1939. Social Studies of Science 19(3), 387-420.

Steele, T. W., & Stier, J. C. (2000). The impact of interdisciplinary research in the environmental sciences: A forestry case study. Journal of the American Society for Information Science 51(5), 476-484.

Sterling, T. D. (1988). Analysis and reanalysis of shared scientific data. Annals of the American Academy of Political and Social Science 495, 49-60.

Sterling, T. D., & Weinkam, J. J. (1990). Sharing scientific data. Communications of the ACM 33(8), 113-119.

Stokes, D. E. (1997). Pasteur's quadrant: Basic science and technological innovation. Washington, DC: Brookings Institution Press.

Suchman, L. A. (1987). Plans and situated actions: The problem of human-machine communication. Cambridge: Cambridge University Press.

Suchman, L. A., & Trigg, R. H. (1993). Artificial intelligence as craftwork. In S. Chaiklin and J. Lave (Eds.), Understanding practice: Perspectives on activity and context (pp. 144-178). New York: Cambridge University Press.

Sugden, A., and Pennisi, E. (2000). Diversity digitized. Science 289(5488), 2305.

Taylor, S. J., & Bogdan, R. (1998). Introduction to qualitative research methods: A guidebook and resource, 3rd ed. New York: John Wiley.

Tenopir, C., & King, D. W. (2001). Lessons for the future of journals. Nature 413(6857), 672-673.

Thiele, H. (1998). The Dublin Core and Warwick Framework: A review of the literature, March 1995-September, 1997. D-Lib Magazine (January). Retrieved February 1, 2003 from, http://www.dlib.org/dlib/january98/01thiele.html

Tilman, D. (1989). Ecological experimentation: Strengths and conceptual problems. In G. E. Likens (Ed.), Long-term studies in ecology: Approaches and alternatives (pp. 136-157). New York: Springer-Verlag.

Trybula, W. (1997). Data mining and knowledge discovery. Annual Review of Information Science and Technology 32, 197-229.

U.S. Department of Commerce. (1992). Spatial Data Transfer Standard. Washington, DC: U.S. Department of Commerce, National Institute of Standards and Technology, Federal Information Processing Standard 173.

U.S. National Commission on Libraries and Information Science. (1984). To preserve the sense of earth from space. Washington, DC: U.S. National Commission on Libraries and Information Science.

Van Alstyne, M., & Brynjolfsson, E. (1996). Could the Internet balkanize science? Science 274(5292), 1479-1480.

Van House, N. A. (2002). Digital libraries and practices of trust: Networked biodiversity information. Social Epistemology 16(1), 99-114.

Van House, N. A. in press. Digital libraries and collaborative knowledge construction. In A. P. Bishop, B. Buttenfield, & N. A. Van House (Eds.), Digital library use: Social practice in design and evaluation. Cambridge, MA: MIT Press.

Van House, N., Butler, M., & Schiff, L. (1995). Needs assessment and evaluation of a digital environmental library: The Berkeley experience. Retrieved February 1, 2003 from, http://info.sims.berkeley.edu/~vanhouse/dl96.html

Van House, N. A., Butler, M. H., & Schiff, L. R. (1998). Cooperative work and practices of trust: Sharing environmental planning data sets. In CSCW '98: Proceedings: ACM conference on computer supported cooperative work, Seattle, WA, November 14-18 (pp. 335-343). New York: Association for Computing Machinery.

Vellucci, S. L. (1998). Metadata. Annual Review of Information Science and Technology 33, 187-222.

Von Schomberg, R. (1993). Introduction. In R. von Schomberg (Ed.), Science, politics and morality: Scientific uncertainty and decision making (pp. 1-4). Dordrecht, The Netherlands: Kluwer.

Walsh, J. P. & Bayma, T. (1996). The virtual college: Computer-mediated communication and scientific work. The Information Society 12(4), 343-363.

Warnow-Bluett, J. & Weart, S. R. (1992). AIP study of multi-institutional collaborations: Phase I: High-energy physics: Report no. 1: Summary of project activities and findings/project recommendations. New York: American Institute of Physics.

Weil, V. (1988). Policy incentives and constraints on scientific and technical information. Science, Technology, & Human Values 13(1 & 2), 17-26.

Weingart, P. (1997). From “finalization" to “mode 2": Old wine in new bottles? Social Science Information 36(4), 591-613.

Weiss, R. S. 1994. Learning from strangers: The art and method of qualitative interview studies. New York: The Free Press.

Weller, A. C. (1996). The human genome project. In S. Y. Crawford, J. M. Hurd., & A. C. Weller (Eds.), From print to electronic: The transformation of scholarly communication (pp. 35-64). Medford, NJ: Information Today.

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

Whillans, T. H., Regier, H. A., & Christie, W. J. (1990). F. E. J. Fry’s field studies: Good field data provoke new questions. Transactions of the American Fisheries Society 119(4), 574-584.

White, H. D. (1982). Citation analysis of data file use. Library Trends 30, 467-477.

Worster, D. (1994). Nature’s economy: A history of ecological ideas, 2nd ed. New York: Cambridge University Press.