

Not by Metadata Alone: The Use of Diverse Forms of Knowledge to Locate Data for Reuse

Ann Zimmerman

*Collaboratory for Research on Electronic Work, School of Information,
University of Michigan, 1075 Beal Avenue, Ann Arbor, MI 48109-2112, USA*

734-764-1865

asz@umich.edu

<http://www.si.umich.edu>

Abstract An important set of challenges for eScience initiatives and digital libraries concern the need to provide scientists with the ability to access data from multiple sources. This paper argues that an analysis of scientists' reuse of data prior to the advent of eScience can illuminate the requirements and design of digital libraries and cyberinfrastructure. As part of a larger study on data sharing and reuse, I investigated the processes by which ecologists locate data that were initially collected by others. Ecological data are unusually complex and present daunting problems of interpretation and analysis that must be considered in the design of cyberinfrastructure. The ecologists that I interviewed found ways to overcome many of these difficulties. One part of my results shows that ecologists use formal and informal knowledge that they have gained through disciplinary training and through their own data-gathering experiences to help them overcome hurdles related to finding, acquiring, and validating data collected by others. A second part of my findings reveals that ecologists rely on formal notions of scientific practice that emphasize objectivity to justify the methods they use to collect data for reuse. I discuss the implications of these findings for digital libraries and eScience initiatives.

Keywords *Data reuse · Data sharing · Ecology*

1 Introduction

The goals of eScience applications and cyberinfrastructure (CI) are extremely ambitious.¹ The hope is that they will support “...faster, better, and different” science and serve individuals, teams, and organizations in ways that will

¹ Clifford Lynch makes a useful distinction between the terms eScience and cyberinfrastructure [25]. He notes that eScience has been used in the United Kingdom to refer to changes in scientific practice that are enabled by computational resources for modeling, simulation, analysis, and data capture; by sophisticated instrumentation; and by advanced networks. In the United States, however, cyberinfrastructure signifies what is needed to support those changes. In recognition of Lynch's definitions, I use the phrase eScience applications as synonymous with cyberinfrastructure.

transform *what they can do, how they do it, and who participates...* [2,21]. Even though these efforts are in their infancy, there are already examples of exciting findings, life-saving systems, and unique learning environments enabled by high performance computing, networking, and information tools and resources [4,17,21].

In contrast to these inspiring successes there are stories of small- and large-scale information infrastructure development and deployment that led nowhere, along with many cases that likely never got reported [36]. Past research on failed projects suggests that incompatibilities between how systems work and how users expect them to work are an important threat to successful adoption and use of CI [38]. In order for users to adopt a new technology, it must offer advantages over current practices, positively change the way work can be performed, and be easy to implement and straightforward to use [9]. If current work practices, user requirements, system usability, and the social aspects of disciplinary communities are ignored, the technology may be of little use, and the huge investments made in eScience applications and CI may be wasted.

eScience initiatives and digital libraries face many of the same challenges when it comes to creating systems and services that support new practices while recognizing traditional ones. An important subset of these challenges concerns the need to organize, maintain, and provide access to scientific data and to support their reuse. A capability of particular interest for both CI and digital libraries going forward is providing users with the ability to access data from multiple sources. The goal is to make it easier for users to search and navigate the widely dispersed universe of scientific data and to bring data together to address new kinds of scientific questions. A number of changes to scientific culture and practice are possible as scientists move from assembling and analyzing their own data to relying more on data collected by others [8,19,22,40]. In order to predict or analyze such change, we must understand how scientists currently locate and use data they did not collect themselves, so systems can be designed that do not inadvertently “break” those practices [35]. This paper argues that an analysis of scientists' reuse of data prior to the advent of eScience can illuminate the requirements and design of digital libraries and CI.

As part of a larger study on data sharing and reuse, I investigated the processes by which ecologists locate data that were initially collected by others. I focused on ecological data because they present significant obstacles to sharing and reuse and little is known about how secondary users overcome these challenges. Elsewhere, I have analyzed the role that standardized methods of data collection have on ecologists' abilities to understand and judge the quality of data that they did not collect themselves [42]. In this paper, I focus on the processes that ecologists use to locate data for reuse. I begin, in Section 2, with an overview of the culture and practice of ecology, especially as they relate to data sharing and reuse. In Section 3, I describe the methods I employed to investigate ecologists' approaches to find data for reuse. My findings, which I present in Section 4, show that the knowledge gained through their own collection of particular types of data, along with familiarity with the literature and with general research trends, provides ecologists with the sense that data exist; helps them to pose research questions that employ those data; guides their search for data; and aids them in developing strategies that increase their chances to obtain data. Further, 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 of scientific practice leads ecologists to seek strategies that *bound* their collection of data, so that the data they gather for reuse can be seen to be drawn from a representative sample. The ecologists I interviewed relied primarily on the published literature, time, and geography to frame their collection of data. Ecologists also choose methods that increase their access to data, that reduce the potential for error, and that provide rationales for their choices that can be defended publicly. In Section 5, I discuss the implications of these findings for digital libraries and eScience applications.

2 Ecological Practice and Data

There is widespread interest across disciplines in using data gathered for one purpose to study new and different problems. Technological advances such as improved techniques and instrumentation for gathering data, policies that encourage data sharing and reuse, and the formation of collaborative projects to tackle problems on larger scales mean that the amount of data available for

secondary use continues to increase. In addition, the sharing and reuse of data are gaining prominence in disciplines where they were previously uncommon. This is the case for fields like ecology where pressing environmental problems have become large in scope and require new methods of inquiry and integration of heterogeneous data.

Ecology is the study of the interrelationships between the earth's organisms and their environment. Ecologists conduct experiments in field and laboratory settings and observe phenomena in the natural world using a wide array of tools and approaches that they learn as part of their enculturation to the discipline [33-34]. Wherever they work, ecologists are confronted with variance among individuals within a population, with the unpredictable character of the natural environment, and with immensely complex systems with large numbers of variables and many subtle interactions. For example, even with a field guide and specimens in hand, it can be difficult to distinguish one tree from another because a guidebook cannot show all the possible variations in leaves, bark, or structure that occur over time. The result is that ecological data are unusually complex, presenting daunting problems of interpretation and analysis that pose particular challenges for secondary use [6-7,27-28].

Ecology is at a unique moment in time in which a potentially major shift is occurring. The complexities of environmental problems, funding opportunities and pressures, and the availability of new technologies have led some ecologists to begin to ask new questions and to look for ways to expand the spatial and temporal scales of their science. For example, the National Center for Ecological Analysis and Synthesis (NCEAS) was established in 1995 to help researchers find and use data collected by others in order to answer larger scientific questions [3]. Through its participation in projects such as the Knowledge Network for Biocomplexity, NCEAS has developed software to help ecologists manage, locate, and integrate ecological data [1]. Another major initiative is the proposed National Ecological Observatory Network (NEON), which is intended to revolutionize the collection of new ecological data through a vast network of sensors and associated technologies [29]. Cyberinfrastructure will be an integral part of environmental observatories like NEON—connecting people to each other, to remote instrument and sensor networks, to data, and to software. NEON and

NCEAS reflect a growing recognition on the part of ecologists, funding agencies, and policy-makers that major scientific advances and collaboration through the sharing of ideas, data, information, and technology will be essential to predict and manage environmental problems with regional, national, or global effects [16,29].

While ecology is gearing up to employ new methods, technologies, and collaborations, most current practice is characterized by relatively modest single-investigator or small-group studies conducted at limited spatial and temporal scales [1,27-28]. Data are generally stored in local file systems, and methods used to document and manage data are largely idiosyncratic; these factors make it difficult to locate and aggregate data from one study to another. Ecology's multiple subdisciplines have varying practices, terminology, and levels of collaboration, and there are few instances of large-scale, shared-instrument platforms.

Furthermore, ecologists do not have an established infrastructure for sharing data, although efforts are underway to change this. In general, sharing takes place between close associates and relies heavily on social interaction [13]. The reasons for this are complex and include the characteristics of ecological data and few incentives for sharing. The practical aspects of conducting research in a particular discipline also limit the questions that can be asked. Like researchers in other field sciences, ecologists conduct their experiments at least partially out of doors in uncontrolled circumstances [24]. Such studies are labor intensive, which constrains the size of the area that can be studied [28]. Since a single ecologist or a small group of researchers typically gathers data, proprietary rights are also an issue as ecologists feel that they own the data they collect and that sharing is not rewarded [28]. Finally, ecological data present particular challenges in terms of documentation for secondary use [6-7,27-28]. Data sets tend to be small and highly diverse, and the methods and techniques used to obtain and manage data vary. This variability makes it hard to describe ecological data adequately for others to use them. 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 sharing and reuse of ecological data.

3 Methods

As the previous section showed, the secondary use of ecological data presents a number of challenges that must be considered in the design of CI. In order to better understand how ecologists currently approach these difficulties, I investigated the experiences of ecologists who used data they did not collect themselves. In this section, I discuss the methods used to address this topic.

My primary method of data collection was semi-structured in-depth interviews with 13 ecologists who reused data to answer a new research question and who published the results of that work in Ecology or Ecological Applications. I made determinations about data reuse by reviewing the methods and acknowledgments sections of papers published from 1999-2001 in these two prominent Ecological Society of America (ESA) journals. 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 for the purposes of my investigation. 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 reuse experience could be remembered. In addition, I selected papers to achieve diversity in the types of data reused, in data sources accessed, and in experience levels of ecologists. I also interviewed four data managers in order to obtain another view of ecological data.

Data for my research were collected through a combination of face-to-face and telephone interviews that took place between June 2001 and February 2002. On average, interviews lasted 90 minutes. Individuals were defined as ecologists if they were members of the ESA, if their institutional affiliation or professional title contained the word *ecology* (or variant of it), or if they identified themselves as ecologists.

The questions I asked ecologists focused on their experiences in locating, accessing, understanding, and judging data and on general attitudes toward data sharing. I conducted one interview with each respondent, and I interviewed one author associated with each journal paper. In all cases except one, the ecologists I interviewed were also the first authors of the published papers. 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.² I developed an initial coding scheme to analyze the interview data, and I refined these codes after applying them to a subset of the interviews. The codes were based on my research questions and on prior literature; a list of the codes I used is available in [41].

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 1 describes the key data reused by the ecologists I interviewed.³ Ecologists accessed and received data in both electronic and print forms. They created their own data sets from the data they collected, typically using Microsoft Excel® or SAS®. While many ecologists gathered similar types of data in small amounts from multiple sources, several ecologists used existing data sets. Data grid technologies that are often mentioned in the context of large-scale eScience are not yet common in ecology, so it is not surprising that they were not used by the ecologists I interviewed to access, manage, or store data [31].

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. 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 [26].

4 Findings

Ecological data are widely dispersed, heterogeneous, and complex, which make them difficult to locate and hard to reuse. These challenges are complicated by

² Except in one case, all the ecologists I interviewed were the first author of the published paper. In the one exception, the interviewee was the third author and the chief person who gathered the data that were reused.

social factors that hinder data sharing, such as issues of ownership and a lack of reward for sharing. The ecologists that I interviewed found ways to overcome many of these difficulties, often without the benefits provided by organized data repositories that emphasize open access, standardized metadata, and quality control. One part of my results shows that ecologists use formal and informal knowledge that they have gained through disciplinary training and through their own data-gathering experiences to help them overcome hurdles related to finding, acquiring, and validating data collected by others.

A second part of my findings reveals that ecologists rely on formal notions of scientific practice that emphasize objectivity to justify the methods they use to collect data for reuse. Ecologists are aware that other scientists will examine the methods they use to collect and interpret data, so they work hard to "make their measurements demonstrably rational and accountable" [32]. This is true whether they create new knowledge based on data they have collected themselves or use data gathered by others.

The findings summarized above are discussed in greater detail in the sections that follow. I begin with an analysis of the knowledge that ecologists acquire from having collected their own data in the field or laboratory and that serves as important background to their experiences.⁴ From here, I discuss the approaches that ecologists use to gather data for reuse and the rationales that drive these methods. I conclude the section with an examination of the limitations of ecologists' approaches.

4.1 Preludes to data reuse

Data are the basic building blocks of scientific argument, and researchers must understand them, or they risk misinterpretation based on inappropriate use of data. For this reason, ecologists go to great lengths to ensure that they understand data collected by others. They accomplish this by selecting data for reuse which are similar to those they have collected themselves. Formal disciplinary training along with insights gained in the field lead to familiarity with particular types of data.

³ All names are pseudonyms.

⁴ Even those ecologists who performed laboratory analyses possessed substantial field experience. For example, in order to analyze plant nitrogen content one has to gather specimens from the field.

As one ecologist explained in talking about the data he reused, "I chose it to be the kind of information that is readily available and that I am familiar with." Familiarity with particular types of data provides ecologists with the specialized knowledge to distinguish true patterns or variations in nature from artifacts of data that are the result of inaccurate observations or experimental error. Roth and Bowen [34] observed that fieldwork experience has a formative function in evolving the formal (academic) and informal (anecdotal) knowledge of field ecologists. 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" [34]. 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. Specifically, it helps provide ecologists with a sense that data are available, and it offers direction to their search for data.

Familiarity with particular types of data, acquaintance with general research trends, and specific knowledge about who is working in what areas provide ecologists with insights into the types of data that are available for reuse. At the outset, most ecologists thought they would be able to acquire the data they needed. As one interviewee explained:

Within my area, I knew that I could do it 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... 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 spatial and temporal scales.

Ecologists' knowledge was also useful in providing initial direction to their search for sources of data. For instance, ecologists who extracted data from the published literature were already aware of publications that contained data. As Michael said, before 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, so he was aware early on that his strategy for obtaining data would need to include sources beyond the published literature. Since Alan had some previous experience with museums, he "knew that if they had specimens at all, they likely would have body weight measurements." Over the years, scientific questions change and technologies improve, and knowledge of these shifts can further 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 for his purposes:

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.

Conversely, Katherine noted that "this particular topic, most of the work on it has been done in the last twenty years," so she focused her search on this time period. Armed with a general sense of where the data they need might reside, ecologists were ready to begin their search.

4.2 Finding data: satisfying multiple requirements

In any research project, the acquisition of data is guided by a doable research problem [11]. Ecologists' familiarity with particular types of data and areas of research helped them to form doable research questions, which in turn enabled them to develop specific criteria for data that guided their search. The methods that ecologists used to find data went beyond simply helping them to locate relevant data; they had to meet other important criteria as well. In particular, ecologists' methods were designed to meet standards of scientific practice, increase their access to data, and reduce concerns about data quality.

4.2.1 Meeting standards of scientific practice

Among my interviewees, public sources of data and information about data included natural history museums, published literature, and bibliographic databases, as well as databases available on CD-ROM, over the Internet, or through a public data center. These public sources sometimes led to "private" stores of data available from people or institutions that ecologists contacted by letter, telephone, or electronic mail. The public data sources that ecologists used served multiple requirements, one of which was that they satisfied ecologists' need to meet standards of scientific practice related to objectivity and representativeness.

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 sampling choices. 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."⁵

⁵ Biosis is a major electronic database for researchers in diverse fields of life sciences.

In some cases, the use of an existing data set that provides its own bounds satisfies the requirement for a representative sample. For example, Cal used a publicly available Internet database compiled by other ecologists that contained every recorded instance of the biological control of insects. This database was considered to be comprehensive and reliable. Ecology has few such databases available for secondary data use, however, so ecologists often have to find other means to place boundaries around their acquisition of data. The ecologists I interviewed accomplished this objective through the use of the published literature, time-frame limitations, and/or geographical restrictions.

Almost half of the ecologists that I interviewed (n=6) utilized the published literature to provide a frame around a segment of the world of data. In some cases, ecologists searched bibliographic databases to identify published literature, particularly journal papers that might contain data of interest.⁶ Often, searches of bibliographic databases were used in combination with hand searches of the literature. Ellen described a study she conducted 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.

Literature cited in relevant papers added to ecologists' confidence that their methods retrieved data from the existing pool. This approach also helped ecologists get beyond what Stephen referred to as "the digital curtain." Although database coverage is improving, many databases are limited in time. For example, the Biosis database covers the literature published from 1969 to the present. Sometimes, the literature served another role; it clued ecologists into the existence of unpublished data held by scientists. Some ecologists pursued these leads to obtain more data.

⁶ Ecologists extracted relevant data from graphs, tables, appendices, or text.

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. These searches were sometimes restricted to a particular time period as a means to further frame the data sample. For example, Cal collected data from three journals to use in a meta-analysis he had conducted shortly before I interviewed him. He was concerned about biasing his results based on the journals he chose. This was a special problem since he works in the tropics, an area of ecology that is not covered evenly by ecological journals. Cal chose a well-known but relatively new journal on tropical ecology to anchor his sampling design. This choice 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 that I interviewed, 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, they 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. Statistical techniques exist to address these issues [14]. 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. For

instance, practical issues of time versus effort and the nature of a particular area of ecology 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." 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. Peer-reviewed publications were generally also a good source of information about data, although ecologists did not rely on peer review to certify data quality. Nearly all the ecologists I interviewed cited examples of suspicious data reported in peer-reviewed publications. They recognized that one rarely sees the actual raw data on which a paper is based. Ecologists did not recommend abandoning peer review, but they recognized its limitations, particularly in terms of assuring data quality.

Geography was another frame that ecologists utilized to place boundaries around their collection of data for reuse. Sometimes, these bounds 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 found in bird journals, a bird-banding database, and museum records to gather as much data as possible on bird sightings 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.

4.2.2 Increasing access to data and decreasing concerns about data quality

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. For one, the acquisition of low volumes of data from multiple sources reduced the chances of error. 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 of bad data, hopefully it doesn't make a difference.

Further, 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. Ecologists who requested data directly from individuals or from institutions, such as natural history museums, made very specific requests. Ecologists' explicit 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.

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 e-mail 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.

The amount of data requested and the proposed use of data may also 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. Scientists may be willing to share portions of a data set, but not the entire data set.

The source of the data played a role in data acquisition choices, too. When ecologists obtained data directly from publications, there was no need to seek further permission to use the data. 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.

4.3 Limitations of ecologists' methods

The methods that ecologists use to locate data for reuse fulfill many of their requirements, but they are not without limitations. In this section, I discuss two

issues faced by the ecologists that I interviewed. One difficulty stems from the lack of approaches that offer both precision and recall when it comes to locating data for reuse. A second, and more subtle, challenge relates to data ownership and its affect on the easy and efficient retrieval of data or information about data.

The multiple purposes for which data are collected hinder their retrieval. In particular, there is a mismatch between ecologists' criteria for data, which are very specific, and the level of detail provided in abstracts or in other information indexed by bibliographic databases. Ecologists noted that the information presented in bibliographic databases is often insufficient to determine the purpose of a particular study. This factor makes it difficult to discern if a publication will contain the necessary data. Michael's experience was common among those ecologists who searched bibliographic databases to locate data in publications. Michael's study required zooplankton data that could be used to address questions of population abundance. His search for data was complicated by the fact that ecologists might collect zooplankton in order to identify the species that exist in a particular lake, or they may gather zooplankton as a way 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, so only surveys based on systematic sampling schemes could provide the data he needed:

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.⁷

As Katherine summarized the problem, "There are a certain number of experiments that will not fit your criteria because they have a different goal."

The literature says that ecological data have a high level of ownership, which is seen as a significant barrier to their sharing. Not surprisingly, the

⁷ Web of Science is an electronic bibliographic database produced by the Institute for Scientific Information.

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 they discussed their experiences in acquiring data owned by individuals or institutions. For example, 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 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.

Ellen's experience shows that data owned by institutions can present similar challenges. Ellen reused forestry data collected by an organization with which her boss was affiliated. He was committed to the reuse of the data, and Ellen was co-located with and knew many members of the group responsible for the data. Even so, she stepped carefully to gain access to the information she needed 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.

Nearly all ecologists I interviewed 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."

In summary, ecologists' approaches to find data to reuse rely heavily on knowledge they have gained through their own data-collection experiences. This knowledge provides them with a sense that data exist, offers clues to the location of relevant data, directs their search for data, and helps them to obtain data in spite of cultural, scientific, and technical obstacles. However, knowing that data exist and can be found is not sufficient. Ecologists also choose methods that meet standards of scientific practice and that can be defended publicly. The emphasis on objective norms of scientific practice leads ecologists to seek strategies that *bound* their collection of data, in order to draw data to reuse from some representative sample. Taken together, the methods that ecologists use to locate data for reuse are intended to help increase their access to data, reduce the potential for error, and provide believable rationales for their choices.

5 Conclusions

In this paper I analyzed the methods that ecologists use to locate data for reuse. My findings show that ecologists are able to overcome many of the challenges associated with the dispersed, complex, and heterogeneous nature of ecological data. My results also point to limitations in the approaches that ecologists use. In this section, I discuss the near-term implications of these findings as they relate to traditional and expanded roles for digital libraries in helping scientists to locate and retrieve data for reuse and to opportunities to explore partnerships between digital libraries, technologists, and scientists in the realm of scientific data. I also analyze some of the more intractable challenges that require a longer time-scale to resolve. I begin with a discussion of the types of scientific databases as it provides a useful frame of reference for my conclusions.

Visions of eScience tend to emphasize large-scale databases that require massive storage capabilities, robust infrastructure for data management and movement, and sophisticated tools for visualization and analysis. While data sets with these requirements are already prevalent in some scientific communities, and are likely to become more common in the future, they are only one type of data collection [18]. A recent report by the National Science Board (NSB) classified

digital data collections into three categories: 1) research data collections, 2) resource or community data collections, and 3) reference collections [30].⁸

Research data collections, the kind currently most common in ecology, are created as part of a focused research project and are intended to serve a specific and limited number of users and may not be preserved beyond the end of the study. In addition, these data are usually stored in diverse formats and lack formal metadata; this lack of standardization and documentation makes them difficult to share outside the research group that generated the data. The NSB noted that there are thousands of research databases across many fields. The other two types, community and reference data collections, are alike in their attention to standards and in their focus on making data available for reuse. Community databases serve a single domain, however, whereas reference data collections are meant to serve large portions of the education and research communities. Reference databases are more likely to have long-term support and to be preserved and maintained indefinitely. The NSB classification, together with the findings from this study, provides a practical framework in which to consider the data-related roles that digital libraries might play alone or with others.

Arguably, the most difficult data to locate and to reuse are those from research data collections because they are not collected with the intention to serve a broader community, nor are they packaged and presented in a way that makes them easy to incorporate into new projects [23]. My study shows, however, that these data are valuable in answering new research questions, that ecologists find ways to locate and reuse them, and that libraries play a valuable role in providing ecologists with access to these data because they collect and maintain the documents in which research data are often reported. Almost half the ecologists I interviewed relied on the published literature as their primary source to locate and obtain data for reuse. In addition, they used bibliographic databases available through their libraries to search for publications that might contain data of interest to them. These methods are common outside ecology as indicated by the development of software to extract data from the literature and as shown in the wide use of meta-analysis, which typically relies on data collected from journal

⁸ The NSB uses the phrase "data collection" to refer to a database or group of databases and to the infrastructure, organization, and individuals essential to manage the collection.

papers [37,39]. Thus, the results from this study have broad implications for digital libraries, and they emphasize the importance of traditional roles for libraries and the potential for new ones.

Traditionally, libraries have provided access to comprehensive collections, tools that help people locate information in those collections, support for the use of collections, and long-term preservation of the materials under their care. These long-standing roles remain important in the context of data sharing and reuse for the reasons listed above. The steady increase in the direct links that libraries provide from records in a bibliographic database to the full-text sources should help scientists to determine more quickly if a particular publication contains the data they need. There are opportunities for digital libraries to go beyond this role, however, and to more actively serve the needs of those seeking data for reuse.

Although ecologists currently rely on published literature because it meets their needs for bounded sources of data, they will use unpublished data if they can find a means to frame their data collection and if they can obtain the information necessary to understand the data. Ecologists also obtain data from community and reference data collections where they exist, as do scientists in other disciplines. This argues for digital libraries to expand their collections and services in two specific ways. First, digital libraries should continue to increase the diversity of information they collect, and they should recognize that some of this information will be valuable as sources of data for reuse. This is not a new idea. In the 1980's, for example, 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 [10]. More than 20 years later, the political, scientific, and social demands for data, in concert with technical advancements, may finally make this vision a reality. Second, librarians should encourage the use of their collections as sources for data by integrating new tools into the digital library, such as those mentioned above that help individuals to extract data from publications. This goal is likely to require cooperation with technologists, scientists, and others and to lead to ideas for new eScience applications; I discuss these possibilities in more detail below.

The NSB categorization can help digital libraries to focus their efforts in terms of whether to provide information about data, direct access to data, or the

active collection of data, including preservation functions [20]. For example, digital libraries may need only to furnish pointers to reference data collections since the providers of these data typically offer user support and long-term preservation. On the other hand, many libraries are creating institutional repositories, and through this mechanism they are likely to have a role in providing direct access to data.⁹ If digital libraries want to provide researchers with bounded data collections, then they should work together to link information about like databases together across their institutional repositories. This is a much more challenging problem than simply providing links to reference or community collections and one that is likely to require partnerships with those outside the digital library community to develop new eScience tools. These are only a few examples of how the database typology can help librarians define their roles and their potential collaborators. Budget, staffing, and technical capacity will also affect their decisions.

The findings from this study make clear that the reuse of data to create new knowledge is "doing science," and is therefore subject to the same norms, requirements, and challenges that affect all research. Therefore, it should not be surprising that the use of existing data to create new knowledge is a complex, difficult, and iterative process. While there is much progress to be made in making it easier for scientists to find, retrieve, aggregate, and understand data, the practice of science can never be completely automated. There are two particularly challenging issues that lack simple solutions in the short term. These areas are ripe for cooperation between librarians, scientists, and technologists.

First, the findings from this study show that ecologists use a combination of formal and informal knowledge to locate and reuse data collected by others. Even the most sophisticated ontologies are unlikely to capture the full range of information that scientists need to find and use data collected by others. For example, one challenge for the ecologists in this study was distinguishing between the purposes for which similar types of data were collected. Informal knowledge, which is important in all domains, is especially difficult to capture and to share

⁹ An institutional repository is a digital archive of the intellectual products created by the faculty, research staff, and students of an institution and accessible to end users both within and outside of the institution [15].

[12]. Cyberinfrastructure must find ways to support the sharing of tacit knowledge and to encourage interactions around data between users of data and those with knowledge of them [5]. These tools should be part of digital libraries and other places where knowledge creation occurs. The development of these technologies is a significant challenge for the future for digital libraries and eScience.

Second, the incentives to share and deposit scientific data are closely tied to the rewards that scientists receive for this activity. Digital libraries can include mechanisms and tools that might encourage scientists to share data, but they alone cannot resolve issues related to incentives for data sharing. Previous research shows that successful data sharing and reuse 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, including leadership from key individuals, community acknowledgment of the importance of data sharing, and requirements from key journals and funding agencies [13].

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, such as theoreticians, for ecologists who use data outside their community, and for non-ecologists who reuse ecological data. However, in all cases the reuse of data is hard work. Data are interpretable in multiple ways and immersed in context, and thus, all types of knowledge are needed to reuse them. Ecology teaches us that to be effective vehicles of data sharing, digital libraries and data repositories must capture all the dimensions of knowledge necessary for data reuse. Reaching this goal will require collaborations among librarians, CI developers, scientists, and others.

Acknowledgements I gratefully acknowledge the ecologists who so willingly shared their experiences with me. For their comments on the manuscript, I thank Katherine Lawrence, Thomas Zimmerman, and three anonymous reviewers.

References

1. Andelman S, Bowles C, Willig M, Waide R (2004). Understanding environmental complexity through a distributed knowledge network. *BioScience* 54: 240-246
2. Atkins D, Droegemeier K, Feldman S, Garcia-Molina H, Messerschmitt D, Messina P, Ostriker J, Wright H (2003) *Revolutionizing Science and Engineering through*

Pre-publication version

Final version appeared in: *International Journal on Digital Libraries*, 2007, vol. 7, nos. 1-2, pp. 5-16. <http://www.springerlink.com/content/p42u8177421u1477/references/>

- Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Panel on Cyberinfrastructure. National Science Foundation, Washington, DC
3. Baskin Y (1997) Center seeks synthesis to make ecology more useful. *Science* 275(5298): 310-311
 4. Berman F, Brady H (2005). Final Report: NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Sciences
<http://vis.sdsc.edu/sbe/reports/SBE-CISE-FINAL.pdf> [last visited November 2006]
 5. Birnholtz JP, Bietz MJ (2003) Data at work: supporting sharing in science and engineering. Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work. ACM Press, New York, pp 339-348
 6. Bowker GC (2000) Biodiversity datadiversity. *Social Studies of Science* 30(5): 643-683
 7. Bowser CJ (1986) Historic data sets: lessons from the past, lessons for the future. In Michener WK (ed) *Research Data Management in the Ecological Sciences*. University of South Carolina Press, Columbia, pp 155-179
 8. Brown C (2003) The changing face of scientific discourse: analysis of genomic and proteomic database usage and acceptance. *Journal of the American Society for Information Science and Technology* 54 (10): 926-938
 9. Buetow KH (2005). Cyberinfrastructure: empowering a "Third Way" in biomedical research. *Science* 308(5723): 821-824
 10. Chen, C, Herson, P (eds) (1984) *Numeric databases*. Ablex, Norwood, NJ (1984)
 11. Clarke AE, Fujimura JH (1992) What tools? Which jobs? Why right? In Clarke AE, Fujimura JH (eds) *The Right Tools for the Job: At Work in Twentieth-Century Life Sciences*. pp. 3-44. Princeton University Press, Princeton, NJ (1992)
 12. Collins HM 1992 [1985] *Changing Order: Replication and Induction in Scientific Practice*. University of Chicago Press, Chicago, IL
 13. 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). Ecological Society of America, Washington, DC
 14. Cooper H, Hedges LV (1994) *The Handbook of Research Synthesis*. Russell Sage Foundation, New York
 15. Crow R (2002) *The Case for Institutional Repositories: A SPARC Position Paper*. The Scholarly Publishing & Academic Resources Coalition, Washington, DC
 16. Ecological Visions Committee (2004) *Ecological Science and Sustainability for a Crowded Planet: 21st Century Vision and Action Plan for the Ecological Society of America*. Ecological Society of America, Washington, DC
 17. Ellisman MH (2005) Cyberinfrastructure and the future of collaborative work. *Issues in Science and Technology* 22(1): 43-50
 18. Emmott S (2006) *Towards 2020 Science*. Microsoft Research, Redmond, WA
 19. Glasner P (2002) Beyond the genome: reconstituting the new genetics. *New Genetics and Society* 21(3): 267-277

20. Gray AS, Dodd SA (1984) The roles of libraries and information centers in providing access to numeric databases. In Chen C, Hernon P (eds) *Numeric Databases*. Ablex, Norwood, NJ, pp 247-262
21. Hey T, Trefethen AE (2005) Cyberinfrastructure for e-Science. *Science* 308(5723): 817-821
22. Hilgartner S (1995) Biomolecular databases: new communication regimes for biology? *Science Communication* 17(2): 240-263
23. Hine, C (2006) Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science* 36(2): 269-298
24. Kuklick H, Kohler RE (1996) Science in the field: introduction. *OSIRIS*, Second series 11:1-16
25. Lynch, C (2006) Research libraries engage the digital world: A US-UK comparative examination of recent history and future prospects. *Ariadne* 46
<http://www.ariadne.ac.uk/issue46/lynch/> [last visited November 2006]
26. Michener WK, Brunt JW (eds) (2000) *Ecological Data: Design, Management, and Processing*. Blackwell Science, Oxford, UK
27. Michener WK, Brunt JW, Helly JJ, Kirchner TB, Stafford SG (1997) Nongeospatial metadata for the ecological sciences. *Ecological Applications* 7(1): 330-342
28. National Research Council (1995) *Finding the Forest in the Trees: The Challenge of Combining Diverse Environmental Data: Selected Case Studies*. National Academy Press, Washington, DC
29. National Research Council (2003) *NEON: Addressing the Nation's Environmental Challenges*. National Academy Press, Washington, DC
30. National Science Board (2005) *Long-Lived Data Collections: Enabling Research and Education in the 21st Century*. National Science Foundation, Arlington, VA
31. Pouchard L, Woolf A, Bernholdt D (2005) Data grid discovery and semantic web technologies for the earth sciences. *International Journal on Digital Libraries* 5: 72-83
32. Roth W-M, Bowen GM (1999) Digitizing lizards: the topology of 'vision' in ecological fieldwork. *Social Studies of Science* 29(5): 719-764
33. Roth W-M, Bowen GM (2001) 'Creative solutions' and 'fibbing results': enculturation in field ecology. *Social Studies of Science* 31(4): 533-556
34. Roth W-M, Bowen GM (2001) Of disciplined minds and disciplined bodies: on becoming an ecologist. *Qualitative Sociology* 24(4): 459-481
35. Schiff LR, Van House NA, Butler MH (1997) Understanding complex information environments: a social analysis of watershed planning. In Allen RB, Rasmussen EM (eds) *Proceedings of the Second ACM International Conference on Digital Libraries*. ACM Press, New York, pp 161-168
36. Smith HJ, Keil M (2003) The reluctance to report bad news on troubled software projects: a theoretical model. *Information Systems Journal* 13: 69-95
37. Smith, JT (1996) Meta-analysis: the librarian as a member of an interdisciplinary research team. *Library Trends* 45(2): 265-279

Pre-publication version

Final version appeared in: *International Journal on Digital Libraries*, 2007, vol. 7, nos. 1-2, pp. 5-16. <http://www.springerlink.com/content/p42u8177421u1477/references/>

38. Star SL, Ruhleder K (1996) The ecology of infrastructure: problems in the implementation of large-scale information systems. *Information System Research* 7: 111-134
39. Weeber M, Kors JA, Mons B (2005) Online tools to support literature-based discovery in the life sciences. *Briefings in Biomedical Informatics* 6(3): 277-286
40. Wouters P, Reddy C (2003) Big science data policies. In Wouters P, Schröder, P (eds) *Promise and Practice in Data Sharing*, NIWI-KNAW, Amsterdam, pp 13-40
41. Zimmerman, A (2003) *Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists*. Unpublished dissertation, University of Michigan, Ann Arbor, MI
42. Zimmerman A (in press) New knowledge from old data: the role of standards in the sharing and reuse of ecological data. *Science, Technology, and Human Values*

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 1: Key data used by ecologists