Title
Little Science Confronts the Data Deluge: Habitat Ecology, Embedded Sensor Networks, and Digital Libraries

Journal
Center for Embedded Network Sensing, 7(1–2)

Authors
Borgman, C L
Wallis, J C
Enyedy, N

Publication Date
2006-11-25

DOI
10.1007/s00799-007-0022-9

Peer reviewed

Little Science Confronts the Data Deluge: Habitat Ecology, Embedded Sensor Networks, and Digital Libraries

Christine L. Borgman
Department of Information Studies
Graduate School of Education & Information Studies, UCLA
borgman@gseis.ucla.edu

Jillian C. Wallis
Center for Embedded Networked Sensing, UCLA
jwallisi@ucla.edu

Noel Enyedy
Department of Education
Graduate School of Education & Information Studies, UCLA
enyedy@gseis.ucla.edu

Abstract

e-Science promises to increase the pace of science via fast, distributed access to computational resources, analytical tools, and digital libraries. "Big science" fields such as physics and astronomy that collaborate around expensive instrumentation have constructed shared digital libraries to manage their data and documents, while "little science" research areas that gather data through hand-crafted fieldwork continue to manage their data locally. As habitat ecology researchers begin to deploy embedded sensor networks, they are confronting an array of challenges in capturing, organizing, and managing large amounts of data. The scientists and their partners in computer science and engineering make use of common datasets but interpret the data differently. Studies of this field in transition offer insights into the role of digital libraries in e-Science, how data practices evolve as science becomes more instrumented, and how scientists, computer scientists, and engineers collaborate around data. Among the lessons learned are that data on the same variables are gathered by multiple means, that data exist in many states and in many places, and that publication practices often drive data collection practices. Data sharing is embraced in principle but little sharing actually occurs, due to interrelated factors such as lack of demand, lack of standards, and concerns about publication, ownership, data quality, and ethics. We explore the implications of these findings for data policy and digital library architecture. Research reported here is affiliated with the Center for Embedded Networked Sensing.

Introduction

Scientists are facing a deluge of data beyond what can be captured, managed, or interpreted with traditional tools. While "big science" fields such as physics and astronomy (Price, 1963) have begun to construct tools and repositories to address this deluge, "little science" areas dependent upon fieldwork lack the tools and infrastructure to manage the growing amounts of data generated by new forms of instrumentation. The lack of an integrated framework for managing these types of scientific data presents significant barriers not only to those scientists conducting the research, but also to those who would subsequently reuse the data.

Scientific data are expensive to produce, but can be of tremendous future value. Data associated

with specific times and places, such as ecological observations, are irreplaceable. They are valuable to multiple communities of scientists, to students, and to nonscientists such as public policy makers. Research on scientific data practices has concentrated on big science such as physics (Traweek, 1992; 2004) or on large collaborations in areas such as biodiversity (Bowker, 2000a; b; c). Equally important in understanding scientific data practices is the study of science areas in which small teams produce observations of long-term, multi-disciplinary, international value. Results from local projects can be aggregated across sites and times, offering the potential to advance the environmental sciences significantly.

Habitat ecology is an optimal case to address these issues, as this research area is in a transition phase from hand-crafted fieldwork to highly instrumented data collection via embedded sensor networks. The choice of research problems and methods in environmental research were greatly influenced by the introduction of remote sensing (satellite) technology in the 1980s and 1990s (Kwa, 2005). Thus one of our research concerns is how habitat ecology may evolve with the use of *in situ* sensing technologies. These scientists are deploying dense sensor networks in field locations to study plant growth, bird behavior, water quality, micrometeorological variations, and other ecological factors. This research community needs consistent, generalizable, scalable tools to manage and share data.

However, little is understood about how scientists in these areas produce, use, or manage data, or how data management practices vary between these scientists and their partners in computer science and engineering. We are studying data practices in newly instrumented areas of habitat ecology and closely related areas of the environmental sciences as a means to learn more about how small science can benefit from e-Science. Findings from our data practices research are used to illustrate digital library requirements for e-Science.

The Role of Data in e-Science

The volume of scientific data being generated by highly instrumented research projects (linear accelerators, sensor networks, satellites, seismographs, etc.) is so great that it can be captured and managed only with the use of information technology. The need to manage the "data deluge" is among the main drivers of e-Science and cyberinfrastructure (Hey & Trefethen, 2003; 2005; Lord & Macdonald, 2003). If these data can be stored in reusable forms, they can be shared over distributed networks. Data are becoming an important end product of scholarship, complementing the traditional role of publications.

Big Science, Little Science

"Big Science," as coined by Weinberg (1961), reflects the large, complex scientific endeavors in which society makes major investments. These are characterized by expensive equipment that must be shared among many collaborators, such as particle accelerators or space stations. e-Science and cyberinfrastructure are big science in this sense, as they are major societal investments.

Derek de Solla Price (1963), in his canonical work *Little Science, Big Science,* distinguished between little and big science not by size of projects but by the maturity of science as an enterprise. Modern science, or big science in Price's terms, is characterized by international, collaborative efforts and by invisible colleges of researchers who know each other and who exchange information on a formal and informal basis. Little science is the 300 years of independent, smaller scale work to develop theory and method for understanding research problems.

Differences between little and big science are more qualitative than quantitative. Big science encourages standardization of processes and products, and thus the growth of digital libraries and data repositories and of metadata standards are predictable outcomes of the trajectory from little to big. The technical infrastructure of e-Science is especially suited to supporting large-scale international collaborations by providing distributed access to instruments, to computational resources, and to digital

libraries of data. Not surprisingly, science domains such as physics and astronomy were among the first to build distributed digital libraries in support of collaborative research (*ArXiv.org e-Print archive*, 2006; Ginsparg, 2001; *International Virtual Observatory Alliance*, 2006). Distributed digital libraries are being created to support many other scientific domains, including the environmental sciences and water resources (*Collaborative Large-Scale Engineering Analysis Network for Environmental Research*, 2006; *Consortium of Universities for Advancement of Hydrologic Science*, 2006; *Global Earth Observation System of Systems*, 2006; *National Ecological Observatory Network*, 2006).

Digital libraries for scientific documents and data can facilitate collaboration and promote the progress of science. They also can hinder progress by forcing standardization prematurely (Bishop, Van House & Buttenfield, 2003; Bowker, 2005). Many scientific research areas continue to be productive without the use of shared instrumentation, shared repositories, or agreements on standards for data description. As research areas such as habitat ecology become more instrumented, they are facing many challenges associated with the transition from little science to big science, including what to standardize, when, and for what purposes.

The Role of Data in Science

Modern science is distinguished by the extent to which its practices rely on the generation, dissemination, and analysis of data. These practices are themselves distinguished both by the massive scale of data production and by the global dispersion of data resources. The rates of data generation in most fields are expected to increase even faster with new forms of instrumentation such as embedded sensor networks. Consequently, scientists need assistance in identifying and selecting data that are useful in individual contexts, and preserving and curating data that are of future value, whether to the originators or to others.

Notions of what are "data," to whom, when, and for what purposes vary widely. The following is a simple and widely cited technical definition:

> *Data*: A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen (*Reference Model for an Open Archival Information System*, 2002, p. 1-9).

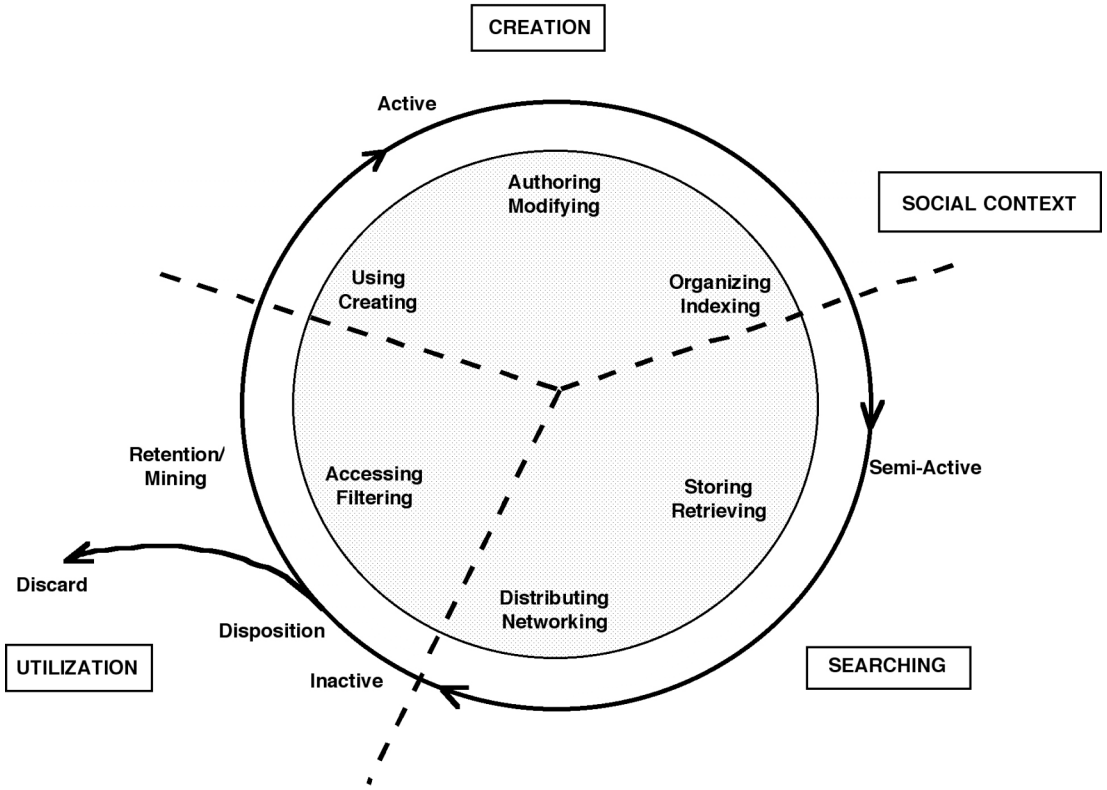
Scientific data can be grouped into the categories of observations, computations, experiments, and record-keeping (Hodge & Frangakis, 2005; *Long-Lived Digital Data Collections: Enabling Research and Education for the 21st Century*, 2005). *Observational* data include weather measurements, which are associated with specific places and times; they can also be used in cross-sectional or longitudinal studies. *Computational* data result from executing a computer model or simulation, whether from a physics experiment or from acoustical arrays that locate the position of singing birds. Replicating the model or simulation in the future may require extensive documentation of the hardware, software, and input data. In some cases, only the output of the model might be preserved. *Experimental* data include results from laboratory studies such as measurements of chemical reactions or from field experiments on plant growth under different light and soil conditions. Whether sufficient data and documentation are kept to reproduce the experiment varies by the cost and reproducibility of the experiment. *Records* of government, business, and public and private life also yield useful data for scientific, social scientific, and humanistic research.

These technical descriptions of "data" obscure the social context in which data exist, however. Observations that are research findings for one scientist may be background context to another. Data that are adequate evidence for one purpose (e.g., determining whether water quality is safe for surfing) are inadequate for others (e.g., government standards for testing drinking water). Similarly, data that are synthesized for one purpose may be "raw" for another (Borgman, 2007; Bowker, 2005). These are among the many complexities of data practices that we are exploring in this research.

Data from Embedded Sensor Networks

The need to develop and deploy an integrated framework for data management is no more keenly felt than by scientists who generate massive quantities of data via wireless sensor networks (Akyildiz et al., 2002; Culler & Hong, 2004; Elson & Estrin, 2004; Pottie & Kaiser, 2000; 2006). These are systems of sensors that are embedded in the environment of the phenomena on which data are sought, and connected via communication networks so that data from numerous locations can be collated and analyzed either within the network or external to it. In habitat ecology, scientists use multiple types of sensors to observe phenomena, each at different sampling rates. Sensors vary widely in type and capability. Some sensors capture data continuously or at discrete intervals for indefinite periods of time; some sensors are activated only when triggered by an event (e.g., the movement of an animal into the field of vision of a camera), requiring Bayesian statistical models. Once collected, sensor data may be analyzed at various frequencies and levels of granularity, depending on the scientific topic and research question. Earthquake data, for example, is typically of immediate interest to scientists, while ecological data is often of interest only when a sufficient period has elapsed to collect a time series. In the early stages of a project, however, scientists usually assess their data at short intervals to calibrate their data collection and instruments.

Figure 1: Information Life Cycle (Borgman et al., 1996)



NOTE: The outer ring indicates the life cycle stages (active, semi-active, and inactive) for a given type of information artifact (such as business records, artworks, documents, or scientific data). The stages are superimposed on six types of information uses or processes (shaded circle). The cycle has three major phases: information creation, searching, and utilization. The alignment of the cycle stages with the steps of information handling and process phases may vary according to the particular social or institutional context.

Data-Intensive Science as a New Paradigm

Scientific progress increasingly will depend on the existence of a common information infrastructure that enables domain scientists to exploit available data effectively and efficiently. Among the potential benefits of e-Science are: (i) new data analysis methods and smarter algorithms to tackle the ever-increasing amount of data; (ii) science centers that allow for computation on the data-server side, while supporting data access, interchange, and integration; (iii) sophisticated metadata for data access that supports physical and logical independence; and (iv) semantic convergence of data tools, crossing disciplinary and epistemological boundaries (Gray et al., 2005). All too often, scientists must become computer scientists and statisticians in addition to their chosen discipline. They need a "tool layer" to support the information life cycle from initial research design through instrumentation, data capture, data management, analysis, publication, and curation (see Figure 1 above).

Digital Libraries for Data

A key component of an integrated framework for data management is automated support for the description and annotation of data, so that those data remain easily identifiable, discoverable, and available in a useful form. At the minimum, a digital library for scientific data will involve the following activities: (i) specification of a standard communication framework (such as is provided by XML) for the communication and exchange of metadata, both among the members of the immediate research community, and between the immediate community and others; (ii) specification of the semantics (meaning) and syntax (structure) of a standard metadata schema (i.e., a standard set of metadata elements), for use by all members of the immediate research community; and (iii) implementation of tools enabling members of multiple communities to supervise the creation (manual, semi-automatic, and automatic) of metadata, as well as the analysis, use, and preservation of data. Metadata provide data independence by separating the data from the database architecture and from analysis software, thus increasing the longevity of those data.

Several technical standards developed by the digital library community will underpin distributed access to scientific data, documents, and composite objects: Reference Model for Open Archival Information Systems (OAIS), Open Archives Initiative Protocols for Metadata Harvesting (OAI-PMH), OpenURL, the Info URI scheme, and Object Reuse and Exchange (ORE) (Bekaert & Van de Sompel, 2006; Chudnov et al., 2005; *Document Action: 'The "info" URI Scheme for Information Assets with Identifiers in Public Namespaces' to Informational RFC*, 2005; *Object Reuse and Exchange*, 2006; Van de Sompel et al., 2004). OAI-PMH facilitates the creation of a platform-independent content layer to support discovery services. OpenURL provides context-sensitive services in coordination with OAI and OAIS. The Info URI scheme maps legacy namespaces into a URI format. The original structure is preserved, while enabling Web services to incorporate extant namespaces and content. Object Reuse and Exchange integrates the aforementioned standards and protocols into an interoperability framework.

These technical standards also facilitate open access to data, which is essential to continued scientific progress. A rich content layer of scientific information on the Internet also creates opportunities for public and private entities to produce tools and services (Esanu & Uhlir, 2004).

e-Science and Data Practices

e-Science initiatives state the requirement for better tools, but say little about what the criteria should be for building them. More understanding is need about practices, behaviors, and incentives associated with the collection, use, and management of scientific data. These findings are important input to the design of effective digital library systems, services, and policies.

The development of digital libraries for scientific data and the policies of funding agencies to promote deposit of data in those systems is predicated on the assumption that these data will be reused by

others (Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility, 2003; 2006). However, data sharing is more common in big science than in small science fields. Scholars in smaller science fields often assume that their data are not of value beyond a specific study or research group. Heads of small labs often have difficulty reconstructing datasets or analyses done by prior lab members, as each person used his or her own methods of data capture and analysis. Local description methods are common in fields such as environmental sciences where data sources vary widely by study (Estrin, Michener & Bonito, 2003; Zimmerman, in press-a; Zimmerman & Nardi, 2006; Zimmerman, 2003).

The degree of instrumentation of data collection is a factor in data sharing due both to the cost of equipment and to the potential reduction in manual effort to generate data. Sharing expensive equipment is among the main drivers for collaboration. In these cases, collaboration, instrumentation, and data sharing are likely to be correlated (David & Spence, 2003). The relationship between instrumentation and data sharing may be more general, however. A small but detailed study conducted at one research university found that scholars whose data collection and analysis were most automated were the most likely to share raw data and analyses; these also tended to be the larger research groups. When data production was automated but other preparation was labor-intensive, scholars were less likely to share data. Those whose data collection and analysis were the least automated and most labor-intensive were most likely to guard their data. These behaviors held across disciplines; they were not specific to science (Pritchard, Carver & Anand, 2004).

Despite the assumed value of data sharing for e-Science, scientists have a number of disincentives to sharing their data. Firstly, they are rewarded for publication, not for data management. Secondly, documenting data sufficiently for others to use them requires considerably more time and effort than documenting them only for the use of a small research team. Documenting data for later use also requires much more effort than what is required to publish data summaries in a journal article or conference paper. Scientists may be willing to share their data, but only after publication and only with certain conditions (e.g., attribution, non-commercialization). If scientific data are to be leveraged for larger communities, data digital libraries must reflect scientific practices in ways that make documenting and sharing data attractive. These may include mechanisms for personal digital libraries, attribution and provenance support, embargo periods for access, and security (Arzberger et al., 2004; Borgman, 2004; 2007; Bowker, 2005; Hilgartner & Brandt-Rauf, 1994).

Habitat Ecology as a Science in Transition

Ecology is defined as "the scientific study of the interrelationships among organisms and between organisms, and between them and all aspects, living and non-living, of their environment" (Allaby, 2006). While people tend to think of ecology in modern context, the term was first coined in 1866. "Habitat" is defined as "The living place of an organism or community, characterized by its physical or biotic properties" (Allaby, 2006). Habitat ecology researchers study relationships among plants and animals in their native environments. Their educational background is usually in biology or in areas of the environmental sciences such as environmental engineering and public health. Instrumented data collection such as embedded sensor networks is relatively new, and is leading to new methods and new research questions.

CENS as a Context to Study Data Practices

Research reported here is affiliated with the *Center for Embedded Networked Sensing* (CENS), a National Science Foundation Science and Technology Center [http://www.cens.ucla.edu/]. CENS conducts collaborative research among scientists, technology researchers, and educators, crossing many disciplinary boundaries. Five universities are partners in CENS. Faculty, students, and staff from other

institutions also participate in research and outreach activities. The Center's goals are to develop, and to implement in diverse contexts, innovative wireless sensor networks. CENS' scientists are investigating fundamental properties of these systems, designing and deploying new technologies, and exploring novel scientific and educational applications. CENS' commitment to sharing its research data, combined with its interdisciplinary collaborations, make it an ideal environment in which to study scientific data practices and to construct digital library architecture to support the use and reuse of research data. The combination of science and technology research offers a rare opportunity to address questions such as differences in criteria for what constitutes "data" and what constitutes a "finding."

CENS' research crosses four scientific areas: terrestrial ecology, marine biology, environmental contaminant transport, and seismology, plus applications in urban settings and in the arts. The research reported here addresses the use of embedded networked sensor technology in the first three of these application areas. Specific research questions in these terrestrial ecology, marine biology and environmental contaminant transport are closely related to habitat ecology. Research methods are based on *in situ* monitoring, with the goal of revealing patterns and phenomena that were not previously observable. While the initial framework for CENS was based on autonomous networks, early results revealed the difficulty of specifying field requirements in advance well enough to operate systems remotely. Autonomous networks also require robust technology that can be left unattended in the field. In contrast, prototype technology, which often is delicate and expensive, can be used in controlled deployments of a few hours or a few days. For these reasons, CENS has moved toward more "human in the loop" models where investigators can adjust monitoring conditions in real time with a wider array of sensor technology.

Research Methods in Habitat Ecology

Habitat ecology research tends to iterate between induction and deduction. Exploratory research is mainly inductive, relying on observations and other data collected in the field to generate hypotheses, which are later tested through deductive experiments performed in the lab (Maurer, 2004). Their methods are rooted in the natural sciences, but also are constrained by available resources. Collecting biological samples is time-sensitive and time-intensive, often requiring sophisticated instrumentation for processing and analysis. Field methods are reflexive to the observations obtained, with many experimental design decisions being made on-site in response to current conditions.

Two brief CENS examples will illustrate the data requirements of this research area. One team is studying toxic algal blooms using a combination of sensor arrays and biological samples of the plankton and algae present in the water. These algae photosynthesize and acquire nutrients on a 24-hour cycle. When conducting field research on this project, biologists on the team set up wet-labs onsite to preserve the samples for future analysis. They must work on-site for at least one continuous 24-hour cycle to capture a complete time series. Technology partners on the team are developing aquatic sensing arrays to capture hydrographic samples, such as temperature, chlorophyll, and light at varying depths. The hydrographic profile serves two functions: these data can provide context for biological sampling, and can flag interesting phenomena that may be worthy of biological sampling and analysis. The aquatic sensor arrays can capture more samples in a short time span than is possible by manual methods. In this project, the sensor networks augment human collection of physical samples, but do not supplant requirements for on-site wet labs or the presence of biology researchers.

Another CENS team is studying soil processes with the use of embedded sensor networks. Biologists and technology researchers jointly developed a means to track underground soil activity by placing clear tubes in the ground. Digital video cameras placed in the tubes take pictures of the soil system surrounding the tube. These images track the growth of roots and the fungal structures that bind to them and act as a system, trading nutrients and water collected by the fungi for sugars produced by the plant through photosynthesis. The sensor networks serve several functions in this project. Sensors are used to capture micrometerological conditions associated with the immediate habitat; the networks

process those data. In the initial stage of the project, images were collected manually by sticking cameras down the tubes and then hand-coding the images for root growth. As the project progresses, better cameras are being mounted in the tubes, which will send images via the sensor networks. Another part of the project is to classify the soil images using computer vision algorithms. Sensor networks will enable more data to be captured, at a much higher sampling rate, than with manual methods. The project also is evolving toward more remote data collection, with less need for human monitoring of soil conditions.

Local vs. Global Science

Ecology can be studied on a local scale, such as the relationships among species in a given habitat, and on a global scale, such as patterns of species migration or of crop conditions. The transformation of methods in large-scale ecology research with the introduction of remote sensing via orbiting satellites was neither fast nor painless. Complex agreements in policy and standards were required, and practices evolved accordingly (Kwa, 2005). Habitat ecology is in the early stages of technological transition through the use of embedded sensor networks that capture data *in situ*. If these data can be captured in standardized forms that other scientists consider trustworthy, they can be stored in digital libraries and made available for shared use via the distributed networks of e-Science. Data from similar habitats could be compared across places and over time. This transition also has been slow, and not without pain.

Data gathering for comparative research on habitat and local ecology has advanced over the last several decades. The U.S. Long Term Ecology Research Network celebrated its 25th anniversary in 2005 (*U.S. Long Term Ecological Research Network*). NEON is a new effort to coordinate ecological observations across the U.S. (*National Ecological Observatory Network*, 2006). Observatories of the ocean exist in a tiered network of local systems (*Southern California Coastal Ocean Observing System*, 2006), which are part of a national system (*Integrated Ocean Observing System*, 2006; *National Office for Integrated and Sustained Ocean Observations*, 2006), which is part of an international system (*Global Ocean Observing System*, 2006). The international system for oceans, in turn, is part of a yet larger international effort to coordinate ecological data (*Global Earth Observation System of Systems*, 2006). Related international projects include the *International Biological Program* (IBP; established in 1964 by the International Council of Scientific Unions) and the *Man and the Biosphere* program (MAB; established in 1971 by the United Nations) (Michener & Brunt, 2000). These systems support multiple data types (numerical measurements, text, images, sound and video) and interact with other systems that manage geographical, meteorological, geological, chemical, and physical data.

Several XML-based standards and protocols exist for managing biocomplexity data but none have been adopted widely. The Knowledge Network for Biocomplexity, for example, offers a data management system and a metadata standard for ecological data (*Ecological Metadata Language*, 2004). The Sensor Modeling Language, a product of the OpenGIS Consortium, can be used to express ecology data captured by sensor technology. SensorML is in the final stages of being accepted as a formal standard, after many years of development (Botts, 2004; 2006; Botts & McKee, 2003; *Sensor Modeling Language*, 2005). The observatory systems are making steady progress on capturing data for which standardized measurements have been agreed, such as micrometerological records. Even here, data collection can be contentious, as basic elements such as temperature and humidity can be measured in many ways, and weather stations often are moved. One of the biggest challenges in developing effective digital libraries in habitat ecology is the "data diversity" that accompanies biodiversity (Bowker, 2000b).

Observatory data can be research results in and of themselves, but in habitat ecology they often serve as context to other research questions. Habitat ecologists observe phenomena at a local scale using relatively ad hoc methods (Zimmerman, in press-b). In CENS, for example, multiple research teams are using the micrometeorological data from sensor networks as context for their own research questions about when, why, and under what conditions do toxic algal blooms occur and root activity changes occur in the soil. These researchers collect additional data with other instruments to address specific research questions.

Thus the study of biodiversity and ecosystems remains a complex and interdisciplinary domain (Schnase et al., 1997). Mechanisms used to collect and store biological data are almost as varied as the natural world those data document. Data collection is guided more by best practices than by formal standards. To the extent that scientific maturity is characterized by the standardization of tools and practices, habitat ecology is a young field (Maurer, 2004).

Empirical Studies of Data Practices

The overarching goal of our research program on data practices is to construct systems to capture and manage scientific data in ways that will facilitate immediate use by the data creators and later reuse by the creators and others, and which will reflect fair policies for access. Multidisciplinary collaboration, which is among the great promises of e-Science, depends heavily on the ability to share data within and between fields. Research in habitat ecology is much different in character than research in computer science and engineering, and thus these collaborators vary widely in data practices.

Our research also has educational components in which data from sensor networks are used in teaching middle-school and high-school science. These aspects of the project are reported elsewhere (Borgman, 2006; Sandoval & Reiser, 2003; Thadani et al., 2006; Wallis et al., 2006), but questions about the use of scientific data for educational purposes are included in our interviews on data practices.

Collaborative research often takes much more time than individual research, due to the effort and experience required for collaborators to learn each others' terminology, methods, and research problems well enough to work together effectively (Borgman, 2006; Cummings & Kiesler, 2005; Finholt, 2003; Olson, 2005; Olson & Olson, 2000; Sonnenwald, 2007; Van House, 2003). This project is no exception. We have been working on these problems since 2001, as the CENS grant proposal was developed, and actively since CENS was funded in August, 2002. Our research problem in data practices has evolved in parallel with the maturity of the technology and of the science being conducted. In the first year (2002-2003), we sat in on team meetings across CENS to learn about scientific and engineering activities and we inventoried data standards for each area (Shankar, 2003). In the second year (2003-04), we interviewed individual scientists and teams and continued to inventory metadata standards. We used the results of the first two years to design an ethnographic study of habitat biologists, conducted in the third year (2004-05). Those results informed the design of a more comprehensive study in CENS' fourth year (2005-06). This paper includes results from the ethnographic study of 2004-05 and selected results from the 2005-2006 interviews with habitat ecologists, marine biologists, environmental scientists, and their partners in computer science and engineering.

Our current research questions address the initial stages of the data life cycle in which data are captured, and subsequent stages in which the data are cleaned, analyzed, published, curated, and made accessible (Figure 1). The questions can be categorized as follows:

- Data characteristics: What data are being generated? To whom are these research data and to whom are these context data? To whom are these data useful?
- Data sharing: When will scientists share data? With whom will they share data? What are the criteria for sharing? Who can authorize sharing? How do data sharing criteria vary between scientific, technological, and educational applications?
- Data policy: What are fair policies for providing access to these data? What controls, embargoes, usage constraints, or other limitations are needed to assure fairness of access and use? What data publication models are appropriate?
- Data architecture: What data tools are needed at the time of research design? What tools are needed for data collection and acquisition? What tools are needed for data analysis? What tools are needed for publishing data? What data models do the scientists who generate the data need? What data models do others need to use the data?

Data Collection Methods

The ethnographic work from the first three years of the study (interviewing teams and individuals, participating in working groups, etc.) is documented in notes, internal memoranda, and a white paper

(Shankar, 2003). We did not audiotape or videotape those meetings to avoid interfering with the local activities. Results of the initial studies were used to identify metadata standards relevant to the scientific domains and instruments of these research teams. In turn, these metadata standards were reviewed with the research teams to determine their suitability.

We used the results of interviews and documentary analyses in the first two years of our research to design the ethnographic study conducted in 2004-05. Interview questions were based on Activity Theory, which analyzes communities and their evolution as "activity systems" (Engeström, 1987; Engeström, Miettinen & Punamaeki, 1999). Activity systems are defined by the shared purposes that organize a community and by the ways in which joint activity is mediated by shared tools, rules for behavior, and divisions of labor. Contradictions are analyzed as the engine for organizational change. We asked participants about their motives, their understanding of their community's motives, the tools they used in daily work, ways in which they divided labor, power relations within their community, and rules and norms for the community.

The results of this small ethnographic study were combined with results of other interviews and notes from meetings of participants (Borgman, Wallis & Enyedy, 2006). Building upon those results, we designed a larger study of five ecology projects. For each project, we interviewed a complementary set of science and technology participants, including faculty, post-doctoral fellows, graduate students, and research staff. One member of our team participated in a 4-day field deployment for one of these projects to observe how data are produced, captured, and managed, both from sensor networks and other methods.

Participants

CENS is comprised of more than 80 faculty members and other researchers, including a varying number of post-doctoral researchers, student researchers, and full-time research staff affiliated with the five participating universities. Many other individuals participate in CENS research activities via collaborations with members of CENS' teams, summer internships, industry partnerships, or other relationships. Of this large community, about 50 people are working on scientific or technological aspects of habitat ecology and related areas of environmental sciences.

The ethnographic study consisted of in-depth interviews with two participants. We interviewed one subject on three occasions for a total of four hours. The other subject was interviewed one time for approximately two hours. The intensive interview study consisted of 22 participants working on the five ecology projects; half the participants were from the science domains and half were from computer science and engineering. Interviews ranged from 45 minutes to two hours in length, averaging about 60 minutes. Also included in the analysis presented here are notes from a group meeting (about 20 people) to discuss data sharing policy.

Qualitative Data Analysis

The ethnographic interviews with two participants were transcribed from the audiotapes. The transcripts are complemented by the interviewers' memos on topics and themes (Lofland et al., 2006). Analysis proceeded in sequence from the first interviews with each participant to identify emergent themes, then to test and refine these themes in coding of subsequent interviews. With each refinement, the remaining corpus was searched for confirming or contradictory evidence. Analysis focused on themes rather than extensive coding of variables. These results, complemented by other meetings and interviews, were sufficiently informative to design the fuller study on data practices in habitat ecology.

The 22 interviews were audiotaped and transcribed. These interviewers also wrote memos after each interview to aid in interpreting the results. The transcripts are now fully coded using NVIVO; analysis of this extensive data set (more than 300 pages of transcripts) is still in progress. Both studies use the methods of grounded theory (Glaser & Strauss, 1967) to identify themes and to test them in the full corpus of interview transcripts and notes.

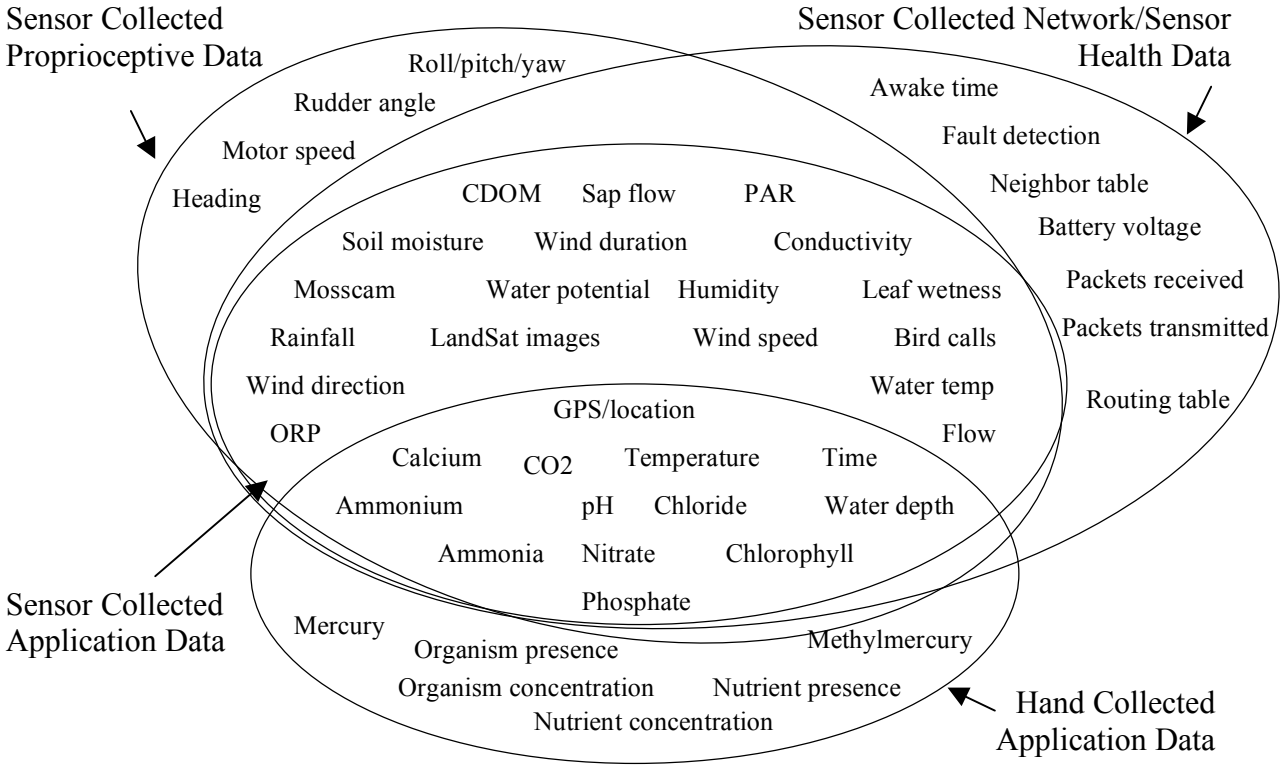
Lessons Learned

Findings from our studies of data practices are used to identify design and policy considerations for digital libraries in e-Science, which is the theme of this special issue. Habitat ecology research is in the early stages of transition from hand-crafted to technology-intensive data collection. Lessons learned from their experience, and that of closely related environmental science domains, can inform e-Science design and policy for other "small science" research areas. Findings are organized by the research questions identified above: Data characteristics, data sharing, data policy, and data architecture.

Data Characteristics

Our interview questions explored what data are being generated, to whom are these research data, to whom are they context data, and to whom are they useful. CENS technologies are being evaluated and field-tested concurrently with scientific data collection. Figure 2 illustrates the diversity of data being collected and the purposes for which they are being used. We grouped the variables into four overlapping categories of sensor-collected application data, hand -collected application data, sensor-collected performance data, and sensor-collected proprioceptive data. These are example variables rather than a complete list. Note that the application scientists often are collecting the same variables with sensors and with other technologies, usually to calibrate the sensors. Data collected for scientific purposes also is useful for computer science and engineering research, either to assess the performance of the sensing technology (e.g. packets transmitted and received, battery life) or to guide robotic sensors in boats or other devices (e.g., motor speed, rudder angle). Whether the converse is true is a question we are pursuing further. Performance data and proprioceptive data about the sensors may be of value to the scientists as context for their research questions, but these are not data they would normally collect for scientific purposes.

Figure 2: Diversity of Data Variables Collected by CENS Researchers



The scientists appear to be collecting the same variables by different means both to calibrate sensors and because the sensor technology was insufficiently reliable. They were losing too much data to trust the output of the sensors (Borgman et al., 2006). Issues for further study are to identify scientists' requirements for trust and validation of sensor data, and to distinguish between issues of scientific validity and issues of technology maturity.

While both the science and engineering teams use the scientific data in their research questions, they do so for different purposes and at different levels of granularity. The scientists assess the numerical data to discover trends, whether in growth patterns of plants or diurnal cycles of algae in lakes. The engineering teams may find the presence or absence of data from a sensor sufficient to monitor system performance. Requirements for data accuracy and "cleaniness" vary considerably between the science and technology research teams.

The ethnographic interviews offered insights into the distinction between experimental data and contextual data. Experimental data are those that reflect the hypotheses and research questions of the investigator. Contextual data include micrometeorological measurements (temperature, humidity, etc.) and calibration of tools and instruments for the study, such as the density of shade cloth for a field experiment.

Complicating matters further are the many states in which these data exist and the many places in which they are stored. The states of data can be grouped into raw data, processed data, verified data, certified data (such as water quality data that meets government standards), models, and software and algorithms (which often are required to interpret the data). Digital data in any of these forms may reside on the computers of the investigators and students who collected them, on laboratory servers, or on shared servers. Data typically are stored in multiple places, in multiple versions. Data in the form of printouts or field notes are stored in desks or file cabinets. Data in the form of specimens or samples are stored in refrigerators or freezers.

Our interviewees were much more concerned with publications as the end product of their research than with the sensor data per se. Several scientists explained how they design a field experiment with a particular story in mind, and how the story determines roughly the amount of data needed. One of our subjects in the ethnographic interviews was very explicit, telling us that "to tell that story I'm going to need an average of five figures and a table." Thus the amount of data required for a study is a consideration in the design of tools and services.

Data Sharing

Of particular value from these studies are insights about what data scientists will share, when, with whom, and by what criteria. These scientists are most willing to share data that already have been published and least willing to share data that they plan to publish, as these data represent claims for their research. This basic result confirms lessons from the social studies of science (Latour, 1987; Latour & Woolgar, 1986). More interesting are the specific examples of criteria for sharing and their implications for digital libraries and e-Science.

Scientists, computer scientists, and engineers alike all were forthcoming with long lists of variables being collected, such as those listed in Figure 2. When asked about which data they would be willing to share, all referred to the scientific variables as being the data of interest to others. The computer science and engineering teams did not appear to view the performance data or proprioceptive data as being of much interest or value to anyone else.

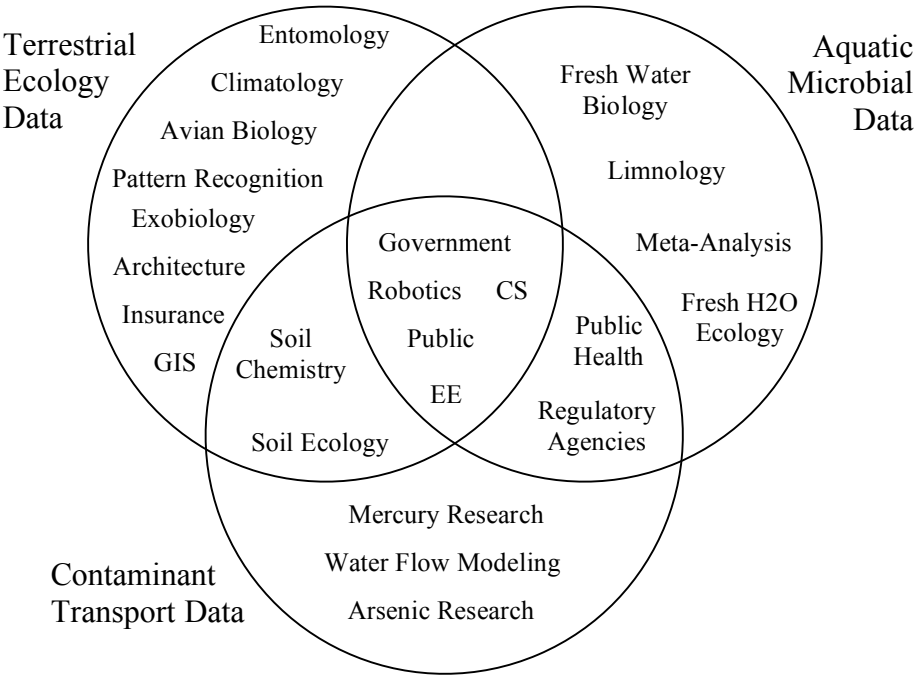
Our subjects varied widely in what data they were willing to release, to whom, and under what conditions. We found all combinations of distinctions by state of the data (e.g., raw, processed, certified), requestor conditions (to anyone, no restrictions; if no commercial reuse, share and share alike; to anyone, provided source is acknowledged or cited; if co-authorship credit given for providing the data; if research questions do not compete with ours), and temporal conditions (release after articles are published; after we've finished mining the data; after a certain time period, e.g., 3-5 yrs; or "it depends").

In cases where subjects made a distinction between experimental and contextual data, they were more willing to release contextual data. These data can be essential when comparing results of multiple studies at a research site. For example, the James Reserve, which is part of the University of California Natural Reserve System and a partner in CENS, makes available a wide array of data collected on-site (*James San Jacinto Mountains Reserve*, 2004).

One of our ethnographic subjects explained that his willingness to share is influenced by the effort required to collect the data. His hand-collected data are more "hard won" than sensor-generated data, and he is less likely to share them. This finding is consonant with that of Pritchard, et al. (2004) who found that data sharing increased with the degree of automation in all disciplines studied.

We also asked subjects to give examples of when they had acquired data from others, based on Zimmerman's (2003) findings that those most likely to borrow were most likely to share. Few of our subjects reported experience in borrowing data from others. Several respondents said they were not sharing data because their data was not of value to others. However, many of them were able to identify current and potential future audiences for data resulting from their research (Figure 3).

Figure 3: Diversity of Current and Potential Users of CENS Data



Data Policy

The lack of sharing data in small science fields is due at least partly to lack of incentives, as discussed above. Digital libraries to manage e-Science data must reflect appropriate policies about who has access to what data, in what form, and under what conditions. To learn more about these issues, we asked questions about data authorship, about who has the authority to share data, about conditions for sharing, and about what constitutes publishable results.

Questions about who was the "author" of a dataset for purposes of publication or data release were of limited utility. Our subjects acknowledged that teams were still struggling with notions of authorship for papers resulting from multi-disciplinary research efforts, and had not dealt directly with the notion of "data authorship." Asking who owned the data was a more fruitful question, despite the ambiguities of legal ownership of data resulting from government-funded research (most of CENS research is funded by the National Science Foundation). Responses to that question can be grouped into

four categories (the funding or supporting institution; the principal investigator; anyone with any intellectual contribution; don't know/haven't considered), with the last two being the most common. The lack of clarity in who "owns" or has authority to release data is problematic for the design of a digital library to support such data, and an issue we will clarify in more detail in the next rounds of data analysis and interviews.

The wide range of opinions about what data would be released, under what conditions, when, and to whom (listed above under data sharing) also complicates data policy. While CENS has a general commitment to sharing the data from its research, the default setting on access obviously cannot be that everything is public, immediately. Rather, CENS' digital libraries of data will need to provide multiple levels of access controls that can be set by principal investigators or by whomever a project designates as having proper authority. We also found that quality of data being released is considered to be an ethical issue. In the large group meeting about ethics and policy of data sharing, responsibility for data quality was a central issue. Concerns also arose about whether any sort of liability disclaimers or rights claims (e.g., Creative Commons licenses for attribution and non-commercial reuse) should be applied. Some commented that if they feel they are forced to share data they will, knowing that raw data may not be of much use to others.

In one of our ethnographic interviews, our questions about data policy elicited an enlightening scenario that contrasts scientific and engineering views about data use policy. At issue was whether data from an instrument belonged to the designers of the instrument or to the designers of the experiment. Although the instrument in question was designed and installed by a member of an engineering faculty, that investigator did not analyze or publish the resulting data. After several years of data production, which were being posted on a public website, one of the biological scientists sought permission to use the data. Authority for release of that data did rest explicitly with the director of the field site where the instrument was installed. The director granted permission to the biological scientist, who then invested effort in cleaning and analyzing the data for his own research questions. When the results appeared promising for publication, the scientist and site director invited the engineering professor to participate in the publication. However, the engineering professor objected to the use of those data for biological research on the grounds that they were his data because he had deployed the instrument. The situation later was resolved in an ad hoc way without making general policy, and the resulting data were published.

Data Architecture

Our analyses to date suggest several lessons about tool requirements at each stage in the life cycle. The challenge is to integrate the lessons from data characteristics, sharing, and policy into data architecture. At the initial stages of research design, habitat ecologists indicated a need for tools to guide sensor placement. Maps of research sites that include the location of sensors and the types of data that each sensor could produce would be helpful. Once in the field, they need to test and calibrate sensors and monitor the data those sensors produce. Scientists and engineers all expressed a desire for better tools to assess individual sensors and overall "network health." They want tools in the field to verify the validity and reliability of the data stream. They mentioned the desire for tools to help them identify when values are duplicated or missing, when sensors are failing, and other anomalous situations. They want to annotate the data in the field to provide important context that cannot be anticipated in data models. Habitat ecologists often modify the instruments or field conditions on-site. They may change the location of equipment during an experiment in response to field conditions, for example. Documenting which data were collected at which location, when, and with what instruments is essential to later interpretation.

Many different types of sensors are used concurrently, each generating different variables. Sensors change from deployment to deployment as technology improves and as research questions change. Some of these sensors are off-the-shelf commercial technology and others are developed by the research teams or are prototypes. The data produced and the form in which they are structured varies widely. Scientists need tools to reconcile these many data formats. They often need to reconcile data

from multiple sensors using external variables such as time stamps. Sometimes several people have to handle data before they can be analyzed by the investigator, severely hampering the ability to make real-time adjustments in research deployments. The need to capture data as cleanly as possible was mentioned frequently. Capturing data in a single consistent format is unlikely, given how quickly the technology and the research questions change. However, improvements can be made in consistency of data capture, and data can be mapped forward into common data structures such as SensorML and Ecological Metadata Language, mentioned earlier.

The scientists also want tools for data analysis. They clean their data with respect to their research questions, thus data that are extraneous to a given study may be stored, but not cleaned or analyzed. Data files and analysis tools proliferate, as each individual creates data files, analyses, and transformations, using preferred tools such as Excel spreadsheets or Matlab. Practices are not standardized either within or between research teams. Again, mapping the sensor data and other data from field deployments into common formats would aid in reconciling data from individuals, groups, and deployments.

Conclusions

Habitat ecology and closely related areas of environmental sciences research are in a state of transition from "small science," characterized by hand-crafted data collection, to "big science," with instrumented data collection, larger volumes of data, and distributed, multi-disciplinary research teams. The scientific value of technologies such as sensor networks is recognized for the potential to ask new questions, in new ways, and to get results more rapidly. The value of the resulting data for longitudinal and comparative research also is recognized widely. This is a new way of "doing science," which requires new kinds of practices. These practices are emerging as scientists, computer scientists, and engineers gain experience in working together and in learning what each can bring to the collaboration.

The present stage of development reflects a paradoxical situation. Researchers are generally in favor of sharing the data from their research, but they do not agree on what those data are, on the conditions under which data should be released, nor who has the authority to release them. Their differences in opinion represent valid differences in practices between participating disciplines. Their reluctance to release data immediately also reflects valid concerns about ethics, about misuse of their data, and about others "scooping" their research findings.

Also paradoxical is researchers' awareness of extant metadata standards for reconciling, managing, and sharing their data, but their lack of use of such standards. The present dilemma is that few of the participating scientists see a need or ability to use others' data, so they do not request data, they have no need to share their own data, they have no need for data standards, and no standardized data are available. Our working hypothesis is that digital libraries and associated tools that are designed to reflect the practices of the participating communities may provide some solutions to this dilemma.

We have identified several design factors to expedite the development of digital libraries for the habitat biology and smaller scale environment sciences communities with which we work. Tools and services can be used to map current datasets and structures into common metadata formats. We are exploring how much of the older data can be mapped with metadata "crosswalks" (Godby, Young & Childress, 2004). New data can be mapped by capturing it initially in forms as close to these standard structures as possible. Once mapped into common formats, data will be more amenable to common tools and to aggregation.

Policies that establish shared ownership of the data also will encourage the contribution and sharing of data. Metadata can identify who contributed to the production and analysis of each dataset, for example. This may mean some sort of public trace of the development and history of the data. Access and release rights could be based on responsibility for data creation. Rich descriptions of data will assist in their being discovered by people with new research questions. Authority for releasing data can be built

into the system in several ways: Individuals can authorize release of datasets. Publications of articles using a dataset could trigger data release. Embargo periods could trigger release (e.g., all data will be released 2 years after it enters the digital library unless specific action is taken to prevent release). Data can be labeled for degree of verification or certification. Liability disclaimers and attribution requirements can be incorporated into conditions for release.

e-Science offers great promise for improving access to scientific data via distributed digital libraries. "Little science" areas that focus on local research problems, collect data in response to local conditions, and work in small teams are not yet well served by e-Science technologies. The data they produce is irreplaceable and has tremendous potential for longitudinal and comparative research on scientific problems of global significance. However, not enough is known about how scientists in these areas produce, use, or manage data, or how data management practices vary between these scientists and their partners in computer science and engineering. Studies of data practices in newly instrumented research areas will contribute to understanding more about how e-Science can benefit little science. The Center for Embedded Networked Sensing offers a microcosm in which to address these issues. If e-Science infrastructure can facilitate understanding of earth's rapidly changing ecosystem, the investments will be deemed well spent by society.

Acknowledgements

CENS is funded by National Science Foundation Cooperative Agreement #CCR-0120778, Deborah L. Estrin, UCLA, Principal Investigator. CENSEI, under which much of this research was conducted, is funded by National Science Foundation grant #ESI-0352572, William A. Sandoval, Principal Investigator and Christine L. Borgman, co-Principal Investigator. Jonathan Furner and Stasa Milojevic of the Department of Information Studies provided valuable material to the fieldwork on which this study is based. Robin Conley, Jeffrey Good, and Matthew Mayernik, UCLA graduate students, assisted in interviewing and data transcription.

References

Akyildiz, I. F., Su, W., Sankarasubramaniam, Y. & Cayirci, E. (2002). Wireless sensor networks: A survey. *Computer Networks,* 38: 393-422.

Allaby, M. (Ed.). (2006). A Dictionary of Ecology (3rd ed.). Oxford, U.K.: Oxford University Press.

ArXiv.org e-Print archive. (2006). Visited http://arxiv.org/ on 27 April 2006.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. & Wouters, P. (2004). An International Framework to Promote Access to Data. *Science,* 303(5665): 1777-1778.

Bekaert, J. & Van de Sompel, H. (2006). Access Interfaces for Open Archival Information Systems based on the OAI-PMH and the OpenURL Framework for Context-Sensitive Services. *PV 2005: Ensuring Long-term Preservation and Adding Value to Scientific and Technical data*, The Royal Society, Edinburgh Visited http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/032.pdf on 28 September 2006.

Bishop, A. P., Van House, N. & Buttenfield, B. P. (Eds.). (2003). Digital library use: Social practice in design and evaluation. Cambridge, MA: MIT Press.

Borgman, C. L. (2004). The Interaction of Community and Individual Practices in the Design of a Digital Library. *International Symposium on Digital Libraries and Knowledge Communities in Networked Information Society*, University of Tsukuba, Tsukuba, Ibaraki, Japan., University of Tsukuba Visited http://www.kc.tsukuba.ac.jp/dlkc/e-proceedings/papers/dlkc04pp9.pdf on 10 April 2006.

Borgman, C. L. (2006). What can studies of e-Learning teach us about e-Research? Some findings from

digital library research. *Journal of Computer Supported Cooperative Work,* 15(4): 359-383.

Borgman, C. L. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Borgman, C. L., Wallis, J. C. & Enyedy, N. (2006). Building digital libraries for scientific data: An exploratory study of data practices in habitat ecology. *10th European Conference on Digital Libraries*, Alicante, Spain, Springer.

Borgman, C. L., Bates, M. J., Cloonan, M. V., Efthimiadis, E. N., Gilliland-Swetland, A. J., Kafai, Y., Leazer, G. L. & Maddox, A. (1996). *Social Aspects of Digital Libraries. Final Report to the National Science Foundation; Computer, Information Science, and Engineering Directorate; Division of Information, Robotics, and Intelligent Systems; Information Technology and Organizations Program*. Visited http://is.gseis.ucla.edu/research/dl/index.html on 28 September 2006.

Botts, M. (2004). *Sensor Model Language (SensorML) for in-situ and remote sensors: Version 1.0.0 beta. Recommended paper, no. OGC 04-019r2*.Open Geospatial Consortium.

Botts, M. (2006). *Sensor Modelling Language (SensorML) Status*. Visited http://stromboli.nsstc.uah.edu/SensorML/status.html on Nov 20, 2006.

Botts, M. & McKee, L. (2003). A Sensor Model Language: Moving sensor data onto the Internet. *Sensors* 20(Issue). Visited http://www.sensorsmag.com/articles/0403/30/main.shtml

Bowker, G. C. (2000a). Mapping biodiversity. *International Journal of Geographical Information Science,* 14(8): 739-754.

Bowker, G. C. (2000b). Biodiversity datadiversity. *Social Studies of Science,* 30(5): 643-683.

Bowker, G. C. (2000c). Work and information practices in the sciences of biodiversity. *VLDB 2000, Proceedings of 26th international conference on very large data bases*, Cairo, Egypt, Kaufmann.

Bowker, G. C. (2005). *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

Chudnov, D., Cameron, R., Frumkin, J., Singer, R. & Yee, R. (2005). Opening up openURLs with autodiscovery. *Ariadne,*(43). Retrieved from http://www.ariadne.ac.uk/issue43/chudnov/ on 29 September 2006.

Collaborative Large-Scale Engineering Analysis Network for Environmental Research. (2006). Visited http://cleaner.ncsa.uiuc.edu/home/ on 16 August 2006.

Consortium of Universities for Advancement of Hydrologic Science. (2006). Visited http://www.cuahsi.org on 15 November 2006.

Culler, D. E. & Hong, W. (2004). Wireless sensor networks. *Communications of the ACM,* 47(6): 30-33.

Cummings, J. N. & Kiesler, S. (2005). Collaborative research across disciplinary and organizational boundaries. *Social Studies of Science,* 35(5): 703-722. Retrieved from <Go to ISI>://000232598300003

David, P. A. & Spence, M. (2003). *Towards Institutional Infrastructures for E-Science: The Scope of the Challenge*. Oxford Internet Institute Research Reports: University of Oxford. Visited http://129.3.20.41/eps/le/papers/0502/0502002.pdf on 30 September 2006.

Document Action: 'The "info" URI Scheme for Information Assets with Identifiers in Public Namespaces' to Informational RFC. (2005). Internet Engineering Task Force. Visited http://www1.ietf.org/mail-archive/web/ietf-announce/current/msg01746.html on 8 March 2006.

Ecological Metadata Language. (2004). Visited http://knb.ecoinformatics.org/software/eml/ on 25 November 2004.

Elson, J. & Estrin, D. (2004). Sensor networks: A bridge to the physical world. In C. S. Raghavendra, K. M. Sivalingam & T. F. Znati (Eds.). *Wireless sensor networks*. Boston, Kluwer Academic.

Engeström, Y. (1987). *Learning by Expanding: An activity-theoretical approach to developmental research*. Helsinki: Orienta-Konsultit.

Engeström, Y., Miettinen, R. & Punamaeki, R.-L. (Eds.). (1999). Perspectives on activity theory. New York, NY, US: Cambridge University Press.

Esanu, J. M. & Uhlir, P. F. (Eds.). (2004). Open Access and the Public Domain in Digital Data and Information for Science: Proceedings of an International Symposium. Washington, DC: The National Academies Press. Visited http://books.nap.edu/catalog/11030.html on 30 September 2006.

Estrin, D., Michener, W. K. & Bonito, G. (2003). *Environmental cyberinfrastructure needs for distributed sensor networks: A report from a National Science Foundation sponsored workshop*. Scripps Institute of Oceanography. Visited http://www.lternet.edu/sensor_report/ on 12 May 2006.

Finholt, T. A. (2003). Collaboratories as a new form of scientific organization. *Economics of Innovation and New Technology,* 12(January).

Ginsparg, P. (2001). Creating a global knowledge network. *Second Joint ICSU Press - UNESCO Expert Conference on Electronic Publishing in Science*, Paris, UNESCO Visited http://people.ccmr.cornell.edu/~ginsparg/blurb/pg01unesco.html on 12 May 2006.

Glaser, B. G. & Strauss, A. L. (1967). *The discovery of grounded theory; strategies for qualitative research*. Chicago,: Aldine Pub. Co.

Global Earth Observation System of Systems. (2006). Visited http://www.epa.gov/geoss/ on 30 April 2006.

Global Ocean Observing System. (2006). Visited http://www.ioc-goos.org/ on 5 June 2006.

Godby, C. J., Young, J. A. & Childress, E. (2004). A repository of metadata crosswalks. *D-Lib Magazine,* 10(12). Retrieved from http://www.dlib.org/dlib/december04/godby/12godby.html on 22 May 2006.

Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A., DeWitt, D. & Heber, G. (2005). Scientific data management in the coming decade. *CT Watch Quarterly,* 1(1). Retrieved from http://www.ctwatch.org/quarterly/articles/2005/02/scientific-data-management/ on 25 August 2006.

Hey, T. & Trefethen, A. (2003). The Data Deluge: An e-Science Perspective. *Grid Computing – Making the Global Infrastructure a Reality*, Wiley. Visited http://www.rcuk.ac.uk/escience/documents/report_datadeluge.pdf on 20 January 2005.

Hey, T. & Trefethen, A. (2005). Cyberinfrastructure and e-Science. *Science,* 308: 818-821.

Hilgartner, S. & Brandt-Rauf, S. I. (1994). Data access, ownership and control: Toward empirical studies of access practices. *Knowledge,* 15: 355-372.

Hodge, G. & Frangakis, E. (2005). *Digital Preservation and Permanent Access to Scientific Information: The State of the Practice*. The International Council for Scientific and Technical Information (ICSTI) and CENDI: U.S. Federal Information Managers Group. Visited http://cendi.dtic.mil/publications/04-3dig_preserv.html on 30 September 2006.

Integrated Ocean Observing System. (2006). Visited http://www.ocean.us/what_is_ioos on 5 June 2006.

International Virtual Observatory Alliance. (2006). Visited http://www.ivoa.net/ on 30 September 2006.

James San Jacinto Mountains Reserve. (2004). University of California. Visited http://www.jamesreserve.edu/ on 30 November 2004.

Kwa, C. (2005). Local ecologies and global science: Discourses and strategies of the International Geosphere-Biosphere Programme. *Social Studies of Science,* 35(6): 923-950.

Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, MA: Harvard University Press.

Latour, B. & Woolgar, S. (1986). *Laboratory Life: The Construction of Scientific Facts*, (2nd ed.). Princeton, N.J.: Princeton University Press.

Lofland, J., Snow, D., Anderson, L. & Lofland, L. H. (2006). *Analyzing Social Settings: A Guide to Qualitative Observation and Analysis*. Belmont, CA: Wadsworth/Thomson Learning.

Long-Lived Digital Data Collections: Enabling Research and Education for the 21st Century. (2005). National Science Board. Visited http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf on 1 October 2006.

Lord, P. & Macdonald, A. (2003). *E-Science Curation Report--Data Curation for E-science in the UK: An Audit to Establish Requirements for Future Curation and Provision*. JISC Committee for the Support of Research. Visited http://www.jisc.ac.uk/uploaded_documents/e-scienceReportFinal.pdf on 1 October 2006.

Michener, W. K. & Brunt, J. W. (Eds.). (2000). Ecological Data: Design, Management and Processing. Oxford: Blackwell Science.

National Ecological Observatory Network. (2006). Visited http://neoninc.org/ on 3 October 2006.

National Office for Integrated and Sustained Ocean Observations. (2006). Visited http://www.ocean.us/

on 5 June 2006.
Object Reuse and Exchange. (2006). Visited http://www.openarchives.org/ore/ on 15 November 2006.
Olson, G. M. (2005). Long distance collaborations in science: Challenges and opportunities. *First International Conference on e-Social Science*, Manchester, UK, National Center for e-Social Science.
Olson, G. M. & Olson, J. S. (2000). Distance matters. *Human-Computer Interaction,* 15(2-3): 139-178.
Pottie, G. J. & Kaiser, W. J. (2000). Wireless integrated network sensors. *Communications of the ACM,* 43(5): 51-58.
Pottie, G. J. & Kaiser, W. J. (2006). *Principles of embedded networked systems design*. Cambridge, England: Cambridge University Press.
Price, D. J. d. S. (1963). *Little Science, Big Science*. New York: Columbia University Press
Pritchard, S. M., Carver, L. & Anand, S. (2004). *Collaboration for knowledge management and campus informatics*. University of California, Santa Barbara. Visited http://www.library.ucsb.edu/informatics/informatics/documents/UCSB_Campus_Informatics_Project_Report.pdf on 5 July 2006.
Reference Model for an Open Archival Information System. (2002). Recommendation for Space Data System Standards, Consultative Committee for Space Data Systems Secretariat, Program Integration Division (Code M-3), National Aeronautics and Space Administration. Visited http://public.ccsds.org/publications/archive/650x0b1.pdf on 4 October 2006.
Sandoval, W. A. & Reiser, B. J. (2003). Explanation-driven inquiry: integrating conceptual and epistemic supports for science inquiry. *Science Education,* 87: 1-29.
Schnase, J. L., Lane, M. A., Bowker, G. C., Star, S. L. & Silberschatz, A. (1997). Building the next generation biological information infrastructure. In P. H. Raven & T. Williams (Eds.). *Nature and Human Society: The Quest for a Sustainable World*. Washington, DC, National Academy Press**:** 291-300. Visited http://darwin.nap.edu/books/0309065550/html on 4 October 2006.
Sensor Modeling Language. (2005). Visited http://vast.uah.edu/SensorML/ on 16 January 2006.
Shankar, K. (2003). *Scientific data archiving: the state of the art in information, data, and metadata management*. Visited http://cens.ucla.edu/Education/index.html on 19 January 2005.
Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility (2003). *Meeting organized by the Wellcome Trust*, Fort Lauderdale, Florida, Wellcome Trust Visited http://www.wellcome.ac.uk/assets/wtd003207.pdf on 25 July 2005.
Sonnenwald, D. H. (2007). Scientific collaboration: Challenges and solutions. In B. Cronin (Ed.). *Annual Review of Information Science and Technology*. Medford, NJ, Information Today. 41**:** 643-682.
Southern California Coastal Ocean Observing System. (2006). Visited http://www.sccoos.org/ on 5 June 2006.
Thadani, V., Cook, M., Millwood, K., Harven, A., Fields, D., Griffis, K., Wise, J., Kim, K. & Sandoval, W. A. (2006). *Eyes on the prize: considering how design research can lead to sustainable innovation*. Paper presented at the Annual Meeting of the American Educational Research Assn., San Francisco, April 7-12.
Traweek, S. (1992). *Beamtimes and Lifetimes: The World of High Energy Physicists*, (1st Harvard University Press pbk. ed.). Cambridge, Mass.: Harvard University Press.
Traweek, S. (2004). Generating high energy physics in Japan. In D. Kaiser (Ed.). *Pedagogy and Practice in Physics*. Chicago, University of Chicago Press.
U.S. Long Term Ecological Research Network. (2006). Visited http://lternet.edu/ on 5 June 2006.
Van de Sompel, H., Nelson, M. L., Lagoze, C. & Warner, S. (2004). Resource harvesting within the OAI-PMH framework. *D-Lib Magazine,* 10(12). Retrieved from http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html on 5 October 2006.
Van de Sompel, H., Hammond, T., Neylon, E. & Weibel, S. L. (2006). *RFC 4452: The "info" URI Scheme for Information Assets with Identifiers in Public Namespaces*. Requests for Comments, Internet Engineering Task Force. Visited http://www.rfc-archive.org/getrfc.php?rfc=4452 on 5 October 2006.
Van House, N. A. (2003). Digital libraries and collaborative knowledge construction. In A. P. Bishop, N.

Van House & B. P. Buttenfield (Eds.). *Digital Library Use: Social Practice in Design and Evaluation*. Cambridge, MA, MIT Press**:** 271-296.

Wallis, J. C., Milojevic, S., Borgman, C. L. & Sandoval, W. A. (2006). The special case of scientific data sharing with education. *American Society for Information Science & Technology*, Austin, TX, Information Today.

Weinberg, A. M. (1961). Impact of large-scale science on the United States. *Science,* 134(3473): 161-164.

Zimmerman, A. (in press-a). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse *International Journal of Digital Libraries*.

Zimmerman, A. & Nardi, B. (2006). Whither or whether HCI: Requirements analysis for multi-sited, multi-user cyberinfrastructures. *CHI 2006*, Montreal, Association for Computing Machinery.

Zimmerman, A. S. (2003). *Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists* School of Information. University of Michigan. Ann Arbor, MI.

Zimmerman, A. S. (in press-b). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology, & Human Values*.