Building Digital Libraries for Scientific Data: An Exploratory Study of Data Practices in Habitat Ecology

Christine Borgman[1], Jillian C. Wallis[2], and Noel Enyedy[3]

[1] Department of Information Studies
Graduate School of Education & Information Studies, UCLA
borgman@gseis.ucla.edu
[2] Center for Embedded Networked Sensing, UCLA
jwallisi@ucla.edu
[3] Department of Education
Graduate School of Education & Information Studies, UCLA
enyedy@gseis.ucla.edu

Abstract. As data become scientific capital, digital libraries of data become more valuable. To build good tools and services, it is necessary to understand scientists' data practices. We report on an exploratory study of habitat ecologists and other participants in the Center for Embedded Networked Sensing. These scientists are more willing to share data already published than data that they plan to publish, and are more willing to share data from instruments than hand-collected data. Policy issues include responsibility to provide clean and reliable data, concerns for liability and misappropriation of data, ways to handle sensitive data about human subjects arising from technical studies, control of data, and rights of authorship. We address the implications of these findings for tools and architecture in support of digital data libraries.

1 Introduction

The emerging cyberinfrastructure is intended to facilitate distributed, information-intensive, data-intensive, collaborative research [1]. Digital libraries are essential to the cyberinfrastructure effort. As scholarship in all fields becomes more data-intensive and collaborative, the ability to share data becomes ever more essential [2, 3]. Data increasingly are seen as research products in themselves, and as valuable forms of scientific capital [4]. "Big science" fields such as physics, chemistry, and seismology already are experiencing the "data deluge" [5, 6]. Data repositories and associated standards exist for many of these fields, including astronomy, geosciences, seismology, and bioinformatics [7-10]. "Little science" fields such as habitat ecology are facing an impending data deluge as they deploy instrumented methods such as sensor networks. Progress toward repositories and information standards for these fields is much less mature, and the need is becoming urgent.

2 Research Domain

We have a unique opportunity to study scientific data practices and to construct digital library architecture to support the use and reuse of research data. The Center

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 170 – 183, 2006.

for Embedded Networked Sensing (CENS), a National Science Foundation Science and Technology Center based at UCLA, conducts collaborative research among scientists, technologists, and educators. CENS' 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' research crosses four primary scientific areas: habitat ecology, marine microbiology, seismology, and environmental contaminant transport, plus applications in urban settings and in the arts. The research reported here addresses the use of embedded networked sensor technology in biocomplexity and habitat monitoring, supplemented by findings about data sharing from other CENS' areas. In these scientific areas, the goals are to develop robust technologies that will operate in uncontrolled natural settings and in agricultural settings. The science is 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. Thus we have moved toward more "human in the loop" models where investigators can adjust monitoring conditions in real time.

3 Background

3.1 Data Digital Libraries and the Data Deluge

Science is a technical practice and a social practice [11]. It is the interaction between technological and social aspects of scientific research that underlies the design challenge. Modern science is distinguished by the extent to which its products rely on the generation, dissemination, and analysis of data. These practices are themselves distinguished by their massive scale and global dispersion. New technologies for data collection are leading to data production at rates that exceed scientists' abilities to analyze, interpret, and draw conclusions. No respite from this data deluge is foreseen; rather, the rate at which data are generated is expected to increase with the advancement of instrumentation [5]. Consequently, scientists urgently require assistance to identify and select data for current use and to preserve and curate data over the long term. Data resources are dispersed globally, due to more international collaboration and distributed access to computing resources for analyzing data. Cyberinfrastructure is expected to provide capabilities to (i) generate scientific data in consistent formats that can be managed with appropriate tools; (ii) identify and extract—from vast, globally distributed repositories—those data that are relevant to their particular projects; (iii) analyze those data using globally distributed computational resources; (iv) generate and disseminate visualizations of the results of such analyses; and (v) preserve and curate data for future reuse. An effective cyberinfrastructure is one that provides distributed communities-- scientific and nonscientific--with persistent access to distributed data and software routinely, transparently, securely, and permanently.

3.2 Data Management Practices

The willingness to use the data of others may be a predictor of willingness to share one's own data. Scholars in fields that replicate experiments or that draw heavily on observational data (e.g., meteorological, astronomical records) appear more likely to contribute data for mutual benefit within their fields. Conversely, scholars in many fields work only with data they have produced. The graph or table that results from analyzing the data may be the essential product of a study. Many scholars assume that the underlying data are not of value beyond that study or that 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 studies where data types and variables may vary widely by study [12, 13].

The degree of instrumentation of data collection also appears to be a factor in data sharing. Sharing expensive equipment is among the main drivers for collaboration, especially in fields such as physics and chemistry. In these cases, collaboration, instrumentation, and data sharing are likely to be correlated. 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 [14].

3.3 Habitat Ecology Data and Practices

The study of biodiversity and ecosystems is a complex and interdisciplinary domain [15]. The mechanisms used to collect and store biological data are almost as varied as the natural world those data document. Over the last thirty years, data management systems for ecological research have evolved out of large projects such as the *International Biological Program* (IBP; established in 1964 by the International Council of Scientific Unions), the *Man and the Biosphere* program (MAB; established in 1971 by the United Nations), and the U.S. *Long-Term Ecological Research* program (LTER; established in 1980 by the National Science Foundation) [16]. These systems need to support multiple data types (numerical measurements, text, images, sound and video), and to interact with other systems that manage geographical, meteorological, geological, chemical, and physical data. Currently one of the biggest challenges to the development of effective data management systems in habitat ecology is the "datadiversity" that accompanies biodiversity [17].

The Knowledge Network for Biocomplexity (KNB) [http://knb.ecoinformatics.org/], an NSF-funded project whose first products became available in 2001, is a significant development for data management in habitat ecology. KNB tools include a data management system for ecologists, based on the Morpho client software and Metacat server software, and a standard format for the documentation of ecological data—the Ecological Metadata Language (EML). SensorML is an equally important development for sensor data [18].

4 Research Problem

While practices associated with scholarly publication vary widely between fields, the resulting journal articles, papers, reports, and books can be described consistently with bibliographic metadata. Data are far more problematic. Disciplines vary not only in their choice of research methods and instruments, but the data gathered may vary in form and structure by individual scholar and by individual experiment or study. Multidisciplinary collaboration, which is among the great promises of cyberinfrastructure, will depend heavily on the ability to share data within and between fields. However, very little is yet known about practices associated with the collection, management, and sharing of data. Despite these limitations, immediate needs exist to construct systems to capture and manage scientific data for local and shared use. These systems need to be based on an understanding of the practices and requirements of scientists if they are to be useful and to be used.

Habitat ecology is a “small science,” characterized by small research teams and local projects. Aggregating research results from multiple projects and multiple sites has the potential to advance the environmental sciences significantly. 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 [19]. Thus one of our research concerns is how habitat ecology may evolve with the use of embedded networked sensing. These scientists are deploying dense sensor networks in field locations to pursue research on topics such as plant growth, bird behavior, and micrometeorological variations.

Our 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. The questions can be categorized as follows:

- **Data characteristics:** What data are being generated? To whom are these 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?
- **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?

5 Research Method

5.1 Data Sources

Our goal is to understand data practices and functional requirements for CENS ecology and environmental engineering researchers with respect to architecture and policy, and to identify where architecture meets policy. The results reported here are drawn from multiple sources over a three-year period (2002-2005). In the first year (2002-2003), we sat in on team meetings across CENS scientific activities and we inventoried data standards for each area [20]. In year 2 (2003-4), 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 year 3 (2004-05). In the current year (2005-6), we are interviewing individual members of habitat ecology research teams, including scientists, their technology research partners in computer science and engineering, and graduate students, postdoctoral fellows, and research staff.

5.2 Process

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 [20]. We did not audiotape or videotape these meetings to avoid interfering with the local activities. Knowledge from this part of the research was used to identify data standards relevant to the research areas. We shared our results with individuals and teams to get feedback on the relevance of these standards. We also constructed prototypes of data analysis and management tools as components of the educational aspects of our research [21]. Thus we are conducting iterative research, design, and development for data management tools in CENS.

5.3 Participants

Our population at CENS is comprised of about 70 faculty and other researchers, a varying number of post-doctoral researchers, and many student researchers. About 50 scientists, computer scientists, engineering faculty, and their graduate students, post-doctoral fellows, and research staff are working in the area of habitat ecology. The data reported here are drawn primarily from in-depth interviews of two participants, each two to three hours over two to three sessions. The direct quotes are from these interviews. Results from one-hour interviews with two other scientists and from a large group meeting (about 20 people) to discuss data sharing policy also are reported here. These results are informed by interviews, team meetings, and other background research conducted in earlier stages of our data management studies.

5.4 Analysis

We used the results of interviews and documentary analyses in the first two years of our research to design the ethnographic study. This study used the methods of grounded theory [22]. The interview questions are based on Activity Theory [23-25], which analyzes communities and their evolution as “activity systems.” Activity systems are defined by the shared purposes that organize a community and by the ways in which joint activity to achieve these purposes is mediated by shared tools, rules for behavior, and divisions of labor. When analyzing how activity systems change and develop, the focus is on contradictions that occur within the system or as a result of the system interacting with other activity systems. These contradictions are analyzed as the engine for organizational change.

Based on this theoretical framework, we developed interview questions about participants' motives, understanding of their community's motives, tools used in daily work, ways they divided labor, power relations within their community, and rules and norms for the community. Interviews were then fully transcribed. In the initial phase of analysis we looked at the first interviews with participants for emergent themes. The analysis progressed iteratively. Subsequent interviews were analyzed with an eye towards testing and further refining the themes identified in the initial coding. With each refinement, the remaining corpus was searched for confirming or contradictory evidence. At this stage, however, the work is still preliminary. As such, no formal coding schemes were developed that were systematically applied to the entire corpus. Rather, what we present below are the emergent themes and representative illustrations in the participants' own words.

6 Results

In the first two years of study, we learned that CENS' habitat biologists perceived a lack of established standards for managing sensor data, specifically those that support the sharing of data among colleagues. They are eager to work with us because they need tools to capture, manage, and share sensor data more effectively, efficiently, and easily. They are committed to participation of developing domain data repositories and standards such as the Knowledge Network for Biocomplexity and Ecological Metadata Language, but are not yet implementing them. A good starting point for exploring the metadata requirements of this community is to assess scientists' experiences with implementing these tools and standards, and to evaluate those tools' utility. The results are organized by the research questions outlined above: Data characteristics, data sharing, data policy, and data architecture.

6.1 Data Characteristics

Our interview questions explored what data are being generated, to whom are these data, and to whom are they useful. CENS is a collaboration between technologists and scientists, thus the technologies are being evaluated and field tested concurrently with the scientific data collection. The scientists interviewed reported unreliability of the sensors. In the early stages, one researcher found that he was losing about 25% of every transmission from every sensor, for example. While the sensors were sending data every five minutes, they only produced usable data every 10 or 15 minutes. Thus they could not always trust what they were getting from the instruments:

> *I'm highly suspicious [of automated data collection]. I mean it works, but then sometimes it doesn't, and then sometimes weird things happen. You get a glitch, and then you start getting the same value twice or something. … when you do averages, it's all funny.*

These researchers have a story in mind when they design a field experiment. They also know about how much data they will need to support that story in a published paper in their discipline. One of our scientists explicitly told us that

> *… to tell that story I'm going to need an average of five figures and a table.*

He sketches out mock figures on paper as part of his research design. We were particularly intrigued by his notion of "least publishable unit" – in this case, five figures and a table. However, he also told us that he tends to collect more data so that he can elaborate on his story:

> *I collect another data set just to round it out rather than [picking] an interesting sub-phenomenon of another sub-phenomenon. That's boring.*

Thus his research design is based on the amount of data he needs to tell a story of this scope and form. The publication is the product of the study, rather than the data per se.

6.2 Data Sharing

We collected some useful commentary on issues of what data scientists will share, when, with whom, and with what criteria. Within this small sample, scientists generally are more willing to share data that already have been published and less willing to share data that they plan to publish. The latter type of data represent claims for their research.

> *Sharing data- if it's already published? It's your data, no problem. I can give that to [you]. If it's something I'm working on to get published or somebody else is working on to get published, or if they want to publish the paper together, it gets a little bit funnier.*

For this scientist, willingness to share also is influenced by the effort required to collect the data. His hand-collected data are more precious than his instrument-generated data:

> *… if you walk out into a swamp .. out in this wacky eel grass, and marsh along with your hip-waders and [are] attacked by alligators ..and then you do it again and again and again... I don't [want to] share that right away. I [want to] analyze it because I feel like it's mine.*

The above dataset was seen as "hard won." When they do share experimental data with collaborators, they feel an ethical and scientific responsibility to clean the dataset sufficiently that it has scientific value. Raw data is meaningless to others. Unless the data are useful and relevant, they would be "just taking up space and nobody's going to be able to use [them]."

If they feel they are forced to share data they will, knowing that it may not be of much use to others. One scientist told us if someone wants his data, he or she can have it in the raw form. His Excel spreadsheets are cryptic and exist in multiple versions, representing each transformation.

Conversely, we found less evidence of proprietary ownership over reference data that provides context, but is not specifically relevant to their research questions. One scientist gave the example of measuring the density of shade cloth for a field experiment. Much work went into determining the amount of shade a particular type of cloth provided in a 24-hour period. He was happy to save other people the effort of reconstructing that number. Similarly, our subjects usually are willing to share software and other tools.

Several of the researchers interviewed did not think their data would be of much use to other researchers. Conversely, some said the data they are collecting already is being shared between themselves and statisticians, engineers, and computer scientists, all with different purposes for the data. One of our subjects recognized the possible uses of his data on water contaminants in a river confluence for such diverse fields as fluid mechanics, public health, ichthyology, and agriculture.

6.3 Data Policy

We encountered a particularly enlightening scenario in which questions arose of whether the data from an instrument belonged to the designers of the instrument or the designers of the experiment. The team member (engineering faculty) who designed and installed the instrument had plans for using the data from the instrument but did not implement those plans. After several years of data production, one of the scientists found the data useful for his own research, and asked the head of the research site for permission to use the data in a publication. Given that no other claims were being made on the data, and that these data were being posted openly on the website of the research site, permission was granted. After some investment in cleaning and analysis, the data looked promising for publication, so the scientist and site director invited the instrument designer to participate in the publication. However, the designer objected strongly on the grounds that they were his data because he had deployed the instrument. The situation was complicated further by the existence of a pending grant proposal involving this instrument by the same designer. We learned later that the situation was resolved only when the pending grant was not awarded, relieving some of the proprietary tension. The scientist we interviewed commented that this was the first real intellectual property issue over data that he had encountered.

In the above case, the technology people are faculty partners in the research. Yet they view the status of the data and the control over it rather differently. Authorship credit on publications is a common issue in research. In cases where instrumentation is essential to the data collection, questions sometimes arise as to whether technical support people should be authors. In another interview, the scientist who provided the above example commented that

> *... tech-support people might get an acknowledgement ... but they're not co-authors on a scientific paper.*

The two situations described here are distinctly different. In the former, the technologist was a researcher who had deployed the instrument, and all agreed that he was entitled to some form of authorship credit. The issue appeared to be about who had priority over the data in determining what should be published, when, and by whom. In the latter situation, a scientist is referring to people who assist with equipment but are not themselves researchers. However, situations may arise where the distinction between technical research and technical assistance is not clear.

In the group meeting (about 20 CENS faculty, students, and research staff) to discuss the ethics and policy of data sharing, several interesting issues arose. One frequent question was about the condition of data to be shared. The group generally agreed that those generating the data had responsibility to assure that the data were reliable and were verified before sharing or posting. Most participants were aware of NSF rules for sharing the data from funded research. At the same time, they were concerned about premature release prior to publication, and whether any sort of liability disclaimers or rights disclaimers (e.g., Creative Commons licenses for attribution and non-commercial reuse) should be applied.

Several of the meeting participants were involved in a small project involving cameras triggered by sensors. The purpose of the project was to test the sensors, but capturing identifiable data on individuals might be unavoidable. They were very concerned about privacy and security issues with these data. Because the study was not about human subjects, they had not sought human subjects review. Several people analogized the situation to webcams on university campuses. Images of people are streaming to public websites without the knowledge or permission of those involved. They also discussed technical solutions such as anonymizing faces captured by the cameras.

6.4 Data Architecture

A longer term goal of our research is to design tools to support data acquisition, management, and archiving. A number of questions addressed the use of data and tools at each stage in the life cycle.

6.4.1 Research Design and Hypothesis Testing

Field research in habitat ecology begins with identifying a research site in which the phenomena of interest can be studied. Before scientists begin setting up sensors or deciding which extant sensors might produce data of interest, they would like a map of the research site that includes the location of sensors and the types of data that each sensor could produce. For example, this scientist would like a map of the area and a table of the data from each set of sensors:

> *I want a table that I can skim really easily and say Okay of these ten stations how many of them have temperature data available? I don't want to look around on a map and have to click on each link*

Scientists spend much time on activities such as sensor placement and development prior to fieldwork. They test equipment and sample the quality of observations from the sensors before they start doing any real science. Thus tools for this exploratory phase are desirable.

6.4.2 Data Collection and Acquisition

Due largely to the relative immaturity of the sensor technology, several of our science subjects were suspicious of automated data collection. They expressed reluctance to take data streams straight from the sensors without good tools to assess the cleanliness of the data. They want simple and transparent methods to find potential gaps in the data; also desirable are tools to identify when values are duplicated or missing, when sensors are failing, and other anomalous situations.

Some also expressed the need to annotate the data in the field, which would provide essential context that cannot be anticipated in data models. For example, data collection tools can have default menus for the available data elements generated by any particular sensor. What they cannot do in advance is predict how scientists will modify the instruments or the field conditions. One scientist mentioned moving his equipment to a different location to compare temperatures. Documenting which data were collected at which location and when is essential to later interpretation. Another good example is distinguishing between common instruments with known characteristics and one that was hand-made. Another scientist might use the same off-the-shelf instrument for another experiment, enabling easy comparisons. If the instrument were unique, the results may be much more difficult to interpret. The scientist might calibrate the two instruments and find they could be used interchangeably, but the future user of the data needs to know what instruments were used and how. The example here is

> *air temperature [from] a little therma-couple that I stuck one foot off the ground with a little aluminum shield that I made. Someone else might use [a common purchased instrument] interchangeably, but they should know that one of them was my home-built little therma-couple thing.*

6.4.3 Data Analysis

Cleaning. The two scientists in the ethnographic study would extract data from the sensor network database to perform any correlations. They prefer to use graphing programs with which they are most familiar. One scientist was disinterested in the offer of graphing tools or visualizations, because he was reluctant to trust other people's graphs. He wants his own graphs so that he can make his own correlations. He also wants the data in a form that he can import into his preferred analysis programs. Data are cleaned with respect to specific research questions. If data are extraneous to a current paper, they may not be cleaned or analyzed. Thus the datasets resulting from a field project are not necessarily complete.

Version Control. Even when scientists are willing to share archived data, those data may be poorly labeled, rendering them difficult to use. Multiple versions may exist of the same data set, resulting from different cleaning or analysis processes. Most of these processes are not recorded, making the data even less accessible to the potential consumer. Multiple versions of datasets complicate retrieval for the scientists who created them and for future researchers who may want to use them. One of our researchers acknowledged that each person on the team created his or her own Excel spreadsheets, and the only access to the data of their teammates was to ask for the spreadsheets and explanations of their contents. This scientist worried that when any of her team members left the project, their data essentially would be lost.

Tools. In these interviews, and in prior interviews and meetings over the last several years, we often found that scientists prefer viewing data in tables, especially Microsoft Excel spreadsheets. One scientist in the ethnographic study offered a detailed explanation. Columns and tables enable him to identify holes in data and to determine how to clean them. Graphs and plots show different types of data inconsistencies, such as identifying dead sensors or graphics in the wrong time scale. Graphs are also personal, because scientists reduce data to test their own hypotheses.

These scientists do not appear to trust transformations made by others; they are more likely to apply their own tools and methods to the original data. The scientist noted above would trust fellow biologists to clean the data:

> *I want some radiation measurements ... you can do energy budget calculations that … have to be cleaned up by a biologist in order to get incorporated into the data set.. any old biologist can come by and start doing correlations, because they know what the data is...*

7 Discussion and Conclusions

While the number of interviews reported in this study is small, the results are based on four years of work with these scientists and technology researchers. Collaborative work can be much slower than solo work due to the effort involved in learning a common language and in finding common ground in research interests [26]. The investments pay off in new insights and new approaches. CENS collaborations between scientists and technologists did not lead to new forms of science as quickly as expected. In its fourth year of existence, the Center is now deploying networked sensor technologies for multiple scientific studies. The promised payoffs are beginning to accelerate. Many of the problems with data cleaning and sensor reliability are due to the immature stage of the instruments and networks. As these scientists become more experienced with these technologies, they are likely to experiment with more instruments and configurations, so the data cleaning and calibration issues will not go away. Experience also is likely to make them more discriminating consumers of the technology, making yet greater demands on the technology researchers.

Scholarly publications have been the product of science for several centuries. Viewing data as a product per se is a relatively new idea. The scientists that we interviewed for this study continue to focus on the paper as their primary product, designing their experiments accordingly. Whether the data will become a direct product of their research to be reused and shared is one of our continuing research questions.

Our tentative findings about data sharing are consonant with other research: In this small sample, our scientists are more willing to share data already published than data that they plan to publish [27]. One scientist was explicit about guarding hand-collected data more closely than data from sensor networks [14]. Our subjects expressed responsibility to assure that any shared data are documented sufficiently to be interpreted correctly [27]. If required to share data they will, knowing that raw data are of little value to others without sufficient cleaning and documentation. However, given a choice, most prefer to exploit their research data fully before releasing them to others.

A number of interesting policy issues arose that we are now studying in much more depth. The Center has made a commitment to share its data with the community and is seeking ways to do so. Members want to provide clean and reliable data, but are understandably concerned about liability and misappropriation of data. Among the questions to address are how to handle sensitive data about human subjects that arises from technical studies, who controls data from a project, and who has first rights to authorship. These are common issues in collaboration, especially at the boundaries between fields. The boundary between life sciences, such as habitat biology, and technology may be even more complex. Not only do differences in data practices between these domains arise, but the distinction between technical research and technical assistance may not always be clear.

These findings have some promising implications for architecture and tools for data management in habitat ecology and perhaps to other field research disciplines. One is the need for tools to explore the research site and the availability of extant data sources. These are especially valuable in the early design stages of an experiment. Visualization tools in the field may be less helpful than expected, as these scientists want to get the data into their own, familiar tools. Quick prototyping of data sources in the field is an essential requirement, and one already recognized in CENS' "human in the loop" architecture. These scientists wish to add new instruments and new details about data and instruments in the field on an ad hoc basis. They invent new tools as needed, using aluminum foil, cloth, tape, and other available equipment. They want data analysis in the field so that they can adjust experiments in real time.

Most of the above requirements suggest hand-crafted tools and structures for this research community. The longer term goal, however, is to build generalizable, scalable tools that facilitate sharing and curation of scientific data. We will continue to work with habitat biologists and other CENS scientists to find a balance between local and global requirements for tools and architecture. While the study reported here relies on a small dataset and is exploratory in nature, it identifies a number of important questions for the design of cyberinfrastructure for science. Research to pursue these questions in more depth is currently under way.

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 and Jeffrey Good, UCLA graduate students, assisted in interviewing and data transcription.

References

1. Atkins, D.E., et al., *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon panel on Cyberinfrastructure*. 2003, National Science Foundation: Washington, D.C.
2. Unsworth, J., et al. *Draft Report of the American Council of Learned Societies' Commission on Cyberinfrastructure for Humanities and Social Sciences*. Last visited 5 November 2005 http://www.acls.org/cyberinfrastructure/acls-ci-public.pdf.
3. Berman, F. and H. Brady. *Final Report: NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Sciences*. Last visited 18 May 2005 http://vis.sdsc.edu/sbe/reports/SBE-CISE-FINAL.pdf.

4. Schroder, P., *Digital Research Data as Floating Capital of the Global Science System*, in *Promise and Practice in Data Sharing*, P. Wouters and P. Schroder, Editors. 2003, NIWI-KNAW: Amsterdam. p. 7-12.
5. Hey, T. and A. Trefethen, *The Data Deluge: An e-Science Perspective*, in *Grid Computing – Making the Global Infrastructure a Reality*. 2003, Wiley.
6. Hey, T. and A. Trefethen, *Cyberinfrastructure and e-Science.* Science, 2005. **308**: p. 818-821.
7. *International Virtual Observatory Alliance*. Last visited 2 March 2005 http://www.ivoa.net/.
8. *Incorporated Research Institutions for Seismology*. Last visited 25 November 2004 http://www.iris.edu.
9. *Biomedical Informatics Research Network.* Last visited 19 March 2005 http://www.nbirn.net/.
10. *GEON*. Last visited 19 March 2005 http://www.geongrid.org/.
11. Star, S.L., *The politics of formal representations: Wizards, gurus and organizational complexity*, in *Ecologies of Knowledge: Work and Politics in Science and Technology*, S.L. Star, Editor. 1995, State University of New York Press: Albany, NY.
12. Estrin, D., W.K. Michener, and G. Bonito, *Environmental cyberinfrastructure needs for distributed sensor networks: A report from a National Science Foundation sponsored workshop*. 2003, Scripps Institute of Oceanography.
13. Zimmerman, A., *New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data.* Science, Technology, & Human Values, under review.
14. Pritchard, S.M., L. Carver, and S. Anand, *Collaboration for knowledge management and campus informatics*. 2004, University of California, Santa Barbara: Santa Barbara, CA. Retrieved from http://www.library.ucsb.edu/informatics/informatics/documents/UCSB_Campus_Informatics_Project_Report.pdf on 14 November 2005.
15. Schnase, J.L., et al. *Building the next generation biological information infrastructure*. in *Proceedings of the National Academy of Sciences Forum on Nature and Human Society: The Quest for a Sustainable World*. 1997. Washington, DC: National Academy Press.
16. Michener, W.K. and J.W. Brunt, eds. *Ecological Data: Design, Management and Processing*. 2000, Blackwell Science: Oxford.
17. Bowker, G.C., *Biodiversity datadiversity.* Social Studies of Science, 2000. **30**(5): p. 643-683.
18. Brown, C., *Lineage metadata standard for land parcels in colonial states.* GIS/LIS '95 Annual Conference and Exposition. American Soc. Photogrammetry & Remote Sensing & American Congress on Surveying & Mapping. Bethesda, MD, USA., 1995. Part 1, Vol 1: p. 121-130.
19. Kwa, C., *Local ecologies and global science: Discourses and strategies of the International Geospher-Biosphere Programme.* Social Studies of Science, 2005. **35**(6): p. 923-950.
20. Shankar, K., *Scientific data archiving: the state of the art in information, data, and metadata management.* 2003.
21. Sandoval, W.A. and B.J. Reiser, *Explanation-driven inquiry: Integrating conceptual and epistemic supports for science inquiry.* Science Education, 2003. **87**: p. 1-29.
22. Glaser, B.G. and A.L. Strauss, *The discovery of grounded theory: Strategies for qualitative research*. 1967, Chicago: Aldine Publishing Co.
23. Engeström, Y., *Activity theory and individual and social transformation*, in *Perspectives on activity theory.* 1999, New York: Cambridge University Press: p. 19-38.

24. Engeström, Y., *Learning by Expanding: An activity-theoretical approach to developmental research*. 1987, Helsinki: Orienta-Konsultit.
25. Cole, M. and Y. Engeström, eds. *A Cultural-historical Approach to Distributed Cognition.* 1993, New York: Cambridge University Press.
26. Cummings, J.N. and S. Kiesler, *Collaborative research across disciplinary and organizational boundaries.* Social Studies of Science, 2005. 35(5): p. 703-722.
27. Arzberger, P., et al., *An International Framework to Promote Access to Data.* Science, 2004. 303(5665): p. 1777-1778.