Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries

Jillian C. Wallis[1], Christine L. Borgman[2], Matthew S. Mayernik[2], Alberto Pepe[2], Nithya Ramanathan[3], and Mark Hansen[4]

[1] Center for Embedded Networked Sensing, UCLA
jwallisi@ucla.edu

[2]Department of Information Studies
Graduate School of Education & Information Studies, UCLA
borgman@gseis.ucla.edu, mattmayernik@ucla.edu, apepe@ucla.edu

[3]Department of Computer Science
Henry Samueli School of Engineering & Applied Science, UCLA
nithya@cs.ucla.edu

[4]Department of Statistics
College of Letters & Science, UCLA
cocteau@stat.ucla.edu

Abstract. For users to trust and interpret the data in scientific digital libraries, they must be able to assess the integrity of those data. Criteria for data integrity vary by context, by scientific problem, by individual, and a variety of other factors. This paper compares technical approaches to data integrity with scientific practices, as a case study in the Center for Embedded Networked Sensing (CENS). The goal of this research is to identify functional requirements for digital libraries of scientific data that will serve this community. Data sources include analysis of documents produced by the CENS data integrity group and interviews with science and technology researchers within CENS.

Keywords: data integrity, data quality, trust, user centered design, user experience, scientific data.

1 Introduction

Digital libraries of scientific data are only as valuable as the data they contain. Users need to trust the data, which in turn depends on notions such as data integrity and data quality. How can these notions be applied to the design of digital libraries for scientific data? Scholarly publications are vetted through peer review processes, but comparable mechanisms to evaluate data have yet to emerge. Data that are reported in publications are evaluated in the context of those publications, but that is not the same as evaluating the data per se for reuse. When data are submitted to repositories such as the Protein Data Bank, they are evaluated rigorously. When data are made available through local websites or local repositories, mechanisms for data authentication are inconsistent. Researchers (or teachers or students) who wish to reuse data rely on a variety of indicators such as reputation of the data collector and the institution, quality of papers reporting the data, and sufficient documentation to interpret the data [1]. Criteria and methods that can be applied to data authentication are essential to the design of digital libraries for eScience.

Criteria for data integrity and quality will vary considerably by type of data and by scientific domain. Studies of scientific practices will shed light on how researchers and technologists define and apply these concepts. Design criteria for digital libraries, in turn, can emerge from these studies of practices.

Research reported here is affiliated with the *Center for Embedded Networked Sensing* (CENS), a National Science Foundation Science and Technology Center established in 2002 [http://www.cens.ucla.edu/]. CENS supports multi-disciplinary collaborations among faculty, students, and staff of five partner universities. The Center's goals are to develop and implement wireless sensing systems that enable dense spatial and temporal *in situ* sensing of various environments. CENS' research crosses four primary scientific areas: habitat ecology, marine microbiology, environmental contaminant transport, and seismology, plus applications in urban settings and in the arts. Application of this technology has already been shown to reveal patterns and phenomena that were not previously observable.

2 Problem Statement

CENS' scientific deployments are generating far more data than can be managed by the traditional methods used for field research. CENS researchers are committed in principle to making their data available for reuse by others. However, they are finding that substantial effort is required to capture and maintain these large volumes of data for their own use, and that even more effort appears to be required to make them available for reuse by others. These data are an important end product of scientific research. They can be leveraged for future analyses by the same or other investigators, whether for comparative or longitudinal research or for new research questions. The ability to interpret data collected by others depends, at least in part, on the ability to assess the integrity and quality of those data. Criteria for data integrity vary by context and by individual, however.

Researchers often prefer to use their own data because they are intimately familiar with how those data were collected, the actions that were taken in the field to collect them, what went wrong and what was done to fix those problems, and the context in which the data were collected. As data production becomes an end unto itself, instead of solely another step towards a publication, and researchers use data produced by others in their own publication, we need consistent methods to document data integrity and quality factors in ways that will facilitate data interpretation.

Digital library tools and services will be important mechanisms to facilitate the capture, maintenance, use, reuse, and interpretation of scientific data. This paper draws together studies of data practices of CENS researchers and analyses of technical approaches to managing data integrity and quality, with the goal of establishing functional requirements for digital libraries of scientific data that will serve this community. Two of the authors of this paper are involved primarily in studies of data practices, two primarily in systems design for data integrity, and two primarily in the development of digital libraries.

2.1 Scientific Data

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 [2]. 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.

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 [3, 4] or on large collaborations in areas such as biodiversity [5-7]. 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.

One of the biggest challenges in developing effective digital libraries in areas such as habitat ecology is the "data diversity" that accompanies biodiversity [5]. Habitat ecologists observe phenomena at a local scale using relatively ad hoc methods [8]. 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 [1, 9]. For example, CENS technology researchers may view the presence or absence of data as an indicator of the functionality of the equipment, whereas the application science researchers may require data that accurately reflect the environment being measured [10].

2.2 Sensor Network Data

While researchers in process control have studied faults, failures, and malfunctions of sensors for many years [11], the problem is significantly harder in the case of sensor networks. First, the scale is much larger in sensor networks: the number of sensors is much greater, areas of coverage are much larger, and wireless interconnections may be lossy. Second, the phenomena being observed in many applications of sensor networks are far more complex and unknown than the manufacturing and fabrication plants studied in classical process control. Consequently, model uncertainty is much higher, and often the model is unknown. Lastly, while in process control the inputs provided to the plant are controlled and at least measured, such is not the case with many phenomena that sensor networks observe (environmental phenomena; inhabited buildings or other structures). Together, these differences make the problem of detecting, isolating, diagnosing, and remediating faults and failures, and being resilient to their occurrence, much harder in sensor networks than in traditional plant control.

2.3 Static vs. Dynamic Embedded Sensor Networks

Sensor networks, per se, are not a new technology. Large manufacturing operations and chemical processing plants, for example, rely heavily on sensor networks to manage operations. Similarly, water flow and water quality monitoring relies heavily on embedded sensor networks. Most of these applications of sensor networks are static deployments: sensors are placed in appropriate positions to report data continuously on local conditions. Sensors are monitored, both by humans and by computers, to determine changes in conditions. Autonomous networks can rely on machine actuation to capture scientifically relevant data, to alter data collection (e.g., capture data more frequently if excessive pollution is suspected), or to report emergencies that require intervention (e.g., faults in dams, water contamination).

While the initial framework for CENS was based on autonomous networks, early scientific results revealed the difficulty of specifying field requirements in advance well enough to operate systems remotely. Most CENS' research is now based on dynamic "human in the loop" deployments where investigators can adjust monitoring conditions in real time. CENS' teams have data collection "campaigns" in which they deploy an embedded sensor network in the field for a few hours or a few days. They may return to the same site, or a similar site, repeatedly, each time with slightly different equipment or research questions.

These discrete field deployments offer several advantages to the scientific researchers, and allow for the deployment of prototype, delicate, or expensive equipment. Scientists also can alter the position of their sensors and the frequency of sampling while in the field, and collect samples for in-field verification. However, the dynamic nature of these deployments poses additional challenges to data integrity, as the conditions, context, and sensor technology all may vary by deployment.

2.4 Data Use and Reuse

Our earlier reports from this project focus on how and when CENS researchers use and reuse their data [10, 12, 13]. Our respondents almost exclusively use data collected by themselves or their research group. They also prefer to keep their own data for potential reuse in future research projects. Reuse of their own data was much preferred over using data collected by others because the researchers are familiar with the context of their own data collection, including its subtleties and quirks; such knowledge of data integrity is difficult to obtain for data collected by other researchers.

3 Research Methods

Our initiative within CENS is aimed at improving data practices by providing researchers with a transparent framework of research tools that will allow them to create, describe, store, and share data resources efficiently. To be effective for data management, such tools should document data integrity and quality sufficiently that CENS researchers can interpret the data, and better yet, so that future researchers can reuse these data. We have applied a variety of research methods over a five-year period, including survey studies, field observation, and documentary analyses [10, 12, 13].

In this paper we compare technical approaches to data integrity with scientists' practices associated with data integrity. We draw upon two data sources to address functional requirements specific to data integrity and quality: (1) analysis of documents produced by the CENS data integrity group and interviews with members of that group and (2) interviews with domain scientists in CENS.

3.1 CENS as a Case Study

Research in the first three years of the Center (2002-2005) was driven more by computer science and engineering requirements than by scientific problems. Initial research focused heavily on the design and deployment of sensing technology. Concerns about equipment reliability, capacity, and battery life outweighed considerations of data quality and usefulness. Now that many basic technical problems are resolved, the CENS research program has become more science-driven. Computer science and engineering research can focus on technology improvements that address scientific problems, and all partners can focus more attention on data integrity and value. CENS' immediate concerns for data management, its commitment to sharing research data, and 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.

3.2 Document Analysis

The document analysis involved the identification and reading of documents written by CENS technical researchers (i.e., computer science, engineering, and statistics) to identify implicit and explicit criteria for data integrity and quality. Documents analyzed include published literature, technical reports, internal memoranda, and grant proposals.

3.3 Study of Scientific Data Practices

The interview data reported here are drawn from a study of five environmental science projects within CENS. For each project we interviewed a complementary set of science and technology participants, including faculty, post-doctoral fellows, graduate students and research staff. We interviewed 22 participants, each for 45 minutes to two hours; interviews averaged 60 minutes. Results from interviews with technical researchers are included in the section on technical approaches to data integrity; results from interviews with the scientists are reported in the section on scientific practices.

The interviews were audiotaped, transcribed, and complemented by the interviewers' memos on topics and themes [14]. Analysis proceeded to identify emergent themes. We develop a full coding process using NVIVO, which was used to test and refine themes in coding and subsequent interviews. With each refinement, the remaining corpus was searched for confirming or contradictory evidence. This study used the methods of grounded theory [15] to identify themes and to test them in the full corpus of interview transcripts and notes. Interview questions were grouped into these four categories: data characteristics, data sharing, data policy, and data architecture. In this paper we report only responses that discussed data integrity, quality, trust, or other information that needs to be known to interpret data. Most of these responses were elicited by questions about data characteristics or data architecture.

4 Results

Results are reported in two sections. First we present technical approaches to data integrity, drawing upon the observations and expertise of CENS researchers in computer science, engineering, and statistics. Second we present scientific approaches to data integrity, drawn from the interviews with domain scientists within CENS. Most of these respondents are faculty or doctoral students in the biological sciences.

4.1 Technical Approaches to Data Integrity

As CENS research has matured and many basic technical challenges of sensor systems have been addressed, data integrity and quality have become driving concerns of all parties in this multidisciplinary collaboration. The Integrity Group, consisting of ten students and three faculty (from computer science, engineering, and statistics), was formed to focus specifically on technical approaches to assuring and improving data integrity. This group has (i) surveyed existing approaches to data integrity, (ii) implemented both rule-based and statistical learning algorithms, and (iii) initiated data integrity experiments, either leveraging existing CENS field deployments or designing original experiments. Members of this group are routinely included in pre-deployment design discussions and consulted during post-deployment analysis, for applications as diverse as aquatic sensing [16], a soil observing system for examining CO2 flux [17], and a short-term deployment in a rice paddy in Bangladesh to study groundwater arsenic content [18].

The Integrity group has led two significant development efforts within CENS that influence the design of digital libraries. First is SensorBase.org, a database platform for data from short-term, rapidly deployed experiments and from longer-lived, continuously operating installations [19, 20]. SensorBase.org provides the sensing research community with a framework for both sharing data and for experimenting with models and computation to support data integrity. Many of its diagnostics and alerting capabilities, leveraging RSS and email, facilitate research by the Integrity Group. The second development is a move toward in-field analysis of data to support both system design and monitoring. This project is more diffuse, branching across several Ph.D. projects and not yet producing a single platform. Methods to access models and data in the field are rapidly becoming part of most CENS systems.

Criteria for Data integrity and Quality. Technical approaches to data integrity build upon engineering criteria for integrity and quality. For example, the Federal Standard 1073C (and its updated Telecom Glossary 2000) define data integrity to be "1. The condition existing when data is unchanged from its source and has not been accidentally or maliciously modified, altered, or destroyed. 2. The condition in which data are identically maintained during any operation, such as transfer, storage, and retrieval." However, this definition of data integrity is inadequate for creators and users of embedded sensor network data. A processing error or fault in this context is not just a matter of networking or database operations, but might trace back to the sensor, its calibration, and the conditions surrounding measurement. The notion of a source alluded to in this standard shifts from the sensor node to the environmental phenomenon under observation. With this shift, the network extends to the physical coupling of the sensing apparatus; thus a broader range and set of characteristics of network services are required to guarantee data integrity. Data integrity, in this sense, is related to the social science notion of reliability – that the data are being measured consistently, in a replicable manner. Observations made by a sensor are intact throughout the processing of sensor network.

Data integrity is just one aspect of data quality, and the line between the two can be fuzzy. For embedded sensing, inferences are often about environmental phenomena, and to achieve high quality data, it is necessary to optimize aspects of the physical coupling to get a good perspective on the phenomenon. Data quality, in this sense, is comparable to the social science notion of validity – that the data represent some true or valid indicator of the phenomenon being studied.

Fault Detection. Fault detection is an important technical component of data integrity for embedded networked sensing systems. Technical researchers are concerned that CENS scientists tend to view fault detection as a component of post-deployment analysis. Instead of identifying faults in real-time, scientists often assume they can wait until all the data have been collected, and throw away faulty data after the fact. From a technical perspective, this assumption is flawed for two reasons: i) it is not always easy to tell which data are faulty after the fact. Researchers often need specific information about the context (e.g., there was an irrigation event at 3PM today), or need to take physical measurements (e.g., extracting physical samples to validate the sensor data) to determine if the sensor data are faulty, and ii) especially for soil sensor deployments, where sensors are short-lived and require frequent calibration, the amount of data available is so small that none can be spared. For example, during one deployment, 40% of the data had to be discarded, limiting the amount of scientific analysis possible.

Simple fault detection also includes applying thresholds to data in order to separate good and bad data. This approach is also not ideal because environments are dynamic, and notions of what it means to be faulty change over time, both as the sensor ages, and as environmental processes develop. Further, notions of faults change across different deployments, so the user must set new thresholds for each new sensor and environment.

People tend to use simple solutions to detect faults, such as looking for holes in the data or data that do not correlate across the sensors within the same variable, or estimation using quicker methods to compare with the sensor data. Fault detection is not yet a priority for most CENS researchers, whether for detecting faults in the network or in the sensors.

Tools to Manage Data Integrity. As CENS has moved from static to dynamic deployments of sensor networks, the need to assess data integrity in real time has become ever more apparent. If scientists, with the assistance of their partners in computer science, engineering, and statistics, can interact with the network while in the field to perform data analysis and modeling, they can significantly and quantifiably improve the quality of the collected data. For example, physical soil samples taken at specific times were useful in validating questionable chloride and nitrate data collected by the network of sensors in Bangladesh. These lessons are being incorporated into the design of Confidence [21], a unified system geared towards increasing the quantity and quality of data collected from a large network. Confidence enables users to effectively administer large numbers of sensors and nodes by automating key tasks and intelligently guiding a user to take actions that demonstrably improve the data and network quality. Confidence uses k-means clustering and a carefully chosen set of system metrics (i.e. features) to group similar data points and identify actions a user can take to improve system quality. Clusters are learned over time, thus they adapt to different environments and sensors. The clustering and feature approach is general, and applies to improving both network and sensor quality. Confidence outperforms similar systems by avoiding precise diagnoses and decision trees that employ statically set thresholds, as was the case with a precursor tool called Sympathy that was also developed at CENS [22].

4.2 CENS Application Scientists

"But we have to have confidence, I guess, in what the measurements are collecting for information." This simple statement by a CENS scientist belies the complexity of achieving trust in one's own data. There are many factors that influence a researcher's confidence in data, just as there are many complexities in the processes of generating and capturing data. To completely have confidence in data, a scientist must trust the data collection process from beginning to end. Trust envelops the entire data life cycle, from the selection and calibration of equipment, to in-field setups and equipment tests, to equipment reliability once it's in the field, to human reliability.

Trust is cemented by documentation that covers each step. Results reported in this section answer the question of what scientists need to know about the data collection process in order to interpret and trust the data, an interchange that is implicitly informed by data integrity and quality.

Equipment Selection. As with any task, the equipment used must be able to perform the task adequately. Thus it is necessary to understand the capabilities or limitations of the sensor, in order to determine whether it is the right sensor for the observations to be measured. "You really need to know what its limitations are, what are its confounding factors, so that you can be relatively confident that your reading is correct." Each model of sensor has a level or range of sensitivity, and some applications require a very fine level of sensitivity and some require a more gross reading. Understanding where and how the sensor is to be used informs the choice of equipment and how it is used.

Sensors use many different methods to measure a variable. These sensing methods can either be direct or by proxy. The method will inform the interpretation or the trust.

> "There are hundreds of different ways of measuring temperature. If you just say, 'The temperature is…,' then that's really low-value compared to, 'The temperature of the surface measured by the infrared thermal pile, model number XYZ, is…'. From this I know that it is measuring a proxy for a temperature, rather than being in contact with a probe. And it is measuring it from a distance. I know that its accuracy is plus or minus .05 of a degree based on the instrument itself. I want to know that it was taken outside versus inside in a controlled environment."

Equipment Calibration. Off-the-shelf sensors presumably have been tested for quality before being sold. This testing would include some calibration against the same standards described in the technical specifications. The majority of the off-the-shelf sensing equipment used within CENS are calibrated by the investigators and their technical staff. Some sensing equipment can only be calibrated by the manufacturer, which requires that equipment be sent back upon occasion.

> "We calibrate against a standard. So it depends on the instrument. If it's something simple we can calibrate it here. If it's a more high-tech instrument, like a lot of what we use are infrared gas analyzers for measuring photosynthesis and they're factory calibrated. We've got to send it back to the factory, so that's a 35 to 1,000 dollar suitcase instrument so we send it back once or twice a year to get it calibrated. So they charge us two or $300 but we get it calibrated. So the complicated things we definitely send back."

Each sensor model has a specific process for calibration, as well as specific standards that the sensors are be calibrated to.

> "The parameters that we collect for each sensor, is the upper and lower detection limit. And the slope and the Y-intercept for the calibration equation. So the calibration equation is just a linear Y = MX + B. And we just record the slope and the y-intercept so those are the four parameters that we record."

Calibration information can be captured about each sensor in a succinct manner.

In addition to the slope and Y-intercept, it is important to capture the date of the most recent calibration. The calibration changes with time, thus an important part of interpreting the data would be to know how far the sensor had drifted.

Ground-truthing. Unfortunately once calibrated, equipment does not actually remain calibrated. "Really there is no way to measure in laboratory conditions and have it apply to the field." This is partly due to the uncertainty of field conditions in order to calibrate to them, and the frailty of the equipment that calibration will degrade over time. The methods used to establish in-field calibration are referred to as ground-truthing, and include such techniques as periodically flushing a sensor with a standard solution, over-sampling, and capturing physical samples to validate measures.

Thresholding. The available range is affected by the sensitivity of the instrument, and it may be important to move to a more sensitive instrument if the expected measure is close to one of these thresholds or boundaries, to distinguish between actual measures and faulty data. These boundaries are derived from the calibration equations discussed previously.

> "And then the other thing is that we also use the calibration equation parameters in order to detect faults. And so a simple example is, well if a data point is outside what we call the linear detection range of the calibration equation, so it's just like an upper and a lower limit, then there's something suspicious going on. And so depending on the pattern of the data, we try to narrow it down."

5 Discussion

The early years of wireless sensing research were focused largely on the problems associated with resource-constrained communications and processing of sensed data, and metrics such as quantity and timelines of data collected. Not much attention was paid to the quality of information returned by the system, and the integrity of the system itself. However, with deployment experience building up, this has begun to change. Section 4.2 above illustrates how many researchers and users of wireless sensing systems now recognize data and system integrity as very challenging problems that limit the scalability of this technology. This recognition has moved the principle focus of data integrity activities up from post-deployment to during the deployment. Facilitating cleaner data collection upstream should help to alleviate the problems in identifying potentially errant data, reducing the manual effort required to clean data after the deployment. The technical approaches outlined in section 4.1 above are attacking the problem of data integrity coming from CENS' wireless sensing systems by identifying faulty data during the collection. These techniques facilitate greater trust in those data, and enable scientists to analyze data with the assurance that data is complete and of high quality.

Human-in-the-loop deployment approaches increase flexibility in the field, but face the challenge of documenting the human factor. Deployments currently rely on "oral culture", assisted by varying and individual documentation techniques. A set of complimentary data structures is currently being developed by CENS to capture sensor data and metadata, and form an information ecology. CENS Deployment Center is a deployment planning tool that captures metadata describing the deployment at large, the equipment and sensors used, and the people involved. SensorBase.org, described in section 4.1 above, is a flexible data digital library, which allows for the slogging of data directly off the sensors. The CENS Bibliographic Database describes the publications associated with the sensor data. Publications have traditionally been used as an access point to data, as well as a wrapper containing the description of the equipment and methods used to collect the data, as such they are an important part of understanding the data collected. The information ecology described here can be leveraged before, during, and after the deployment to collect contextual metadata.

6 Conclusion

We are working towards an architecture for data integrity and quality in wireless sensing systems. Through interviews, observation, consultation, and research, we have an accurate idea of the existing and developing data integrity activities. Wireless sensing systems have advanced to the point where the technology is producing data of real scientific value. Data integrity problems must be addressed if the data produced by these sensing systems are to be useful to scientists on a large scale.

Digital libraries can facilitate data integrity by recognizing and accounting for the variety of relevant issues outlined in this paper. Scientists have pre-existing methods for describing the network, sensors, and calibrations, but this information is largely documented separately from the data itself. How can digital libraries store this important information and have it associated with each relevant data point? Sensor faults have a huge impact on the quality and quantity of data generated on wireless sensing system deployments. How can sensor fault detection be reflected in the digital library? Calibration information is essential to post-deployment data analysis, but calibration information varies for each type of sensor, and in some circumstances even between sensors of the same type on the same deployment. What level of granularity in the calibration information needs to be associated with each data set? Future architecture for wireless sensing systems must address capturing, organizing, and accessing this information.

Acknowledgements. CENS is funded by National Science Foundation Cooperative Agreement #CCR-0120778, Deborah L. Estrin, UCLA, Principal Investigator; Christine L. Borgman is a co-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. Alberto Pepe's participation in this research is supported by a gift from the Microsoft Technical Computing Initiative. The CENS Integrity Group is supported by NeTS-NOSS seed funding. SensorBase research in CENS is led by Mark Hansen and Nathan Yau. Support for CENS bibliographic database development is provided by Christina Patterson and Margo Reveil of UCLA Academic Technology Services.

References

1. Borgman, C.L., Scholarship in the Digital Age: Information, Infrastructure, and the Internet. 2007, Cambridge, MA: MIT Press.
2. Hey, T. and A. Trefethen, The Data Deluge: An e-Science Perspective, in Grid Computing – Making the Global Infrastructure a Reality. 2003, Wiley.
3. Traweek, S., Beamtimes and lifetimes : the world of high energy physicists. 1st Harvard University Press pbk. ed. 1992, Cambridge, Mass.: Harvard University Press. xv, 187.
4. Traweek, S., Generating high energy physics in Japan, in Pedagogy and Practice in Physics, D. Kaiser, Editor. 2004, University of Chicago Press: Chicago.
5. Bowker, G.C., Biodiversity datadiversity. Social Studies of Science, 2000. 30(5): p. 643-683.
6. Bowker, G.C., Mapping biodiversity. International Journal of Geographical Information Science, 2000. 14(8): p. 739-754.
7. Bowker, G.C. Work and information practices in the sciences of biodiversity. in VLDB 2000, Proceedings of 26th international conference on very large data bases. 2000. Cairo, Egypt: Kaufmann.
8. Zimmerman, A.S., New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. Science, Technology, & Human Values, under review.
9. Bowker, G.C., Memory Practices in the Sciences. 2005, Cambridge, MA: MIT Press.
10. Borgman, C.L., J.C. Wallis, and N. Enyedy, Little Science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. International Journal on Digital Libraries, in press.
11. Isermann, 2005.
12. Borgman, C.L., et al. Drowning in Data: Digital Library Architecture to Support Scientists' Use of Embedded Sensor Networks. in JCDL '07: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries. in press. Vancouver, BC: Association for Computing Machinery.
13. Borgman, C.L., J.C. Wallis, and N. Enyedy. Building Digital Libraries for Scientific Data: An exploratory study of data practices in habitat ecology. in European Conference on Digital Libraries. 2006. Alicante, Spain.
14. Lofland, J., et al., Analyzing Social Settings: A Guide to Qualitative Observation and Analysis. 2006, Belmont, CA: Wadsworth/Thomson Learning.
15. Glaser, B.G. and A.L. Strauss, The discovery of grounded theory; strategies for qualitative research. Observations. 1967, Chicago,: Aldine Pub. Co. x, 271.
16. Singh, 2007.
17. Ramanathan, N., 2006.
18. Ramanathan, N., et al. Designing Wireless Sensor Networks as a Shared Resource for Sustainable Development. in Information and Communication Technologies and Development. 2006.
19. Chen, G., N. Yau, and M. Hansen, 2007.
20. Chang, K., et al. SensorBase.org - A Centralized Repository to Slog Sensor Network Data. in International Conference on Distributed Networks (DCOSS)/EAWMS. 2006.
21. Ramanathan, N., et al., Fixing Faults in Wireless Sensing Systems with Confidence, in CENS Technical Report. in submission.
22. Ramanathan, N., et al. Sympathy for the Sensor Network Debugger. in SenSys. 2005. San Diego, CA.