Title

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

Authors

Wallis, J C
Borgman, C L
Mayernik, Matthew
et al.

Publication Date

2007-08-29

DOI

10.1007/978-3-540-74851-9_32

Peer reviewed

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) in the use of wireless, in-situ sensing for the collection of large scientific data sets. The goal of this research is to identify functional requirements for digital libraries of scientific data that will serve to bridge the gap between current technical approaches to data integrity and existing scientific practices.

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. Users also need the means to assess the quality of 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 [1], they are evaluated rigorously. When data are made available through local websites or local repositories, mechanisms for data authentication are less consistent. Scientific researchers often prefer to use their own data because they are intimately familiar

L. Kovács, N. Fuhr, and C. Meghini (Eds.): ECDL 2007, LNCS 4675, pp. 380–391, 2007.

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, the context in which the data were collected, and local subtleties and quirks. Such knowledge of data integrity is difficult to obtain for data collected by other researchers. 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 documentation. Standardized criteria and methods that users can apply to assess data quality are essential to the design of digital libraries for eScience [2].

Enabling reuse of scientific data can be of tremendous future value as such data are often expensive to produce or impossible to reproduce. 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 to study small teams that produce observations of long-term, multi-disciplinary, and international value, such as those in the environmental sciences. The emergence of technology such as wireless sensing systems has contributed to an increase in the volume of data that can be generated by small research teams. Scientists can perform much more comprehensive spatial and temporal in situ sensing of environments than is possible with manual field methods. The "data deluge" resulting from these new forms of instrumentation is among the main drivers of e-Science and cyberinfrastructure [8]. Data produced at these rates can be captured and managed only with the use of information technology. If these data can be stored in reusable forms, they can be shared over distributed networks.

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 across disciplines ranging from computer science to biology. The Center's goals are to develop and implement wireless sensing systems as described above, and to apply this technology to address questions in four scientific areas: habitat ecology, marine microbiology, environmental contaminant transport, and seismology. Application of this technology already has been shown to reveal patterns and phenomena that were not previously observable. 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.

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.

Section 2 discusses the characteristics of scientific sensor data collected by CENS. Section 3 presents the research methods used to study data integrity practices of CENS researchers. Research results are presented in Section 4 and discussed in Section 5. Our findings address ways in which digital libraries can assist in improving the integrity of scientific data and in facilitating their reuse.

2 The Data Integrity Problem Redefined

CENS' scientific sensor 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.

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, consistent methods are needed to document data integrity and quality criteria in ways that will facilitate data interpretation. The variety of practices associated with data management and range of understanding of what constitutes "data," which are well known issues in social studies of science [9], present practical problems in the design of digital libraries for wireless sensing data.

2.1 Static vs. Dynamic Embedded Sensor Networks

Sensing systems are not a new technology, *per se*. Common applications of sensing networks in the environmental sciences include monitoring of water flow and quality, for example. Most applications of wireless sensing systems in the environmental sciences 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. In addition to conducting extended "static" sensor deployments, where sensing systems are installed and left for weeks or months at a time with only intermittent physical monitoring, CENS teams regularly conduct short term "campaigns" to collect data, in which they deploy a wireless sensing system in

the field for a few hours or a few days with constant human presence. They may return to the same site, or a similar site, repeatedly, each time with slightly different equipment or research questions.

Discrete field deployments offer several advantages to the scientific researchers, allowing 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 may vary by deployment.

2.2 Data Diversity

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 [10]. 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 [2, 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 [11].

2.3 Wireless Sensing Data

While researchers in process control have studied faults, failures, and malfunctions of sensors for many years [12], the problem is significantly harder in the case of wireless sensing systems. First, the scale is larger in wireless sensing systems in terms of number of sensors and areas of coverage. Second, the phenomena being observed in many applications of wireless sensing systems are far more complex and unknown than the manufacturing and fabrication plants studied in classical process control. Consequently, model uncertainty is higher, and often the model is unknown. Third, the sensors used in scientific experiments are often in nascent stages of development and not yet designed for robust field use. Frequent calibration and sensor damage are among the faults that affect the quality of sensor data. Fourth, whereas sensors in factories obtain power and connectivity over wires, resulting in a robust data-delivery infrastructure, wireless sensing systems rely on batteries and wireless channels. Even well planned deployments experience high rates of packet loss [13], resulting in largely incomplete datasets. Lastly, in process control, inputs to the plant are controlled and measured, which is not the case with many phenomena observed by wireless sensing systems (e.g., environmental phenomena; inhabited buildings or other structures). Together, these differences make the problems of detecting, isolating, diagnosing, and remediating faults and failures, and being resilient to their occurrence, more difficult in wireless sensing systems than in traditional plant control.

2.4 Digitizing the Oral Culture

CENS has relied on a largely oral culture for the exchange of information about how data are collected, the equipment used, and the state of the equipment. As the Center has grown, an oral culture is no longer sufficient to capture and retain institutional memory. The student research population turns over rapidly and tacit knowledge needs to be exchanged within and between a larger number of research teams. These are but two reasons for communication breakdowns to occur in the data lifecycle. Research deployment practices were identified as a critical area that required more consistent documentation and better means of information exchange.

Data sharing is often an interpersonal exchange between data collectors and data requestors. This can be a time- and labor-intensive process to describe and document data appropriately for use by others. Interpersonal exchanges do not scale well to large research centers and frequent data requests [2]. Much of our research is devoted to developing tools, services, and policies that will facilitate data capture, management, use, and sharing, while respecting rights and preferences of researchers in determining what data to release to whom, in what formats, and under what conditions [11, 14, 15].

3 Research Methods

The goal of our research initiative within CENS is to provide researchers with a transparent framework of tools that will allow them to create, describe, store, and share data resources efficiently. The design of these tools, and associated digital library services and policies, is based on studies of data practices. We have applied a variety of research methods over a five-year period, including survey studies, field observation, and documentary analyses [11, 14].

In this paper we compare technical approaches to data integrity with scientists' practices associated with data integrity. We draw upon multiple sources to identify functional requirements for digital libraries, including analysis of documents produced by the CENS data integrity group, interviews with members of that group, interviews with domain scientists, computer scientists, and engineering researchers in CENS, and analysis of existing data sets and data archives.

3.1 Studies 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 [14, 15]. Results from interviews with computer science and engineering 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.

Interviews were audiotaped, transcribed, and complemented by the interviewers' memos on topics and themes [16]. Initial analysis identified emergent themes. A full coding process, using NVIVO, was used to test and refine themes in the interview

transcripts. With each refinement, the remaining corpus was searched for confirming or contradictory evidence, using the methods of grounded theory [17]. Interview questions were grouped into 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 related issues about data interpretation. Most of the responses reported here were elicited by questions about data characteristics or data architecture.

3.2 Integrity Research Group

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, addresses technical approaches to data integrity. This group has surveyed existing approaches to data integrity, implemented both rule-based and statistical learning algorithms, and 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 [18], a soil observing system for examining CO2 flux [19], and a short-term deployment in a rice paddy in Bangladesh to study groundwater arsenic content [20].

4 Results

Results are reported in two sections. First we summarize current scientific practices to ensure data integrity. Evidence of these practices is drawn from the interviews with domain scientists within CENS; most of these respondents are faculty or doctoral students in the biological sciences. Second we present technical approaches to ensuring data integrity, drawing upon the observations and expertise of CENS researchers in computer science, engineering, and statistics. Of the many systems approaches being pursued at CENS, we have identified these as having the most direct impact on digital library design for CENS research.

4.1 Needs of CENS Application Scientists

"We have to have confidence…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. Many factors influence a researcher's confidence in data, most of which arise from the complexities of generating and capturing data. Confidence in data depends upon trust in the entire data life cycle, from the selection and calibration of equipment, to in-field setups and equipment tests, to equipment reliability once it is in the field, to human reliability. Trust can be enhanced by documentation of each step in the process and by recording of tacit knowledge that may be exchanged orally in the field. Lab and field notebooks also are essential forms of documentation, whether in paper or digital form. Results reported in this section

address questions of what scientists need to know about the data collection process to interpret and trust the data, which in turn depends upon 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 a given sensor to determine whether it is appropriate to capture the desired observations. As one scientist put it, *"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 others require a more gross reading. Understanding where and how the sensor is to be used informs the choice and use of equipment.

Sensors can measure variables in multiple ways. Some sensing methods are direct and others are proxy-based. The method chosen will influence both the interpretation of the resulting data and one's trust in them, as illustrated by the following comment of a biologist:

> *"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. Such testing normally includes calibration against the standards described in the technical specifications. The majority of off-the-shelf sensing equipment used by CENS researchers are also calibrated by the investigators and their technical staff. Sensing equipment that can only be calibrated by the manufacturer must be returned periodically for recalibration, as described by this researcher:

> *"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… once or twice a year to get it calibrated… the complicated things we definitely send back."*

Each sensor model has a specific process for calibration and specific standards for calibration, as reflected in this comment:

> *"The [four] parameters that we collect for each sensor [are] the upper and lower detection limit...and the slope and the Y-intercept for the calibration equation…the calibration equation is just a linear Y = MX + B."*

Calibration information for sensors such as these can be captured in a succinct manner. Other important information to capture is the date of the most recent calibration, because once calibrated, equipment does not necessarily remain

calibrated. As another scientist said, *"there is no way to measure in laboratory conditions and have it apply to the field."* Thus an important part of interpreting the data includes knowledge of how the calibration parameters change over time. One approach to capturing changing calibration parameters is periodically to calibrate the sensor in-situ (i.e., without extracting it from the soil or water) by providing a known input and recording the reported output.

Ground-truthing. Unfortunately data from many sensors cannot be blindly trusted. This is partly due to the uncertainty of field conditions and partly to frailty of equipment. Calibration accuracy is known to degrade over time. When possible, scientists periodically validate sensor data by applying a known perturbation to a sensor, over-sampling the phenomena, and capturing physical samples (e.g., water, dirt, leaves, plankton) to validate measures.

4.2 Technical Approaches to Data Integrity

The Integrity Group has led two significant development efforts within CENS that influence the design of digital libraries. First is a move toward in-field analysis of data to support both system design and monitoring. This project is diffuse, branching across several Ph.D. projects and not yet producing a unified platform, but essential because methods to access models and data in the field are becoming part of most CENS systems. Second is SensorBase.org, a database platform for data from short-term, rapidly deployed experiments and from longer-lived, continuously operating installations [21, 22]. SensorBase.org is a central component of CENS' data ecology.

Real-Time, Adaptive Fault Detection. Fault detection is an important technical component of data integrity for embedded networked sensing systems. Often fault detection is viewed solely as a component of post-deployment analysis. Instead of identifying faults in real-time, many users assume they can wait until all the data have been collected, discarding faulty data later. This assumption is flawed for two reasons. First, it is not always easy to tell which data are faulty once the collection process is complete. Researchers may need specific information about the context (e.g., an irrigation event occurred at 3PM during the data collection), or need to take physical measurements (e.g., extracting physical samples to validate the sensor data) to determine if the sensor data are faulty. If scientists interact with the network while in the field to perform data analysis and modeling, data quality can be improved significantly. 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. Second, 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.

In addition to detecting faults in real-time, systems must be dynamic. Simple fault detection includes applying statically defined thresholds to data in order to separate good and bad data. This approach is 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 vary by deployment, so users often must set their own thresholds for each new sensor and environment.

Tools to Improve Data Quality. The above lessons are being incorporated into the design of *Confidence*, a system to improve the quality of data collected from large sensor networks. *Confidence* enables field researchers to administer sensors more effectively by automating key tasks and intelligently guiding a user to take actions that demonstrably improve the data quality. The system uses a carefully chosen set of features to group similar data points and to identify actions a user can take to improve system quality. As users take actions and manually validate sensor data, the system adjusts how data are grouped, thus learning to modify parameters for good and faulty data.

Confidence includes tools to annotate data with actions users have taken and to perform other types of data validation. However, this approach is a primitive implementation of a more complete documentation system; much more information is needed to document the context of sensor data collection adequately.

Building an Information Ecology. A set of complimentary tools and services is being developed by CENS to capture sensor data and metadata, which together form a CENS information ecology. These include *Confidence*, described above, to improve the initial data capture in the field, SensorBase to capture, analyze, and visualize data, the CENS Deployment Center (CENSDC) to capture and share information about deployments, and a bibliographic database of CENS publications.

SensorBase provides the sensing research community with a framework for sharing data and for experimenting with models and computation to support data integrity. SensorBase allows for the "slogging" of sensor data directly from the field into the database. Many of its diagnostics and alerting capabilities, leveraging RSS and email, facilitate research by the Integrity Group. Sensorbase acts as a data digital library, but currently lacks metadata crucial for the interpretation of CENS data. SensorBase will rely on in-field tools such as fault detection to increase the quality of data as they enter the database, and will provide other tools to add necessary metadata.

The CENS Deployment Center is a planning tool for documenting field deployments. It attempts to supplement the "oral culture" of deployments through simple interfaces to record equipment requirements, calibration requirements, personnel requirements, and other contextual information, as well as lessons learned from individual field deployments.

The bibliographic database of CENS publications complements SensorBase and CENSDC. Publications traditionally have served as access points to data and as wrappers containing descriptions of equipment, data collection methods, and other information necessary to interpret results. The internal CENS bibliographic database has been ported over to the University of California eScholarship Repository with its own home page [http://repositories.cdlib.org/cens/], greatly improving public access.

Our goal is a tight integration of SensorBase, CENSDC, and the CENS eScholarship repository. CENSDC will document the SensorBase datasets such that deployment records will link to resulting datasets and vice versa. As research results are published, links can be established between the publication, dataset, and deployment records. This tight coupling will establish a rich value chain through the life cycle of CENS data, documentation, and publications [2].

5 Discussion and Conclusions

The early years of wireless sensing research were focused largely on the problems associated with resource-constrained communications, 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 or the integrity of the system itself. As deployment experience increases, data integrity has become a core concern. Researchers now recognize that data and system integrity are limiting factors in scaling these technologies. The focus of data integrity activities has shifted from post-deployment to concurrent processes within deployments. By capturing cleaner data upstream, later problems in identifying potentially errant data are minimized. These techniques facilitate greater trust in those data and enable scientists to analyze data with the assurance that data are complete and of high quality.

A set of complimentary tools and services are being developed to capture sensor data, metadata, and publications, which together form a CENS information ecology. The information ecology described here can be leveraged before, during, and after deployments to collect contextual information, to provide access to an array of information about CENS research, and to follow the life cycle of a research project.

In sum, we are developing an architecture for data integrity and quality in wireless sensing systems. Through interviews, observation, consultation, and systems development, we are learning enough about scientific data practices to build digital libraries that will facilitate data integrity and will improve the ability of current and future researchers to interpret and trust those data. 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 these data are to be useful to the larger scientific community.

Digital libraries can facilitate data integrity by recognizing and accounting for the scientific practices and requirements identified here. Scientists have established methods for describing the network, sensors, and calibrations, but often this information is documented separately from the data, if it is documented at all. Among the many research questions provoked by our research are how digital libraries can store essential contextual information and associate it with relevant data points. Sensor faults have a huge impact on the quality and quantity of data generated by wireless sensing system deployments. Similarly, we are concerned with how sensor fault detection can be reflected in digital libraries. 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. Issues arise such as 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 CENSDC development is provided by Christina Patterson and Margo Reveil of UCLA Academic Technology Services.

References

1. Protein Data Bank. Visited: (October 4, 2006), http://www.rcsb.org/pdb/
2. Borgman, C.L.: Scholarship in the Digital Age: Information, Infrastructure, and the Internet. MIT Press, Cambridge, MA (2007)
3. Traweek, S.: Beamtimes and lifetimes: the world of high energy physicists. 1st Harvard University Press pbk. xv, p. 187. Harvard University Press, Cambridge, Mass (1992)
4. Galison, P.: Image and Logic: A Material Culture of Microphysics. University of Chicago Press, Chicago (1997)
5. Bowker, G.C.: Biodiversity datadiversity. Social Studies of Science 30(5), 643–683 (2000)
6. Bowker, G.C.: Mapping biodiversity. International Journal of Geographical Information Science 14(8), 739–754 (2000)
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. El Abbadi, A., et al. Cairo, Egypt Kaufmann, pp. 693–696 (2000)
8. Hey, T., Trefethen, A.: The Data Deluge: An e-Science Perspective. In: Grid Computing -- Making the Global Infrastructure a Reality Wiley, Chichester (Visited January 20, 2005), http://www.rcuk.ac.uk/escience/documents/report_datadeluge.pdf
9. Bowker, G.C.: Memory Practices in the Sciences. MIT Press, Cambridge, MA (2005)
10. Zimmerman, A.S.: New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. Science, Technology, & Human Values (in press)
11. Borgman, C.L., Wallis, J.C., Enyedy, N.: Little Science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. International Journal on Digital Libraries (in press)
12. Isermann, R.: Fault diagnosis and fault tolerance. Springer, Heidelberg (2005)
13. Tolle, G., et al.: A macroscope in the redwoods. In: Sensys, San Diego, CA (2005)
14. Borgman, C.L, Wallis, J.C., Enyedy, N.: Building Digital Libraries for Scientific Data: An exploratory study of data practices in habitat ecology. In: Gonzalo, J., Thanos, C., Verdejo, M.F., Carrasco, R.C. (eds.) ECDL 2006. LNCS, vol. 4172, pp. 170–183. Springer, Heidelberg (2006)
15. 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. Vancouver, BC. Association for Computing Machinery (in press)
16. Lofland, J., et al.: Analyzing Social Settings: A Guide to Qualitative Observation and Analysis, Wadsworth/Thomson Learning, Belmont, CA(2006)
17. Glaser, B.G., Strauss, A.L.: The discovery of grounded theory; strategies for qualitative research. Observations, Aldine Pub. Co., Chicago (1967)
18. Singh, A., et al.: IDEA: Iterative experiment Design for Environmental Applications. CENS Technical Report (Visited January 28, 2007) (2006), http://research.cens.ucla.edu/pls/portal/docs/page/cens_resources/tech_report_repository/spots07_idea.pdf

19. Ramanathan, N., et al.: Investigation of hydrologic and biogeochemical controls on arsenic mobilization using distributed sensing at a field site in Munshiganj, Bangladesh. in American Geophysical Union, Fall Meeting. (Visited June 6, 2007) (2006), http://adsabs.harvard.edu/abs/2006AGUFM.U41B0823R
20. Ramanathan, N., et al.: Designing Wireless Sensor Networks as a Shared Resource for Sustainable Development. In: Information and Communication Technologies and Development (2006)
21. Chen, G., et al.: Sharing Sensor Network Data. CENS Technical Report. (Visited January 28, 2007) (2007), http://research.cens.ucla.edu/pls/portal/docs/page/cens_resources/tech_report_repository/share_sn_data.g.chen.pdf
22. Chang, K., et al.: SensorBase.org - A Centralized Repository to Slog Sensor Network Data. In: International Conference on Distributed Networks (DCOSS)/EAWMS (2006)