Whose data do you trust? Integrity issues in the preservation of scientific data

Matthew S. Mayernik
Dept of Information Studies
Graduate School of Education & Information Studies, UCLA
00+1+3102060029
mattmayernik@ucla.edu

Jillian C. Wallis
Center for Embedded Networked Sensing
UCLA
00+1+3102060029
jwallisi@ucla.edu

Alberto Pepe
Dept of Information Studies
Graduate School of Education & Information Studies, UCLA
00+1+3102060029
apepe@ucla.edu

Christine L. Borgman
Dept of Information Studies
Graduate School of Education & Information Studies, UCLA
00+1+3108256164
borgman@gseis.ucla.edu

ABSTRACT

Integrity of content is a generic issue in curation and preservation, but has not been extensively studied in relation to scientific data. Data are now being seen as an important end product of scholarship in themselves. In this paper, we will discuss data integrity issues in relation to environmental and ecological data, and the implications of these issues on the development of data digital libraries. 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. The goal of this research is to identify functional requirements for digital libraries of scientific data, encompassing both technical and social factors that can affect data integrity. Mechanisms to ensure data integrity have to be present at each stage in the data life cycle, from data collection to data preservation and curation. The implications of our research on data integrity are multi-fold for the iSchool research community, and we hope to promote discussion of these issues.

Topics

Information infrastructure development
Information management
Information technology and services
Nature and scope of *i*Schools and *i*Research
Preserving digital information

Keywords

Data integrity, digital libraries, scientific data practices

Copyright and Disclaimer Information

1. INTRODUCTION

Preserving digital information is a central concern to the design of information infrastructure. Digital information takes many forms, and one of increasing importance is scientific data. Data are much less "self-archiving" than are publications, however, and often require more human effort to describe and to provide context for interpretation [5, 10]. When engaged in their own research, scientific researchers take steps to ensure the quality and integrity of the data they capture. However, it is much more difficult for researchers to evaluate data that are collected by others because of the lack of standardization in the ways that data are documented and annotated. These standards are necessary to provide enough information to enable reuse of data by others.

Additionally, researchers will only use data collected by others if they are confident in its quality and integrity. The lack of sufficient methods for ensuring data integrity is a serious impediment to the establishment and use of data repositories in many sciences. In this paper, we will discuss data integrity issues in relation to environmental and ecological data, and the implications of these issues on the development of data digital libraries.

2. PROBLEM STATEMENT

Researchers want to capture, manage, and store their data in ways that assures its integrity. Similarly, librarians, archivists, and future users of those data want to be assured of the integrity of the data over time. But what does "integrity" mean in this environment, and to whom? We draw upon our research on data practices conducted at a large, collaborative, multi-disciplinary science and technology center to compare the notions of "integrity" held by different research participants to suggest implications for iSchool-based research on information infrastructure and on digital preservation [24].

2.1 Scientific Data Practice Research

Research on scientific data practices has concentrated on big science such as physics [13, 22] or on large collaborations in areas such as biodiversity [7, 8, 9]. Equally important in understanding scientific data practices is the study of 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 and complexity 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, opening up new directions for research questions and methods. However, because the use of wireless sensing systems in environmental and ecological research is a relatively new phenomenon, data management techniques for data from such systems are largely local and idiosyncratic in nature.

2.2 Data Standards

Several XML-based standards and protocols exist for this diverse community, but none of them are stable or widely adopted. Structures most relevant to embedded sensor network data in the environmental sciences include the Ecological Metadata Language, supported by the Knowledge Network for Biocomplexity [11, 16], and the Sensor Modeling Language (SensorML), supported by the Open Geospatial Consortium, which describes sensor network equipment and relationships. SensorML is complemented by the Observations and Modeling (O&M) language to express ecology data captured by the sensor network. SensorML and O&M are in the final stages of being accepted as formal standards [23].

The multiplicity of standards in this field poses significant challenges to researchers and has limited the widespread implementation of any individual standard. As the use of complex instrumentation such as wireless sensing systems becomes more widespread, environmental scientists face the questions of what to standardize, when, and for what purposes. Attempts to develop data digital libraries for these domains are hindered by this lack of standardization.

2.3 Data Digital Libraries

Data digital libraries are only as valuable as the data they contain. When data are submitted to prestigious repositories (e.g. [21]) they are evaluated rigorously. This is not the case with data that are made available through local websites or local repositories, where mechanisms for data authentication are less consistent or entirely non-existent. Digital library systems not only must be able to integrate disparate data sets that were collected with different research questions in mind, but must also ensure the integrity and usability of those data.

The diversity of research questions, techniques, and instrumentation complicates the long-term preservation and access to data in data digital libraries. Scientific 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, the context in which the data were collected, and local subtleties and quirks.

A prime goal of data digital libraries is to facilitate widespread use of data by any interested user. What can digital libraries systems do to ensure data integrity? Researchers (or teachers or students) who wish to reuse data rely on a variety of indicators when assessing data collected by others. Indicators include the reputation of the data collector and the institution, the quality of papers reporting the data, instrumentation descriptions and specifications, and any accompanying documentation. When these indicators are not available to users, the task of assessing data integrity is much more difficult.

2.4 Assessing Data Integrity

Assessing data integrity is especially difficult in the environmental sciences, where data practices vary widely from project to project. Ecological research questions are typically focused on specific locales or on particular types of phenomena. Because of this, research methods and instrumentation varies widely between projects. Even multi-site comparatives, such as the Long-Term Ecological Research program (LTERs), are not immune from this diversity, as the highly unpredictable and variable nature of in-field ecological research demands flexibility.

Karasti, Baker, & Halkola describe their work in the LTER program [15]. They have found that it is necessary to give researchers multiple paths and incentives for producing metadata and documentation. Rapid technological change requires the LTER program to take a science-driven view of data management, rather than the technology-driven view that motivates most discussion around e-Science data curation (where the focus is on the "digital obsolescence" problem and digital preservation techniques like migration and emulation) because data regularly outlive the technology used to create them. They emphasize an overall framework for "data stewardship", involving "data definitions, data requirements, and quality assurance as well as user feedback, redesign, and data exchange" (pg. 352).

Researchers in the ecological and environmental sciences must be able to design customized solutions to local and unforeseen challenges. Because of this, assessing the integrity of data is a complex and difficult task.

3. RESEARCH CONTEXT

To examine these issues, we are studying the data practices of researchers within 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 and to apply this technology to address questions in a variety of scientific areas, including terrestrial 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.

Data integrity issues are becoming more salient within CENS as the new wireless sensing technologies and the scientific applications using those technologies have each matured. In the initial stages of CENS research, the primary focus was developing new wireless sensing system technology. As is typically the case with new technology development, the first few generations of systems were often unreliable, sometimes producing data with scientific value and other times not. CENS has now reached the point where the nascent technology is consistently producing data of real scientific value, and as such must now take means to address long-term data preservation and access issues. Additionally, questions about the integrity of data are coming to the fore as more useful data are being produced.

Until recently, CENS relied on a largely oral culture for the exchange of data, and for the exchange of information about how data are collected, such as 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 many more research teams. Future uses of CENS data depend on identifying and implementing solutions for data preservation and access challenges

4. BACKGROUND ON DATA INTEGRITY

CENS data result from sensor deployments where researchers deploy sensors and wireless communication equipment in real-world field locations. Assessing the integrity of data captured during real-world sensor deployments is a complex task, encompassing the entire data life cycle: experimental design, equipment calibration, data capture, cleaning, derivation, integration, analysis, publication, and preservation [25]. The following sections illustrate how the task of ensuring data integrity is both a technical and a social process [24].

4.1 Technical Data Integrity Issues

Within CENS, a group of computer scientists, statisticians, and electrical engineering researchers have formed the "CENS integrity group" devoted to looking at the technical aspects of wireless sensing system data integrity. The CENS integrity group focuses on sensor fault detection - that is, detecting when sensor output does not accurately represent the phenomena being measured. Fault detection is often viewed as a step in post-deployment data analysis, where data is rejected during analysis if it is deemed to be faulty. However, this approach is flawed for a couple of reasons. First, it may not be possible to tell what data are faulty after the fact, and second, the amount of data available after a challenging deployment may be so small that none can be spared. This is particularly the case on experiments where sensors are short-lived and require frequent calibration. CENS researchers have had to discard as much as 40% of the data collected on a particular deployment, limiting the amount of scientific analysis possible.

For these reasons, the CENS integrity group is researching fault detection methods that would identify sensor faults as data is being captured in the field. There are many challenges in this task, as sensor faults can take many forms, such as out of range values, and have a number of possible causes, each requiring a different detection method [19]. Often faults compound each other, such as when a failing battery causes a sensor to give faulty readings. This technical data integrity work is important because it indicates that the embedded networked sensing community is becoming aware of the complexity of data integrity issues, even if the mechanisms and techniques they are developing for automated and human-mediated data integrity checks in sensor systems are in the very beginning stages. Once these mechanisms mature, they will be an integral component of scientific research that uses sensor networks.

4.2 Social Data Integrity Issues

Sensor faults represent a technical impediment to ensuring data integrity. Equally important social, cultural, and economic impediments must also be addressed. Birnholtz and Bietz [4] point out that many factors play into data sharing practices. Understanding a data set requires knowledge of the context in which it was collected. Documenting the data collection process is challenging, however, because much of the knowledge that goes into collecting and interpreting a data set is tacit. Knowledge transfer “is not simply a matter of sharing a set of instructions, but is a highly social process of learning practices that are not easily documented" (pg. 340). They identify three recommendations for the design of systems to share scientific data: 1) Support social interaction around data abstractions and the data themselves, 2) Do not rely on metadata alone, it is also necessary to support the sharing of supplementary materials that enhance the value of the data, and 3) Support social and scientific roles of data. Because tacit knowledge is such an integral part of ecological research, producing data set documentation is time-consuming and labor-intensive process.

4.3 Cultural Data Integrity Issues

Data collections in the environmental and ecological sciences are primarily research data collections [18] that were collected for a specific purpose, and are typically held by the researchers who created them [26]. It is therefore often difficult for researchers to discover and access research data that was collected by someone else. As Zimmerman notes, ecologists face a litany of challenges in finding data. Because few ecological databases exist, the ecologists Zimmerman studied looked to many other sources for data, including peer-reviewed publications, museums, and personal contacts. Once relevant data was found, they went "to great lengths to ensure that they understand data collected by others" [26, pg. 9], relying on their own field expertise to evaluate the potential problems with given data.

Additionally, when combining disparate data sets, researchers often ignored small differences between data sets, with the hope that small amounts of bad data or incompatibilities between data sets will not affect how they can be analyzed. Peer-reviewed literature are generally a good source of information about data, but peer-review typically does not certify data quality per se, as publications rarely report the actual raw data on which they are based. Similarly, databases are challenging to use because they often do not provide enough information to determine the purposes for which data was collected.

Despite the challenges, scientists find ways to get access to the data they need [26]. However, tremendous amounts of time and energy are spent in the process of finding, evaluating, and integrating data. Better methods for data documentation and access would make it much easier for scientists to discover and share data, and, just as importantly, evaluate the integrity of data collected by others. Understanding the requirements for such methods is the focus of our research.

5. RESEARCH METHODOLOGY

We have identified a number of critical research areas regarding CENS data practices, and are pursuing a multi-pronged research approach. We recently completed a series of interviews with scientists from 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. Interview questions were grouped into these four categories: data characteristics, data sharing, data policy, and data architecture. The interviews were audiotaped, transcribed, and complemented by the interviewers' memos on topics and themes [17]. Analysis proceeded to identify emergent themes. This study used the methods of grounded theory [14] to identify themes and to test them in the full corpus of interview transcripts and notes.

The next planned stages of our research are based on findings and gaps identified by this interview study, and encompass ethnographic research, information system design, and quantitative characterization of CENS collaborative activities. We are developing another interview study that will focus on data versioning and provenance within CENS research. This study is in the development stages, but will seek to determine in what states or versions data sets exist, and how these states are documented. Additionally, we are examining ways that wireless sensing information and data can be organized through the development of taxonomies, ontologies, or metadata models.

6. DISCUSSION - IMPLICATIONS FOR DATA DIGITAL LIBRARY DESIGN

Digital libraries can facilitate data integrity by recognizing and accounting for the scientific practices and requirements. Scientists have established methods for describing their communication networks, sensors, and equipment 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.

6.1 The Context of CENS Data Collection

Within CENS, data collection largely takes place on real-world sensor deployments. We have designed and implemented the CENS Deployment Center (CENSDC), a database for CENS deployment information. The CENSDC provides a centralized web-accessible location for researchers to articulate and document deployment activities through the creation of pre-deployment plans and post-deployment feedback/notes. CENS researchers collaboratively plan sensor deployments prior to going out in the field. Information captured in deployment plans includes deployment dates, locations, participants, technology, equipment, tasks, and other notes. The system also facilitates post-deployment information capture, by providing a series of fields for researchers to outline any problems they encountered while in the field, as well as recommendations and suggestions for future deployments.

The development of the CENSDC has followed an iterative rapid-prototyping design process. Requirements for the system were developed through ethnographic study of CENS deployments and discussions with CENS researchers to characterize deployment practices. Much of the in-field data collection process during deployment involves tacit knowledge about equipment setups, deployment locations, and field preparations from past deployments. As CENS technologies mature and current researchers gain deployment experience, new students face a steeper learning curve when joining a project. The CENSDC was designed to capture tacit knowledge and contextual information surrounding in-field data capture, and to serve as a tool for transfer of common knowledge – knowledge gained through field experiences that can be utilized for future deployments. The CENSDC adds value to CENS data by providing a source of contextual information surrounding the data collection. This information can assist researchers in writing papers, proposals, and reviews, as well as in maintaining their data and leveraging them for reuse by others.

6.2 Data Integrity and the Data Life Cycle

We have proposed that digital library services should serve scientists whose data exists in all stages of the data life cycle [6, 25]. Building on this model, mechanisms to ensure data integrity will also have to be present at each stage in the data life cycle.

Prior to data capture, equipment are tested and calibrated. 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 the level of granularity in the calibration information needs to be associated with each data set. At the point of data capture, it will be essential that data digital library implementations accommodate and (ideally) incorporate automated or human-mediated data integrity checks as data are being collected. Sensor faults have a huge impact on the quality and quantity of data generated by wireless sensing system deployments, and researchers must be able to indicate the presence of these faults and their impact on the resulting data.

Similarly, contextual information around the data capture, such as the equipment and software used, is critical to evaluate data. Often this information is not documented on data. Contextual information about the data collection process is particularly important when evaluating data collected by someone else. During the data analysis phase, data sets often undergo changes as scientists clean, integrate, and analyze data. The provenance of these changes is relevant when assessing the integrity of the resulting data, as it may be necessary to backtrack through the data analysis steps that led to a research claim.

Publications are currently the main product of scientific research, but as data and other scientific information is increasingly available online, it would be very valuable to be able to identify relationships between resources, making the scholarly value chain explicit [20]. Finally, curating data and providing effective data stewardship is essential if data are to be used and re-used in the future.

7. CONCLUSION - IMPLICATIONS FOR ISCHOOL RESEARCH

The implications of our research on data integrity are multi-fold for the iSchool research community, and we hope to promote discussion of these issues. Data are a growing component of the scholarly information infrastructure and must be integrated into larger discussions of technology, institutions, practices, and policy [2, 3]. iSchool research has focused much more on documents than on data. Techniques that have been effective in promoting access and interoperability of documents may not be applicable to data and other digital scientific resources. Research relating to scientific data practices and data preservation and curation are small but growing areas of iSchool expertise. The development of a larger research base in these areas is critical to enhance our understanding of the cyberinfrastructure "blank canvas" [12], and to facilitate the development of a trained workforce of data scientists [18].

Integrity of content is a generic issue in curation and preservation, but has not been extensively studied in relation to scientific data. As data are increasingly being made available on the internet, questions about ensuring data integrity across all stages of the data life cycle must be answered. Data that are collected on by ecological and environmental researchers have scientific value both to immediate research questions and long-term longitudinal studies. Distributed longitudinal studies will require standards that ensure the interoperability of the sensor, the network, and the data [1]. These concerns are coming to the fore in data practices; lessons learned here will apply across the scholarly information infrastructure.

8. 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. Development support for the CENSDC provided by Margo Reveil, Christina Patterson, and Andrew Tai of UCLA Academic Technology Services.

9. REFERENCES

[1] Arzberger, P. 2004. Sensors for environmental observatories. Report of the NSF sponsored workshop, December 2004, WTEC, Baltimore, MD. http://www.wtec.org/seo/final

[2] Arzberger, P., Schroeder, P., et al. 2004. An International Framework to Promote Access to Data. Science, 303: 1777-8.

[3] Atkins, D., et al. 2003. National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, Revolutionizing Science and Engineering through Cyber-infrastructure. http://www.communitytechnology.org/nsf_ci_report/

[4] Birnholtz, J. P. and Bietz, M. J. 2003. Data at work: supporting sharing in science and engineering. In M. Tremaine (ed.): GROUP '03. Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, 2003 November 9 to 12, 2003. ACM Press, pp. 330-348.

[5] Borgman, C. L. 2007. Scholarship in the Digital Age: Information, Infrastructure, and the Internet. Cambridge, MA: MIT Press.

[6] Borgman, C. L., Wallis, J. C., Mayernik, M., and Pepe, A. 2007. Drowning in Data: Digital Library Architecture to Support Scientists' Use of Embedded Sensor Networks. JCDL '07: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, Vancouver, BC. Association for Computing Machinery.

[7] Bowker, G. C. 2000a. Biodiversity datadiversity. Social Studies of Science, 30(5): 643-683.

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

[9] Bowker, G. C. 2000c. Work and information practices in the sciences of biodiversity. VLDB 2000, Proceedings of 26th international conference on very large data bases, Cairo, Egypt. Kaufmann: 693-696.

[10] Cyberinfrastructure Vision for 21st Century Discovery. 2007. National Science Foundation. http://www.nsf.gov/pubs/2007/nsf0728/

[11] Ecological Metadata Language. 2004. http://knb.ecoinformatics.org/software/eml/

[12] Freeman, P. A., Crawford, D. L., Kim, S., and Munoz, J. L. 2005. Cyberinfrastructure for Science and Engineering: Promises and Challenges. Proceedings of the IEEE, 93(3): 682-691.

[13] Galison, P. 1997. Image and Logic: A Material Culture of Microphysics. Chicago: University of Chicago Press.

[14] Glaser, B. G. and Strauss, A. L. 1967. The discovery of grounded theory; strategies for qualitative research. Observations. Chicago: Aldine Pub. Co. x, 271.

[15] Karasti, H., Baker, K., and Halkola, E. 2006. Enriching the notion of data curation in e-Science: Data managing and information infrastructuring in the Long Term Ecological Research (LTER) Network. Journal of Computer Supported Cooperative Work, 15(4): 321-358.

[16] Knowledge Network for Biocomplexity. 2004. http://knb.ecoinformatics.org/index.jsp

[17] Lofland, J., et al. 2006. Analyzing Social Settings: A Guide to Qualitative Observation and Analysis. Belmont, CA: Wadsworth/Thomson Learning.

[18] Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. 2005. National Science Board. http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf

[19] Ni, K., et al. in review. Sensor Network Data Fault Types.

[20] Pepe, A., Borgman, C. L., Wallis, J. C., and Mayernik, M. S. 2007. Knitting a fabric of sensor data resources. ACM/IEEE Information Processing in Sensor Networks Workshop on Data Sharing & Interoperability, Cambridge, MA.

[21] Protein Data Bank. 2006. http://www.rcsb.org/pdb/

[22] Traweek, S. 1992. Beamtimes and lifetimes: the world of high energy physicists, (1st Harvard University Press pbk. ed.). Cambridge, Mass.: Harvard University Press.

[23] VAST. 2008. Introduction to SensorML. http://vast.uah.edu/SensorML/

[24] Wallis, J. C., Borgman, C. L., Mayernik, M., Pepe, A., Ramanathan, N. and Hansen, M. 2007. Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries. 11th European Conference on Digital Libraries, Budapest, Hungary. Berlin: Springer. LINCS 4675: 380-391.

[25] Wallis, J. C., Borgman, C. L., Mayernik, M., and Pepe, A. 2007. Moving Archival Practices Upstream: An Exploration of the Life Cycle of Ecological Sensing Data in Collaborative Field Research. 3rd International Digital Curation Conference, Washington D.C.

[26] Zimmerman, A. 2007. Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. International Journal of Digital Libraries, 7: 5-16.