An Exploration of the Life Cycle of eScience Collaboratory Data

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

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

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

ABSTRACT

The success of eScience research depends not only upon effective collaboration between scientists and technologists but also upon the active involvement of information scientists. Archivists rarely receive scientific data until findings are published, by which time important information about their origins, context, and provenance may be lost. Research reported here addresses the lifecycles of data from ecological research with embedded networked sensing technologies. A better understanding of these processes will enable information scientists to participate in earlier stages of the life cycle and to improve curation of these types of scientific data. Evidence from our interview study and field research yields a nine lifecycle phases, and three types of lifecycle depending on the research goal. Findings include highlighting the impact of collaboration on the research processes and potential phases during which the integrity of the captured data is compromised.

Topics

Cultural information systems, Information infrastructure development, Information management, Preserving digital information

1. INTRODUCTION

The success of eScience research depends upon effective collaboration between scientists and technologists. Partners often must learn how to produce data that are meaningful to participants from multiple disciplines. Many decisions are made about data at each stage in its life cycle. Curation of these data and their value for reuse depends heavily on how much is known about their origins, derivation, and provenance.

Archivists typically receive scientific data only after the findings of a study are published or after a researcher retires. Neither of these archival outcomes provides access to scientific data in a timely manner. More importantly, by the time that archivists receive data, much of the information necessary for future interpretation may have been lost. Shifting the practices of archiving such as appraisal, curation, and tracking provenance into earlier stages of a given material's life cycle can increase the likelihood of capturing reliable, valid, and interpretable data [1] and thus improve both short- and long-term access and interpretation.

To determine how early these archiving processes might begin, it is necessary to identify the life cycle of a given type of data. eScience partners often have different responsibilities at each stage of a life cycle. Individual researchers may be insufficiently aware of how others have acted upon the data, or how others may use or interpret the data further down the line. Making the entire life cycle of data more transparent and self-documenting has the potential to simplify data capture, management, interpretation, and curation for all parties involved [2, 3]. Some stages can be augmented by technical means, such as automated tools to identify potential instrumentation errors as they occur. Other stages can be made more transparent by identifying and documenting scholarly practices associated with the data.

The life cycle of business and government documents is characterized by each stage being handled by a different party. The life cycle of data from little science – that is, science performed by an individual or small research group – is characterized by all of the phases being handled by one or a few persons with similar domain knowledge and training. The life cycle of data from big science – that is, science performed by a large number of researchers, such as high-energy physics – is characterized by many researchers participating in each stage of the life cycle. These researchers all have similar domain knowledge and training. In the research reported here, researchers from multiple disciplines play complementary (and sometimes conflicting) roles in data handling.

In keeping with the scientific data research agenda for the next decade set by the Warwick Workshop [4], our goals are to develop: a) more detailed data models for each domain, including intra-domain and inter-domain commonalities, b) automatic processes for data and metadata capture, and c) consistent methods of data description in this scientific and technical environment. Our exploration of the life cycle of scientific data identifies the stakes and stakeholders at each phase to develop a "digital curation infrastructure" [5] that will support the use, re-use, access, and interpretation of ecological sensing data. In this context, we need to understand the processes that lead to the creation, analysis, and publishing of said data for metadata capture, and when major changes occur to data so that we can

build appropriate provenance tracking measures. Born-digital objects leave no physical residue that can be referenced later; too often, useful information is discarded before being properly assessed for archival value [6].

2. BACKGROUND

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, 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.

Our data management research group has been part of CENS since its inception. While few scientific data were generated in the early years, we were planting the seeds of archival practice and preservation. Once data captured by CENS' instrumentation became relevant to our application scientists, we took a more active role in building the necessary infrastructure for long-term access. Our initial research focused on defining what were "data" in this environment. Now that we understand better what are data to whom and when, we are addressing larger data life cycle issues.

2.1 Deployment Scenario

An example of a CENS embedded networked sensing system deployment will provide context for the life cycle of CENS data.

CENS researchers utilize several deployment models. Along with static deployments typical of observatories such as NEON or GEON [7, 8], CENS researchers regularly go on short-term deployments, or "campaigns," where sensing systems are deployed in the field for a few days. Among the benefits of this approach for exploratory research are compatibility with the data collection practices of application science researchers (most are in biology or environmental sciences), the ability to field test delicate and expensive experimental equipment, and the opportunity for science and engineering researchers to work together in the field to trouble-shoot technical problems and improve the overall quality of data.

An example of a CENS deployment is the study of biological processes associated with harmful algal blooms. In designing a deployment, the application science researchers (biologists in this example) identify a viable research site, in this case a lake known for summer blooms. Available background information about the lake includes peak months for algae, a topology of the lakebed, local species of phyto and zooplankton, and nutrient presence and concentration. The engineering researchers determine which equipment are most appropriate for capturing the data desired by the scientists.

Prior to going in the field, the team calibrates equipment in the laboratory based on knowledge of the types of organisms likely to be present in the water. Because of the natural variation of water organisms, calibrations will be augmented with physical water samples taken adjacent to sensors. A "wet lab" will be set up on site to process water samples. Once on site, the team deploys sensors in the lake using static buoys that house a power source, data logger, and wireless communication system. They document GPS coordinates of each buoy, times of placement, and serial numbers of each sensor in a laboratory notebook.

The data collection process is a combination of pre-planned activities and in-field decisions. Because the aquatic phenomena of interest vary on diel or 24-hour cycle, scientists take data for a full 24 hours. Once sensors begin to report data, researchers begin observing interesting phenomena, such as that the water flows more quickly at one end of the lake, and that the water is greener and at a higher temperature where a rock slows the flow. Based on such information, the team may change the data collection strategy, altering plans for sensor placement or for hand collection of water samples. At the end of a deployment, equipment is removed and returned to the lab. Water samples are processed for organism identification and concentration and for nutrient concentrations. Sensor data are compared to the in-lab and in-field calibration curves and to other trusted data sources. Only then are water sample data and sensor data integrated for analysis. After data analysis is complete and papers are published, numerical data are burned to DVDs and shelved with other data. Any remaining water samples are put in cold storage.

2.2 CENS Data, Users, and Uses

Data from CENS' dynamic field deployments can be grouped into four types. Sensors are used to capture numerical data on 1) the scientific application, 2) the performance of the sensors themselves, and 3) proprioceptive data to use in navigation for robotic sensor technology. The fourth category is hand-collected data for the scientific application, such as the water samples described above in the deployment scenario. Each of the four data types has multiple variables; from temperature and barometric pressure to roll, pitch, and yaw to packets sent and received. Some data serve only one purpose, but most serve multiple purposes for instance the scientific data collected can be used to identify sensor faults [9].

When we asked our subjects about capturing, using, sharing, and preserving data from deployments, and about capabilities they desired in archives to support their data, the primary interest was in the scientific data. Computer science and engineering researchers were as concerned about the quality and accessibility of scientific data as were the domain scientists. Conversely, the computer science and engineering researchers took little interest in maintaining access to sensor performance data or proprioceptive data that are essential to their own research. These forms of data appear to serve transient purposes for these researchers, with minimal archival value. However they may be essential for re-use of the application science data by others.

3. METHOD

Our research questions address the initial stages of the data life cycle in which data are captured and subsequent stages in which the data are cleaned, analyzed, published, curated, and made accessible. The interview questions were divided into four categories: data characteristics, data sharing, data policy, and data architecture. This paper reports our further exploration of the scientific data life cycle based on responses to questions about data characteristics and architecture. Findings on other questions are reported elsewhere[10-16].

The findings reported here are drawn from an interview study of five environmental science projects and subsequent field

observations. For each project, we interviewed a complementary set of science and technology participants, including faculty, post-doctoral fellows, graduate students, and research staff. CENS is comprised of about 70 faculty and other researchers, about 120 student researchers, and some full-time research staff who are affiliated with the five participating universities. Our pilot ethnographic study consisted of in-depth interviews with two participants, each two to three hours over two to three sessions. The intensive interview study consisted of 22 participants working on the five ecology projects. Interviews were 45 minutes to two hours in length, averaging roughly 60 minutes.

The interviews were audio-taped, transcribed, and complemented by the interviewers' memos on topics and themes. Transcription totaled 312 pages. Analysis proceeded to identify emergent themes. We developed a full coding process using NVivo 2, which was used to test and refine themes in coding of subsequent interviews. This study used the methods of grounded theory[17] to identify themes and to test them in the full corpus of interview transcripts and notes.

In addition to the interviews, members of our research team have attended numerous field deployments with various CENS research groups. Participant observation was used to gain entree to these work-intensive daylong to weeklong data collection events. While assisting the researchers in their deployment of equipment and data capture we were able to discover the processes leading up to the deployment and the plans for use of the data captured.

User scenarios for how data were captured, processed, and published were extracted from the interview and observational data. These scenarios were used to construct a data flow model, including the data sources, level of derivation, and any computer programs or scripts that were used to transform the data. From the combined flows we were able to extract common procedures and generalize them across our participants. We then verified this life cycle model during our interactions with researchers after the interviews, or observations of their data collections efforts.

4. RESULTS

Described here are both the general lifecycle phases and the three data lifecycles present within CENS research. Initially our research lead us to model a unified data life cycle that applied to all of CENS research. This lifecycle model in described in more detail along with the charge to evaluate the model for generalizability across the center and to shine alight on some of the more entangled phases (DCC). In teasing out the variation present in two of the phases of the lifecycle the generalization broke. There was a clear distinction between what happened to the data created as a part of technology research and those data created when science and technology researchers worked together, and these were both different from the data created by scientific research. These three types of lifecycle are characterized by what the research is meant to accomplish, be that scientific research, technological research, or development of technology for science. This latter is when both science and technology researchers collect data in the field, with the technologists responsible for the equipment or systems being tested and the scientists responsible for collecting other data to make the sensor data worthwhile.

4.1 General Life Cycle Phases

We have identified nine stages that appear to be common to the CENS deployments studied, the researchers, and to the resulting data. The order of the steps is not absolute, as some stages are iterative while others may occur in parallel. For instance, Phases 4-6 appear to happen concurrently, with new outliers emerging only when the data has been integrated

1) Experiment Design. The beginning of the data life cycle is the design of new experiments. CENS researchers design new experiments by reusing data from prior research.
2) Calibration and Setup. Before sensors are deployed, they are calibrated to known solutions or values to identify the offset between the actual measurement and the expected measurement. They are calibrated again in-field, a process referred to as "ground-truthing".
3) Capture or Generation. Once sensors have been deployed successfully in the field, researchers begin to collect observations of physical phenomena. Some sensor measurements are direct (e.g., temperature, wind speed) and others are indirect (e.g., measure of fluorescence as an indicator of chlorophyll activity).
4) Cleaning. After data have been captured, calibration and ground-truthing information need to be applied to the data to normalize any calibration offsets from the sensing equipment.
5) Integration. Few of the observations and samples collected in the field can be interpreted without derivation into more meaningful data points. Data typically must be averaged into composite points before they can be used in analysis.
6) Derivation. Researchers are looking for trends over time and across spatial locations. Datasets each given deployment are integrated by multiple researchers, for multiple reasons, and in multiple combinations.
7) Analysis. Researchers use statistical, modeling, and visualization tools that vary by research specialty and individual preference. They test and generate hypotheses and draw conclusions about data obtained from the deployments.
8) Publication. Data collected during embedded network sensor deployments culminate in scholarly publications such as journal articles, conference papers, posters, and technical reports.

Preservation. Few, if any, of the CENS researchers interviewed had data preservation strategies commensurate with those of the archival community. It is more accurate to say that they back up their data.

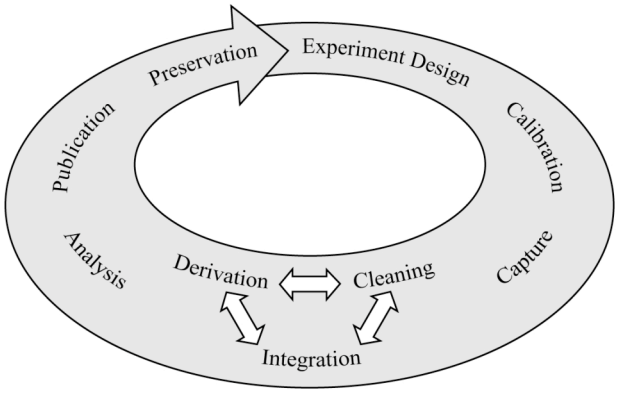

Figure: Life cycle of CENS data.

4.2 Three Lifecycles Model

Table: Three lifecycles with sample tasks for each phase of each cycle

Phase/Cycle	Scientific Research	Sci-Tech Development	Tech Research
Experiment Design	Generate hypothesis; develop methods; choose equipment; plan sampling schedule	Negotiate researchable questions; choose equipment and personnel; schedule tasks	Generate hypothesis; develop methods
Calibration/Setup	Calibrate equipment; collect ground truth samples	Calibrate sensing systems; ground-truthing	Prepare model, data, or algorithm to be used
Data Capture/ Generation	Hand sampling; observation; processing samples	Sensor collection; hand sampling; observing environment; tweaking systems; observing users; checking in across groups	Sensor collection; generating from models; creating new data by running algorithms over data; creating models from data
Cleaning, Integration, & Derivation	Analysis of samples; recording presence and frequency/volume of organisms or chemicals; comparing to environmental models; remove outliers	*Part I: Tech* Remove sensor artifacts; synch time stamps; recalibrate; aggregate data by variable; derive data for compound measures *Part II: Science* Sample analysis; recording presence and frequency/volume of organisms or chemicals; ground-truthing based on hand samples; comparing to environmental models; removing outliers	Debug; investigate error reports; retesting; pass code around to get additional opinions
Analysis	Linear regression of variables captured; hypothesis testing	Visualization; hypothesis testing	Comparisons; regressions; evaluation
Publication	Publish conclusions in science journals; post or reposit data	Publish conclusions in science journals and technical proceedings; post data	Publish conclusions in technical proceedings; post data
Preservation	Refrigerate samples; numerical data kept in databases; printed for hard copies; filed	Refrigerate samples; numerical data kept in databases; move files to a lab server or local machine	Move files to a lab server or local machine

In these models we have combined the formerly separate Cleaning, Integration, and Derivation phases because they are so interrelated. In order to properly remove outliers, the data must be derived or integrated, thus these three phases are really separate tasks that are performed iteratively until the data is clean enough for analysis to begin.

5. DISCUSSION

Technology research has the luxury of not always needing to go out into the field. Much of their research involves testing models or algorithms, generating data from these, and evaluating the performance, none of which requires fieldwork. At the same time the data generated in the lab may be wildly different from those captured in the field, and the occasional reality check is necessary to maintain a clear heading. When the technologists perform fieldwork, they do not need the scientists present to capture data that will progress their research, because their test subject in these cases are the equipment and systems themselves. Similarly the scientists do not need to bring the technologists along for their own data collection efforts, unless they want to use the sensing equipment.

Many Technology research cycles happen between in-field deployments, just as many in-field technology deployments happen between a Science-Technology Development deployment. For scientists on the other hand the frequency of data collection cycle is much lower, with most cycles lasting a year or more. These science research cycles carry a much higher investment for the data collected. The interplay of the cycles are negotiated based on time available and need. Coordinating larger deployments to collect data that is meaningful to both the scientists and the technologists is a significant investment of time and resources. Difficult as they are these are the rare opportunities to see the "users" using the instruments and the systems, allowing the technologists to begin the next iteration of the design.

Comparing these three data lifecycles reveals the affects of collaboration on routine research by either the science or technology researchers. The experiment Design phase is much different during the collaborative cycle, attempting to offer significant research opportunities to both parties. The data capture phases are very different for each cycle; during the science research cycle the emphasis is on hand collected samples, the collaborative cycle is concerned more with the capture of sensor data than the hand sample data even though this data is significant to the science researchers and acts as a ground truth mechanism for the sensor data, and the technology research cycles range from collecting sensor data to generating data from a model. The cleaning, integration, and derivation phase is similarly different across all three cycles, but in this case the significant part is the way the custody of the data is no longer shared. The technology researchers must first make sense of the captured data in reference to the technology and then the science researchers will make sense of them in reference to the science. The final phase to display

divergence is the publication phase, where collaboration forces both the scientists and the technologists to publish in one another's domains. These publications do not carry the same weight as those within their own domain, thus dis-incentivising collaboration.

These affects caused by collaboration appear to be trouble spots where the future value of the data is compromised, due to variations in methods and processes across collaborating disciplines. For instance, the use of linear regression for data analysis does not scale to the volume of sensor-captured data, thus scientists must rely on technologists for new methods of interpreting their data. Linear regression is the established method for ecological data analysis, and has been so for multiple decades. The science researchers in this case are really dependent on the tools being built for them to be as trustworthy as linear regression and stand up to peer-review. Another example of this is the variation in what is considered data by each group, for instance the technology researchers will consider their algorithm or the script developed to be data, whereas the scientists will consider the hand sample or the sensor readings to be data. This may opens gaps where one party is not as invested in the overall quality of a specific data type because it is not central to their own discipline.

Actions taken at each stage of the life cycle influence how the resulting data can be interpreted, hence it is important that these stages be documented and associated with the resulting dataset. There is a cumulative effect of decisions made at each stage of the life cycle. For example, decisions made in the experimental design stage determine what data exist for analysis, or calibration decisions are essential to interpreting the data. There is a delicate balance of decision making between scientific and technology research partners.

6. CONCLUSIONS & FUTURE RESEARCH

The success of eScience depends upon successful collaboration between application scientists and their partners in computer science and engineering. Data resulting from such collaborations is expected to be extremely valuable for reuse by others. However, the value of data for reuse depends upon the quality of those data, which in turn depends on the ability to interpret the origins, provenance, and context of the data. Surprisingly little is known about how data arises from eScience collaborations. Our case study of ecological research in the Center for Embedded Networked Sensing sheds light on the various life cycles of eScience data.

Our future research will continue to explore and refine the data life cycle identified here, and to build systems to support it. In order to determine how generalizable this three-cycle model is across the various research groups, we need to test this model in the field. Additionally we would like to understand the role of publication in collaboration, specifically if the willingness to co-author is skewed towards either the science or technology researchers by exploring whether scientists are more apt to publish in technology papers, or the reverse. Given our access to the co-authorship data from this population answering this question is a logical next step. We are also curious about the role of decisions made in the field on the calibration and capture interplay, as well as the experiment and design interplay. At present, much of the sensing technology is experimental, but commercial off-the-shelf sensors also are in use. Research questions about data provenance will evolve as the technology stabilizes and the scientific research questions broaden.

7. IMPLICATIONS FOR iSCHOOL

Our interest in understanding information from every angle should lead us to studying the processes and methods that lead to the creation of data. Scientific data represents the entire evidentiary basis for scientific information and knowledge. Thus without studying the life cycle of data we are ignoring the premise of scientific information. Research such as this cuts to the heart of the disciplinary tributaries that comprise iSchool, namely Informatics, Archives, and Digital Librarianship, while also falling under Science, Technology, and Society. This is the practical application of learnings from each of these fields and serves as an example of what will advance iSchool education and research agenda.

8. ACKNOWLEDGMENTS

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.

9. REFERENCES

[1] Esanu, J.M., et al., Selection, Appraisal, and Retention of Digital Scientific Data: Highlights of an ERPANET/CODATA Workshop. Data Science Journal, 2004. 3.

[2] Beagrie, N., Digital curation for science, digital libraries, and individuals. International Journal of Digital Curation, 2006. 1(1).

[3] Beagrie, N. and D. Greenstein, A Strategic policy framework for creating and preserving digital collections. 1998, London: Arts and Humanities Data Service: London.

[4] Digital Curation and Preservation: Defining the research agenda for the next decade, in Report of the Warwick Workshop - 7 & 8 November 2005. 2005, Digital Curation Centre: Warwick, UK.

[5] Lord, P. and A. Macdonald, E-Science Curation Report--Data Curation for E-science in the UK: An Audit to Establish Requirements for Future Curation and Provision. 2003, JISC Committee for the Support of Research. p. 85 pages.

[6] Day, M., Metadata for digital preservation: an update. Ariadne, 1999(22).

[7] Geosciences Network. Last visited 16 August 2006 http://www.geongrid.org/.

[8] National Ecological Observatory Network. Last visited 3 October 2006 http://neoninc.org/.

[9] Ni, K., et al., Sensor Network Data Fault Types. ACM Transactions on Sensor Networks, in review.

[10] Wallis, J.C., et al. Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries. in 11th European Conference on Digital Libraries. 2007. Budapest, Hungary: Berlin: Springer.

[11] 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 10th European Conference on Digital Libraries. 2006. Alicante, Spain: Berlin: Springer.

[12] 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, 2007.

[13] Mayernik, M.S., J.C. Wallis, and C.L. Borgman. Adding Context to Content: The CENS Deployment Center. in American Society for Information Science & Technology. 2007. Milwaukee, WI: Information Today.

[14] 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. 2007. Vancouver, BC: Association for Computing Machinery.

[15] Pepe, A., et al. *Knitting a fabric of sensor data and literature*. in *Information Processing in Sensor Networks*. 2007. Cambridge, MA: Association for Computing Machinery/IEEE.

[16] Wallis, J.C., et al. Moving Archival Practices Upstream: An Exploration of the Life Cycle of Ecological Sensing Data in Collaborative Field Research. in *3rd International Digital Curation Conference*. 2007. Washington, D.C.

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