The International Journal of Digital Curation
Issue 1, Volume 3 | 2008

Moving Archival Practices Upstream: An Exploration of the Life Cycle of Ecological Sensing Data in Collaborative Field Research

Jillian C. Wallis,

Center for Embedded Networked Sensing, UCLA

Christine L. Borgman, Matthew S. Mayernik, Alberto Pepe,

Department of Information Studies

Graduate School of Education & Information Studies, UCLA

Abstract

The success of eScience research depends not only upon effective collaboration between scientists and technologists but also upon the active involvement of data archivists. 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 life cycle of data from collaborative ecological research with embedded networked sensing technologies. A better understanding of these processes will enable archivists 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-stage life cycle. Among the findings are the cumulative effect of decisions made at each stage of the life cycle; the balance of decision-making between scientific and technology research partners; and the loss of certain types of data that may be essential to later interpretation.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

The success of eScience research depends upon effective collaboration between scientists and technologists. Partners must often learn how to produce data that are meaningful to participants from multiple disciplines. Many decisions are made about data at each stage in their 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 (Esanu, Davidson, Ross & Anderson, 2004) 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 (Beagrie, 2006; Beagrie & Greenstein, 1998). Some stages can be augmented by technical means, such as the application of 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 (Digital Curation Centre [DCC], 2005), our goals are to develop:

1. more detailed data models for each domain, including intra-domain and inter-domain commonalities;
2. automatic processes for data and metadata capture, and;
3. 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" (Lord & Macdonald, 2003) that will support the use, reuse, 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 (Day, 1999).

Background

Research reported here is affiliated with the Center for Embedded Networked Sensing (CENS)[1], a National Science Foundation Science and Technology Center established in 2002. 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.

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 GEON[2] or NEON[3], 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

[1] Center for Embedded Networked Sensing (CENS) http://research.cens.ucla.edu/

[2] Geosciences Network (GEON) http://www.geongrid.org/

[3] National Ecological Observatory Network (NEON) http://www.neoninc.org/

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 into 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.

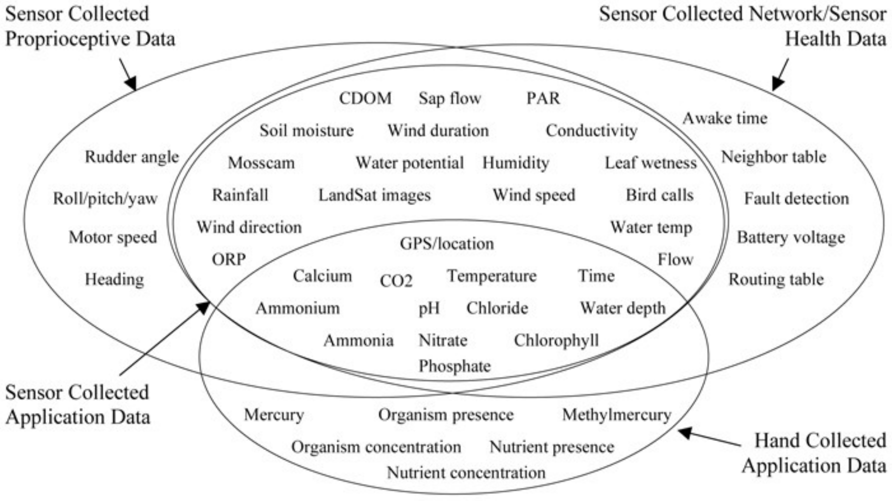

Figure 1. CENS data variation organized by collection method and use.

CENS Data, Users, and Uses

As shown above in Figure 1, data from CENS' dynamic field deployments can be grouped into four types. Sensors are used to collect data on:

1. the scientific application;
2. the performance of the sensors themselves;
3. proprioceptive data to use in navigation for robotic sensor technology;
4. 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; these are examples from a longer list. Some data serve only one purpose, but most serve multiple purposes as illustrated by the intersecting sets in Figure 1. 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 (if not sole) 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 the latter researchers, with minimal archival value. However they may be essential for reuse of the application science data by others.

Methods

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 results on our understanding of the scientific data life cycle based on responses to questions about data characteristics and architecture. Findings on other questions are reported elsewhere (Borgman, Wallis, & Enyedy, 2006, in press; Borgman, Wallis, Mayernik, & Pepe, 2007; Mayernik, Wallis, & Borgman, in press; Pepe, Borgman, Wallis, & Mayernik, 2007; Wallis et al, in press).

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 (Glaser & Strauss, 1967) to identify themes and to test them in the full corpus of interview transcripts and notes.

User scenarios for how data were captured, processed, and published were extracted from the interview 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 against observations of their data collection efforts.

Results

We have identified nine stages that appear to be common to the CENS deployments studied, the researchers, and to the resulting data, as shown in Figure 2 below. The order of the steps is not absolute, as some stages are iterative while others may occur in parallel. 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.

Stage 1: Experimental 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. Application science researchers identify interesting locations, time periods, and variables for data collection. Technology researchers examine performance data from previous experiments to identify new ways of testing the equipment. Researchers tend to use their own data for these purposes, with the intention of comparing or combining new data with prior data. Data from other sources, such as monitoring data from government agencies, is occasionally used.

The back-and-forth between application science and technology researchers has evolved over the five years of CENS. The early years were driven by technology researchers asking application scientists, “We have this equipment; can you do any science with it?” Now this interaction is driven by application researchers asking, “I need to do this science; what equipment can you give me?” Compromises are reached that give both parties something to test. This stage includes selecting sensors that will satisfy the needs of both parties, as each sensor collects specific parameters (e.g., temperature, salinity, nitrate concentration, network connectivity, etc.).

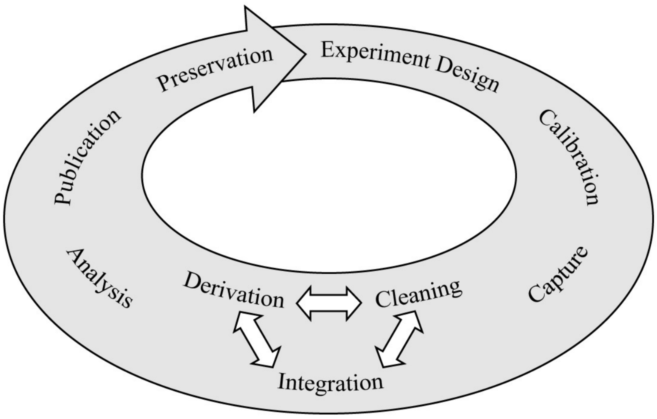

Figure 2. Life cycle of CENS data.

Stage 2: Calibration and Ground-Truthing

Before sensors are deployed, they are calibrated to known solutions or values to identify the offset between the actual measurement and the expected measurement. For our application science researchers, "calibration" refers to establishing offsets in the lab. As equipment is being deployed in the field, it may need to be "ground-truthed" or calibrated again to make sure that the offsets have not changed during the trip from the lab to the field site (Wallis et al, in press).

Calibration information can be described as a function with a slope and y-intercept. A certain amount of calibration drift occurs when equipment is deployed, which is described as another function using the calibration information from before and after deployment. Researchers usually assume that this drift function is either linear or logarithmic in nature. It is impossible to add intermediate calibration points, as removing equipment once it has been deployed can disturb the environment and phenomena being measured, and create gaps in the data stream. To compensate, physical samples are collected to verify data captured from the sensors.

Robotic equipment that houses sensing equipment also requires calibration, as robots must literally get their bearings. Robots that run along a tether or a static line must first establish the horizontal length of the tether, adjust for vertical sag, and program in depth parameters for how far the sensing payload may drop before it will hit the ground. Similarly, untethered robots must learn boundary and obstacle locations, lest they run aground or into some other sensing equipment.

Application science and technology researchers each have stakes in how equipment is calibrated and ground-truthed, although groups may have different tolerances. Earlier in CENS, technology researchers were more concerned with capturing any data than with capturing good data. Visualization tools now enable application scientists to identify data that are problematic from an ecological standpoint. Data quality concerns are now more balanced between application science and technology partners.

Stage 3: Data Capture

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). Some data are collected less for scientific value than for adjusting sensor readings or placing readings in appropriate contexts. These include sensor, network, and system health data, which might indicate the reliability of each measure; and proprioceptive data that indicate changes in location for a sensor through time; and other variables such as temperature or turbidity that may affect sensor performance. Application science researchers collect additional observations by traditional field research methods. Physical samples, such as water or soil samples are collected and processed on site to minimize contamination or organism breakdown. Some of these samples will be used later to double-check instrument calibrations.

Both technology and application science researchers sample observations in real-time to check for data integrity, sensor reliability, variability, and other factors. Because the data have not yet been cleaned using the calibration information, researchers cannot compare readings of different sensors. Instead they look for relational trends in the data such as common rates of change. Application science researchers will check sensing data against their models of how the ecological system should work, just as the technology researchers will check against their model of how the technology should work. If results differ from expectations, equipment setups will be investigated or experiments will be adjusted. Sensors may be moved from other locations to increase the sensor density, thus gaining a higher resolution of data about a phenomenon of interest. Some equipment will be restarted, repaired, or removed if not behaving adequately. This feedback loop continues until the end of the deployment. Careful records must be kept of where sensors were placed, and when, where, and why they were moved, if the data are to be interpreted adequately later.

Because technology used by CENS researchers is highly experimental, rather than hardened, off-the-shelf equipment, it is imperative that technology researchers participate in data collection. Researchers involved in equipment development are better able to debug problems in the field. In some cases the technology is so nascent that only the technology researchers know how to deploy it. While the emphasis may be on the collection of scientific data, the experiments run within the context of CENS operations are as much technological experiments as they are scientific ones.

Stage 4: Cleaning Data

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. Technology researchers mainly perform this task, as they are the most familiar with handling the data in the extremely raw form that streams from the sensors.

Outliers must be identified and flagged or removed. This process can be contentious, as “sensor artifacts” are often introduced into the data. Some faults have characteristic signatures, such as the “stuck-at fault,” where a sensor will get hung up on a value and continue to read the same value until it is “un-stuck”. Other faults are not so easy to identify, and lead to debate between application and technology partners about whether the data reflect a true phenomenon or are merely a sensor fault. CENS technology researchers are developing methods to improve data integrity by identifying and correcting such errors, but these methods have not yet been implemented widely.

Future integration of data from multiple sensors relies heavily on the ability to synchronize timestamps. Sensor clocks often drift and power interruptions or other faults can cause equipment to reboot and reset timestamps. The timestamp synchronization is performed using scripts and is at best an imperfect process. Technology researchers usually perform these tasks because of their familiarity with raw sensor data and writing transfer scripts.

This stage in the data life cycle significantly affects the future interpretation of the data. Mistakes in data cleaning may undermine later conclusions. Application science researchers must allow and trust the technology researchers to perform cleaning tasks that would otherwise fall to scientific partners. Discomfort can arise because technology researchers have a significantly different set of practices to establish data validity and reliability. Several application scientists interviewed were unsure about how well these tasks are performed in comparison to their own well-established set of practices for data verification and validation. Even when unsure about the specifics of their partners' practices, they expressed confidence in their research partners, however.

Stage 5: Deriving Numerical Data

Few of the observations and samples collected in the field can be interpreted without derivation into more meaningful data points. Researchers using nascent sensing equipment typically collect data at a very high frequency, using a deliberate over-sampling technique to minimize the contribution of equipment errors. Data typically must be averaged into composite points before they can be used in analysis. Data from those sensors that capture indirect measures, using one variable as an indicator for another, need to be processed through models that express the relationship between the sensed variable and the indirect variable. Similarly, data from sensors that are affected by external variables, such as temperature, need to be adjusted accordingly.

Physical samples such as water or soil core samples need to be processed by application scientists. Water or soil samples may yield useful data only after being separated in a centrifuge and then cultured in the lab for hours or days. To measure biomass, samples are counted by hand using grids and microscopes or incinerated in a calorimeter to yield a volume number.

Stage 6: Integrating Data from Multiple Sources

The CENS motto is "the network is the sensor." Relationships among observations from individual sensors are the real value from embedded networked sensors, not the individual observations. Researchers are looking for trends over time and across spatial locations. They want to know what happened when and where, in combination with what other events, and what preceded and followed interesting events. Datasets each given deployment are integrated by multiple researchers, for multiple reasons, and in multiple combinations. Application science and technology researchers each integrate the deployment data with respect to their own hypotheses and models.

Integration of sensor data depends on the accuracy of records about changes in sensor placement during the deployment. Sensor data must be integrated with hand-collected sample data. Water samples might be hand-collected four times in 24 hours, whereas water sensors may capture four data points per minute, resulting in incommensurate scales. Digitization and integration of hand-recorded data in field notebooks is a time-consuming task that does not fit within the current workflow of data interpretation.

Stage 7: Data Analysis

Data verification occurs throughout the data life cycle, and especially during the calibration and capture stages. Data analysis occurs after data are cleaned, derived, and integrated. 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.

Technology researchers combine scientific, proprioceptive, and network, system, and sensor health data to evaluate the performance of their technologies. Application scientists focus on understanding biological or chemical phenomena. Therefore, measures of biological or chemical behaviors, both sensor data and hand-collected data, are key to supporting or negating their hypotheses. Other sensor-collected parameters, such as temperature and humidity, are also important for background context. Proprioceptive and network/system/sensor health data are peripheral to science hypotheses, but testify to the trustworthiness of the instrumentation.

Stage 8: Publication

Data collected during embedded network sensor deployments culminate in scholarly publications such as journal articles, conference papers, posters, and technical reports. Publishable products vary between the application science and technology communities. For robotics researchers, the navigation algorithm might be the product. For others it might be the system or the program or the piece of equipment. For technology researchers, scientific data help to interpret and evaluate the functioning of their product, but play only a minimal role in publications. For application science researchers, the product is the tested hypothesis or the proven theory.

Publications serve as records of the methods used to capture, calibrate, clean, derive, integrate, and analyze the data, although rarely is enough detail provided to replicate the study. We did not find a one-to-one mapping between deployments and publications. One deployment may yield multiple papers, and one paper may draw on data from multiple deployments. Rarely are the data themselves published except as tables and figures. Some CENS researchers post their data on their team website or the CENS website after the publication appears; some will make data available on request.

Stage 9: Storage and 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. Some files remain on laboratory servers and may or may not be accessible to others outside the team. Some data are being contributed to a new CENS data repository, SensorBase.org. Scripts, programs, and systems are treated similarly to the data, languishing in folder structures and often lacking the documentation necessary to identify, access, or reuse them. Volatile hand-collected samples are either stored in refrigerator units until researchers run out of room, or are immediately discarded.

Unfortunately this stage of the data life cycle belies the lack of emphasis on proper archival practices on the part of researchers. That said, the data that researchers use most when they initiate research is their own data and typically the most recently collected. Only those application science researchers who compile multi-year datasets from single locations have established archival practices for maintaining their data. Previously these researchers compiled paper copies of data. Newer data from these longitudinal studies are kept in databases and backed up on CDs and DVDs.

Discussion and Conclusions

The success of eScience depends upon successful collaboration between application scientists and their partners in computer science and engineering. Data resulting from such collaborations are 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 life cycle of eScience data; including how they are handled and by whom at each stage. Evidence from our interview study and field research yields a nine-stage life cycle for these data.

Several findings are of particular import for data curation. One is the cumulative effect of decisions made at each stage of the life cycle. Decisions made in the experimental design stage determine what data exist for analysis. Calibration decisions are essential to interpreting the data. Calibration is notoriously difficult, as sensor measurements drift, and experimental sensors are balky in field conditions in unpredictable ways. The effect of calibration decisions may be magnified in the data cleaning process (which is one reason that researchers are attempting to push these decisions further upstream). A spike in data may be an interesting phenomenon or it may be an electrical error. Decisions about keeping or removing outliers affect later analysis. As the data are reduced to numerical values and as those values are integrated, the reliability and validity of each data stream may be obscured. Thus the more that can be known about decisions made at each stage, the more likely that others can interpret data in the future. Documenting these detailed decisions is difficult, of course. We are seeking ways to capture as many of them automatically as possible.

Another finding of interest to data archivists is the balance of decision making between scientific and technology research partners. In traditional field research, biologists and ecologists are accustomed to being in relatively complete control of their data collection methods and their data. In these partnering conditions, they must depend upon decisions made by their computer science and engineering partners about what data can be collected, and about the reliability of those data. Many computer science and engineering researchers are accustomed to using artificial datasets to build their models and test their equipment. Often they see having “real data” from their scientific partners as an advantage, although the reality of these data also introduces uncertainty into their own metrics.

Thirdly, the engineering data may be essential to later interpretation, at least for some uses. Engineering research practice is much less oriented toward data retention or sharing than is biology research practice in this community. How much of the engineering metric data need to be preserved for use in interpreting the scientific data is an open question.

CENS provides a rare opportunity for long-term, in-depth studies of the emergence of eScience practices in scholarly research (Borgman, 2007). Our system development efforts focus on instantiating the value chain of these data by linking the datasets directly with the documentation of associated field deployments and publications (Pepe et al., 2007). Other work in progress is comparisons between CENS and other eScience collaborations. Our future research will continue to explore and refine the data life cycle identified here, and to build systems to support it. At present, much of the sensing technology is experimental, but commercial off-the-shelf sensors are also in use. Research questions about data provenance will evolve as the technology stabilizes and the scientific research questions broaden.

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.

References

Beagrie, N. (2006, November). Digital curation for science, digital libraries, and individuals. International Journal of Digital Curation, 1(1). Retrieved April 7, 2007, from http://www.ijdc.net/

Beagrie, N. & Greenstein, D. (1998). A Strategic policy framework for creating and preserving digital collections., London: Arts and Humanities Data Service: London. Retrieved under 'Archived Documents' May 20, 2008, from http://ahds.ac.uk/about/publications/

Borgman, C.L. (2007). *Scholarship in the digital age: Information, infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Borgman, C.L., Wallis, J.C., & Enyedy, N. (2006). Building digital libraries for scientific data: An exploratory study of data practices in habitat ecology. In *10th European Conference on Digital Libraries*. Alicante, Spain: Berlin: Springer.

Borgman, C.L., Wallis, J.C., & Enyedy, N. (2007). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*.

Borgman, C.L., Wallis, J.C., Mayernik, M., & Pepe, A. (2007). 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.

Day, M., (1999, December). Metadata for digital preservation: an update. *Ariadne, 1999, (22)*. Retrieved July 27, 2007, from http://www.ariadne.ac.uk/issue22/metadata/

Digital Curation Centre. (2005). *Digital curation and preservation: Defining the research agenda for the next decade*. In Report of the Warwick Workshop, November 7-8, 2005. Digital Curation Centre: Warwick, UK. Retrieved July 27, 2007,.from http://www.dcc.ac.uk/events/warwick_2005/Warwick_Workshop_report.pdf

Esanu, J., Davidson, J., Ross, S., & Anderson, W. (2004). Selection, appraisal, and retention of digital scientific data: Highlights of an ERPANET/CODATA workshop. In *Data Science Journal, 2004, 3*. Retrieved July 30, 2007, from http://www.jstage.jst.go.jp/article/dsj/3/0/227/_pdf

Glaser, B.G., & Strauss, A.L. (1967). The discovery of grounded theory; strategies for qualitative research. Observations., Chicago,: Aldine Pub. Co. x, 271.

Lord, P., & Macdonald, A. (2003). *E-science curation report: Data curation for e-science in the UK: An audit to establish requirements for future curation and provision*. (Prepared for JISC Committee for the Support of Research.) Retrieved October 1, 2006, from http://www.jisc.ac.uk/uploaded_documents/e-scienceReportFinal.pdf

Mayernik, M.S., Wallis, J.C., Borgman, C.L., and Pepe, A. (2007). Adding Context to Content: The CENS Deployment Center. in Andrew Grove (Ed.), Proceedings of the 70th ASIS&T Annual Meeting, vol. 44, 2007. Milwaukee, WI: Richard B. Hill.

Pepe, A., Borgman, C.L., Wallis, J.C., & Mayernik, M.S. (2007). Knitting a fabric of sensor data and literature. In *Information Processing in Sensor Networks*. Cambridge, MA: Association for Computing Machinery/IEEE.

Wallis, J. C., Borgman, C. L., Mayernik, M., Pepe, A., Ramanathan, N. & 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.