Drowning in Data: Digital Library Architecture to Support Scientific Use of Embedded Sensor Networks

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

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

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

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

ABSTRACT

New technologies for scientific research are producing a deluge of data that is overwhelming traditional tools for data capture, analysis, storage, and access. We report on a study of scientific practices associated with dynamic deployments of embedded sensor networks to identify requirements for data digital libraries. As part of continuing research on scientific data management, we interviewed 22 participants in 5 environmental science projects to identify data types and uses, stages in their data life cycle, and requirements for digital library architecture. We found that scientists need continuous access to their data from the time that field experiments are designed through final analysis and publication, thus reflecting a broader notion of "digital library." Six categories of requirements are discussed: the ability to obtain and maintain data in the field, verify data in the field, document data context for subsequent interpretation, integrate data from multiple sources, analyze data, and preserve data. Three digital library efforts currently underway within the Center for Embedded Networked Sensing are addressing these requirements, with the goal of a tightly coupled interoperable framework that, in turn, will be a component of cyberinfrastructure for science.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: Systems Issues, User Issues.

H.3.5 [**Online Information Services**]: Data sharing.

J.4 [**Social and Behavioral Sciences**]: Sociology.

General Terms

Design, Human Factors, Standardization.

Keywords

Scientific data, user-centered design, functional requirements analysis, networked sensing, data deluge, data capture, data use, data preservation.

JCDL'07, June 17–23, 2007, Vancouver, British Columbia, Canada.

1. INTRODUCTION

Scientific data are expensive to produce and often have long-lasting value for scientists, students, and public policy makers. Data associated with specific times and places, such as ecological observations, are irreplaceable. The predicted "data deluge" [1] is now a reality for many scientific researchers. The deluge is occurring not in absolute sense, but in a relative one. While "big science" fields such as physics and astronomy [2] have begun to construct tools and repositories to address this deluge, "little science" areas dependent upon fieldwork lack the tools and infrastructure to manage the growing amounts of data generated by new forms of instrumentation. The lack of an integrated framework for managing these types of scientific data presents significant barriers not only to those scientists conducting the research, but also to those who would subsequently reuse the data. A few gigabytes of data daily might be a trickle to a high-energy physicist, but waterfall to a habitat ecologist.

e-Science and cyberinfrastructure initiatives recognize the need for better data management capabilities, but research tends to focus more on technical than social solutions. Researchers are more likely to adopt tools that fit into their practices of data collection and analysis. More needs to be understood about the scientific practices of those whose research is evolving through the use of new technologies that generate data at volumes heretofore unknown. The environmental sciences are among the many fields whose field research methods are being transformed by the ability to capture observations at high spatial and temporal granularity via in situ embedding of sensor network technologies [3, 4]. This paper reports on a study of scientific practices to identify requirements for data digital libraries.

2. DATA DIGITAL LIBRARIES: PROBLEM OR SOLUTION?

Digital libraries, whether for data or documents, typically serve as repositories for content that was ingested at or near the end of its life cycle. That narrow view of digital libraries can be a problem for scientific data management, as scientists often need access to their data much earlier. Conversely, digital libraries can be a solution, if they are conceived more broadly as systems that encompass the entire information life cycle [5]. To understand the appropriate tools and services required for data digital libraries, much more study is needed of scientific practices associated with data storage and retrieval and with their production, analysis, and interpretation.

2.1 Scientific Data Practices

Studies of scientific data practices by us and others suggest that in only a few fields do researchers contribute their data to shared repositories [6-10]. Among the reasons are the additional effort required to document data in standardized formats and concerns about others having access to their data prior to publication. Few repositories offer the tools and services that scientists appear to desire, such as the ability to store data for personal analysis and use, and tools to monitor and interpret data in the field so that changes can be made to experiments in progress. The majority of scientific researchers in our studies save all of their data and they reuse those data when applicable to future research. However, their available tools support analysis of data much better than they do preservation and sharing. As a result, scientists often store data with minimal documentation and do little toward preservation other than to back up files. These approaches are largely local and do not adequately support future access and use. Few of the current scientific tools are scaling up well to the volume of data now being produced by embedded sensor networks.

2.2 Embedded Networked Sensor Data Structures

Capturing the data from embedded sensor networks in a common data structure would seem to provide obvious benefits for these scientists. A consistent format would facilitate the design of tools and services for data collection, analysis, sharing, and storage. Common data structures also could improve data integrity by flagging observations that are inconsistent with model or equipment parameters. We have been surveying available data structures for description of datasets and of observations since 2002 [9-11]. Several XML-based standards and protocols exist for this diverse community, but none of them are stable or widely adopted. The lack of established data standards contributes to the entrenchment of ad hoc management techniques and minimal documentation.

Structures most relevant to embedded sensor network data in the environmental sciences include the Ecological Metadata Language, supported by the Knowledge Network for Biocomplexity [12, 13], 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 [14]. Other structures of interest are specific to individual research areas, such as WaterML for hydrology research. Complicating matters further, scientists may be involved in multiple national and international projects, each of which subscribes to different metadata development efforts.

Forcing standardization prematurely can hinder scientific progress [15, 16]. Many scientific research areas continue to be productive without the use of shared instrumentation, shared repositories, or agreements on standards for data description. As the environmental sciences become more instrumented, scientists face the questions of what to standardize, when, and for what purposes. The multiplicity of standards in this field poses significant challenges to researchers and has limited the widespread implementation of any individual standard.

2.3 CENS as a Context to Identify Digital Library Requirements

Research reported here is affiliated with the *Center for Embedded Networked Sensing* (CENS), a National Science Foundation Science and Technology Center established in 2002 [http://www.cens.ucla.edu/]. CENS supports multi-disciplinary collaborations among faculty, students, and staff of five partner universities. The Center's goals are to develop and implement innovative wireless sensor networks. CENS' scientists are investigating fundamental properties of these systems, designing and deploying new technologies, and exploring novel scientific and educational applications. CENS' research crosses four primary scientific areas: habitat ecology, marine microbiology, environmental contaminant transport, and seismology, plus applications in urban settings and in the arts. The science is based on *in situ* monitoring, with the goal of revealing patterns and phenomena that were not previously observable.

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

2.4 Static vs. Dynamic Sensing Systems

Sensor networks, per se, are not a new technology. Large manufacturing operations and chemical processing plants, for example, rely heavily on sensor networks to manage operations. Similarly, water flow and water quality monitoring relies heavily on embedded sensor networks. In the U.S. alone, public regulatory agencies monitor several hundred million individual sensors on streams, lakes, and rivers. Extant and emergent observatory networks for research, such as the Long Term Ecological Reserve System [17], WATERS (merging CUAHSI and CLEANER) [18, 19], GEON (Geosciences Network) [20], and NEON (National Ecological Observatory Network) [21] also use embedded sensor networks to collect scientific data. Most of these applications of sensor networks are static deployments: sensors are placed in appropriate positions to report data continuously on local conditions. Sensors are monitored, both by humans and by computers, to determine changes in conditions. Autonomous networks can rely on machine actuation to capture scientifically relevant data, to alter data collection (e.g., capture data more frequently if excessive pollution is suspected), or to report emergencies that require intervention (e.g., faults in dams, water contamination). Data repositories such as ROADnet [22] can capture real time data from autonomous networks.

While the initial framework for CENS was based on autonomous networks, early scientific results revealed the difficulty of specifying field requirements in advance well enough to operate systems remotely. Most CENS' research is now based on dynamic "human in the loop" deployments where investigators can adjust monitoring conditions in real time. CENS' teams have data collection "campaigns" in which they deploy an embedded sensor network in the field for a few hours or a few days. They may return to the same site, or a similar site, repeatedly, each time with slightly different equipment or research questions. These discrete field deployments offer several advantages to the scientific researchers. They can deploy prototype equipment that is much more sophisticated than the robust equipment required for autonomous networks. The researchers who developed these technologies often participate in deployments to test, evaluate, and adjust their equipment in the field.

Brief deployment campaigns also enable researchers to deploy equipment that is too delicate, too expensive, too premature, or has too short a life span to leave unattended in the field. Some chemical sensors, for example, are sufficiently volatile that they lose sensitivity within a few days. Scientists can finalize the precise positioning of equipment in the field, based on current conditions (e.g., moisture, temperature, light, shade). Hand-collected samples of water and soil often are required to calibrate sensors. Scientists also can alter the position of their sensors and the frequency of sampling while in the field. If the water depth chosen is not yielding interesting data, they may raise, lower, or move the sensors. Dynamic deployments also benefit the computer science and engineering researchers as equipment may be tested sooner and more iteratively than with autonomous networks. By collaborating in the field, researchers and students from all the participating disciplines learn about each others' problems and needs very quickly.

While dynamic, human-in-the-loop sensor deployments yield better science for this particular kind of research, the data they generate are much harder to manage by traditional methods. Each deployment may have different research questions, methods, equipment, and data. This research framework differs greatly from the approach assumed by systems such as ROADnet, in which participants are able to agree on common semantics, data structures, services, ontologies, and preservation policies [22].

3. METHODS

The findings reported here are drawn from an interview study of five environmental science projects. For each project, we interviewed a complementary set of science and technology participants, including faculty, post-doctoral fellows, graduate students, and research staff.

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 questions are categorized as follows:

- **Data characteristics:** What data are being generated? To whom are these data? To whom are these data useful?
- **Data sharing:** When will scientists share data? With whom will they share data? What are the criteria for sharing? Who can authorize sharing?
- **Data policy:** What are fair policies for providing access to these data? What controls, embargoes, usage constraints, or other limitations are needed to ensure fairness of access and use? What data publication models are appropriate?
- **Data architecture:** What data tools are needed at the time of research design? What tools are needed for data collection and acquisition? What tools are needed for data analysis? What tools are needed for publishing data? What data models do the scientists who generate the data need? What data models do others need to use the data?

This paper reports our results on data architecture. Early results on the other questions are reported elsewhere [9, 10], and fuller analyses are forthcoming.

3.1 Participants

CENS is comprised of about 70 faculty and other researchers, about 140 student researchers, and some full-time research staff who are affiliated with the five participating universities. The 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.

3.2 Qualitative Data Analysis

The interviews were audiotaped, transcribed, and complemented by the interviewers' memos on topics and themes [20]. Transcription totaled to 312 pages of interview data. Analysis proceeded to identify emergent themes. We developed a full coding process using NVIVO, which was used to test and refine themes in coding of subsequent interviews. With each refinement, the remaining corpus was searched for confirming or contradictory evidence. This study used the methods of grounded theory [23] to identify themes and to test them in the full corpus of interview transcripts and notes.

3.3 Functional Requirements Extraction

While our interview instrument asked general questions about the data practices of the CENS researchers, specific functional requirements could be inferred and in some cases were specifically requested by our subjects. We identified categories of data being captured, uses and users of those data, and researchers' needs at each stage of the data life cycle.

4. FUNCTIONAL REQUIREMENTS

Results are organized in several subsections. First we describe a typical field deployment based on interview data and on participation in deployments. Then we report on the types of data resulting from deployments, the data life cycle associated with dynamic sensor network deployments, and lastly, specific digital library tools and services.

4.1 Dynamic Deployment Scenario

An example CENS' deployment is one to study biological processes associated with harmful algal blooms, with the long-term goal of preventing such blooms. The deployment takes place at a lake known for summer blooms. Available background information about the lake includes peak months for algae, a topology of the lakebed, organism species, and nutrient presence and concentration. Prior to going in the field, the team calibrates its equipment in the laboratory. Because these aquatic phenomena occur in reference to the diel cycle, they plan to take water samples for a full 24 hours to track the presence of organisms and the context associated with the release of toxins. The team places sensors in the lake using static buoys and a robotic boat. They document GPS coordinates of the sensors, times of placement, and serial numbers of each sensor. Once sensors begin to report data, researchers might find that the water moves more quickly at one end of the lake, or that the water is greener and at a higher temperature where a rock slows the flow. Based on this information, the team may alter its plans for sensor placement or

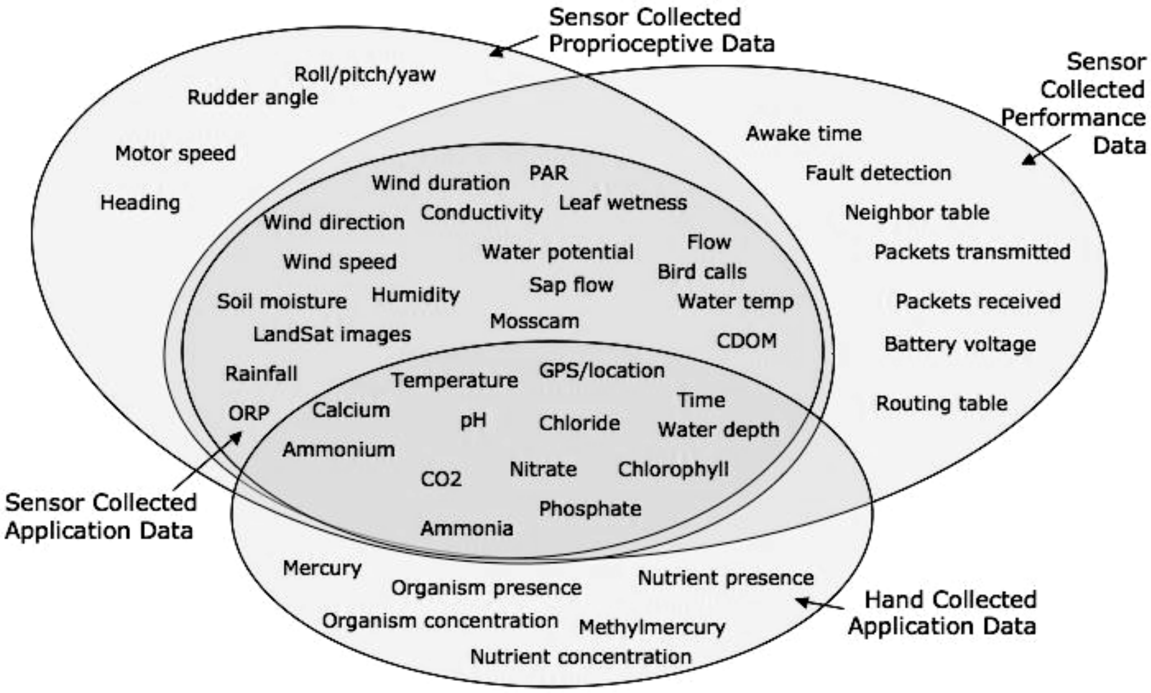

Figure 1: CENS data variation organized by collection method and use.

for hand collection of water samples. They typically set up a "wet lab" on site to process water samples and will use this information to adjust their data collection. Water samples are taken adjacent to sensors for field calibration. At the end of the deployment, the equipment is removed and returned to the lab. Water samples are analyzed 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, data are burned to DVDs and shelved with other data.

4.2 Data Types and Uses

As shown in Figure 1, data from CENS dynamic field deployments can be grouped into four types. Sensors are used to collect data on the scientific application, on the performance of the sensors themselves, or – for robotic sensor technology – proprioceptive data about the world to use in navigation. 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; these are examples from a longer list. Some data serve only one purpose, but most serve multiple purposes as illustrated by the intersecting sets. When we asked our subjects about capturing, using, sharing, and preserving data from deployments, and about capabilities they desired in digital libraries 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 these researchers, with minimal archival value. Thus we have focused our search for digital library requirements on the needs of scientists and on data that address scientific applications.

When we asked our subjects about capturing, using, sharing, and preserving data from deployments, and about capabilities they desired in digital libraries 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 these researchers, with minimal archival value. Thus we have focused our search for digital library requirements on the needs of scientists and on data that address scientific applications. Figure 2: Heterogeneous sensor deployment. Graphic by Jason Fisher.

4.3 Data Life Cycle

A first step in developing digital library tools and services to support the data life cycle is to identify the stages in that cycle. We have identified eight stages that appear to be common to the CENS deployments studied and to the resulting data, as shown in Figure 3. The order of the steps is not absolute, as some stages are iterative.

4.3.1 Experimental Design

This first stage reuses data from prior research to design new experiments. Interesting locations or time periods for data collection are identified, as well as the variables to be collected. Researchers most commonly use their own data for this purpose, considering how to compare or combine prior data with new data. They will occasionally use data from other sources, such as monitoring data from government agencies, but are less likely to combine these with new data that they collect themselves. This stage includes selecting the sensors to deploy, as each sensor

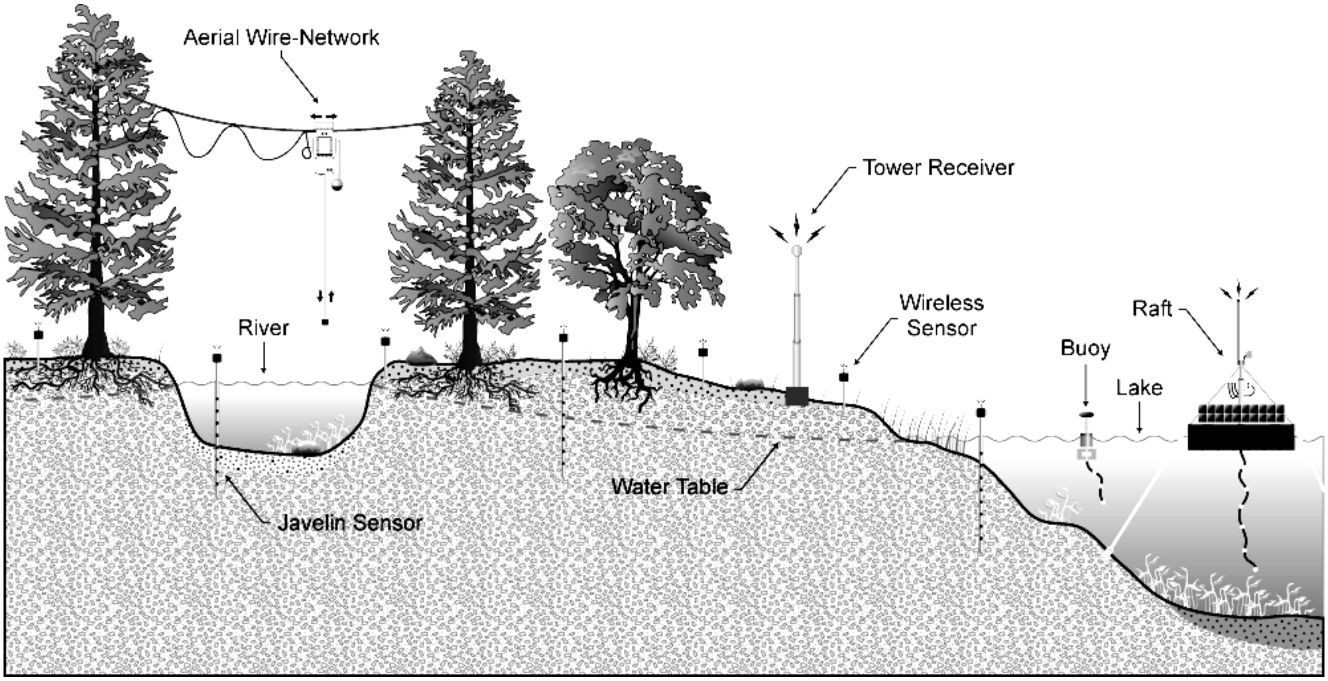

Figure 2: Heterogeneous sensor deployment. Graphic by Jason Fisher.

collects a specific type of observation (e.g., temperature, salinity, nitrate concentration).

4.3.2 Calibration

Before sensors are deployed, they are calibrated to known solutions or values. Once equipment is in the field, it may need to be "ground truthed" or calibrated again. In the deployment scenario above, sensors are calibrated in the lab based on the aquatic organisms expected to be in the lake. However, it is impossible to predict precisely which zooplankton and phytoplankton might be present, requiring additional calibration on site. Similarly, water samples collected next to the sensors are used to calibrate sensors for temperature and salinity. Sensors that are affected by temperature must be calibrated to local conditions. Outliers are investigated and other potential sensor artifacts are addressed during this stage.

4.3.3 Data Capture

Once the sensors are calibrated in the field, the team begins to collect observations of physical phenomena from the sensors and also may collect other observations by traditional field research methods. Some sensor measurements are direct (e.g., temperature, wind speed) and others are indirect (e.g., voltage measure of fluorescence as an indicator for chlorophyll). The scientific team samples the observations as they are being captured to check for data integrity, sensor reliability, variability, and other factors. If results differ from expectations, staff will check further or will adjust the experiment. Team members might move sensors from other locations to increase the sensor density, thus gaining a higher resolution of data about a phenomenon of interest. This feedback loop continues until the end of the deployment. Careful records must be kept of where sensors were placed, and of when, where, and why they were moved, if the data are to be interpreted adequately later.

4.3.4 Deriving data from indicators and samples

Many of the observations and samples collected in the field cannot be interpreted without further processing. Sensors can detect some kinds of chemicals but not others, for example, so the absence or low value of one chemical may indicate the presence of another that cannot be detected by sensors. These types of sensor observations serve as input to models from which the data of interest are derived. Similarly, water samples may yield useful data only after being processed through a centrifuge and then cultured in the lab for hours or days.

4.3.5 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. Scientists 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. Integrating 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. CENS technology researchers are developing methods to improve data integrity by identifying and correcting such errors. Other factors that influence integration of sensor data are the accuracy of records about changes in sensor placement during the deployment. Sensor data also must be integrated with hand-collected data. Water samples might be hand-collected 4 times in 24 hours, whereas water sensors may capture 4 data points per minute, resulting in incommensurate scales.

4.3.6 Data Analysis

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

4.3.7 Publication

Data from embedded network sensor deployments culminates in publications. We did not find a one-to-one mapping between deployments and publications, however. One deployment may yield multiple papers, and one paper may draw on data from multiple deployments. Rarely are the data themselves published. Some CENS scientists post their data after the publication appears and some will make data available on request. The publication serves as a record of the methods used to capture, calibrate, derive, integrate, and analyze the data, although rarely is enough detail provided to replicate the study.

4.3.8 Storage and Preservation

Few, if any, of the CENS researchers interviewed had data preservation strategies commensurate with those of the digital library community. It is more accurate to say that they back up their data and that CDs and DVDs are the preferred storage media. 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 repository, known as SensorBase.

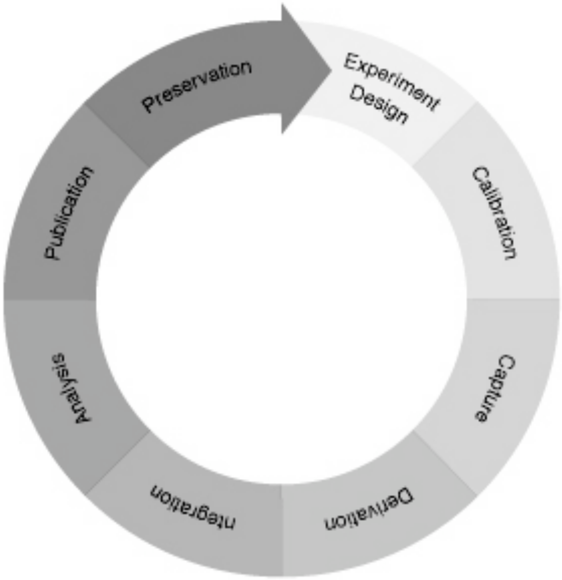

Figure 3: Life cycle of CENS data.

5. TECHNICAL REQUIREMENTS

Designing digital libraries and data structures for static sensor networks is already difficult due to the variety of equipment, observations, spatial and temporal variations, and to the complexities introduced by actuated sensing whereby sensors change data capture parameters on demand. Digital libraries for dynamic deployments must address all of these issues, plus accommodate the disparate datasets produced by each deployment. Because these are experimental deployments that both test technology and gather scientific data, digital libraries must support early access to the data as they are captured, and must incorporate rich contextual information about sensor data such as how, when, and why sensors were moved.

Both CENS' technology and collaborations have matured to the point at which the amount of data generated is overwhelming and the science problems are understood well enough to drive technology development. We distilled many individual requirements from our interviews with CENS domain scientists, computer scientists, and engineering researchers, which we have grouped into six categories: (i) the ability to obtain and maintain data in the field, (ii) verify data in the field, (iii) document the data sufficiently that it can be interpreted, (iv) integrate data from multiple sources, (v) analyze data, and (vi) preserve the data. These requirements are presented in the order they occur in the information life cycle. Several of these requirements already are foci of CENS research and concrete digital library efforts are currently underway to address such requirements.

5.1 Obtain and Maintain Data in the Field

Laptop computers are an essential field technology for recording notes on deployments, and records of hand-collected samples, and often to transfer data from individual sensors. Data on individual laptops can be difficult to reconcile with data on other team members' laptops. Laptops also are vulnerable to field conditions where temperature and moisture vary widely. Teams need safer short-term storage for data until adequate network resources are available. As one subject explained the problem, "*I was just storing it locally on this laptop because I did not have network access... for two weeks, during the entire deployment..*" Data captured on portable machines often tends to stay on those machines rather than being transferred to shared servers on a regular basis.

This problem is being addressed by SensorBase, CENS's central data repository, currently in its development stage. SensorBase allows publishing and sharing of sensor data via "slogging," which is manually or automatically uploading sensor data to the archive in a way that resembles the posting of journal entries to a blog. SensorBase is intended to capture and provide access to raw data as they are captured by sensors, thus serving the early stages of the life cycle in which scientists need to inspect data and made adjustments to field experiments. In the absence of network connectivity, a mobile installation of SensorBase would guarantee remote data acquisition. This will minimize the synchronization and consistency problem between data captured on portable machines and data stored on shared servers, as distributed, mobile installations of SensorBase will automatically synchronize with the main installation at the earliest occurrence of network connectivity. The system also supports refined and analyzed data that are ready for publication. Other features are planned for SensorBase such as an RSS feed to alert scientists to interesting new data as they are captured.

5.2 Verifying Data in the Field

CENS scientific teams need tools in the field to calibrate, verify, and correct data. Sensors are delicate instruments whose measurements change over time. Rarely is it possible to determine the degradation curve without pulling sensors out and recalibrating them. Accurate calibration is necessary to determine the true value of each data point. For example, *"The sensor data will be given to me in a certain form and then I will convert it into a concentration, and then there will be some debate, because the sensors are dying as they're in the field.... So our pre-imposed calibration curves are pretty different from one another, so there will be some debate about whether we use the pre or the post or* *the average, or whether there's something we can do to measure how fast it's changing."*

Similarly, scientists need tools in the field to assess data quality and accuracy so that they make changes to their deployment. These include tools to visualize data quickly and easily. The initial in-field data interface offered only comma-delimited numerical data streaming off the sensors. Data in this form are meaningful to technology researchers who are primarily concerned with the presence, absence, and range of data, but are of limited value to scientists who need to make decisions in real time. CENS technology researchers and statisticians are working on better field tools for capturing calibration and degradation information and for visualization. Records of these processes need to be incorporated into the digital library to ensure accurate data interpretation throughout the life cycle. This could be achieved by building ad hoc ontologies for field instrumentation, that would describe precisely the technical specifications of sensors and actuators as well as particular information specific to the capture, such as calibration and degradation. Technical information organized in such a structured manner has the potential of being interpreted and reused across the entire data lifecycle by all other digital resources.

5.3 Documenting Data Context

Metadata can be used to describe observations and datasets, but may not provide sufficient documentation of the context in which data were collected. Choices of metrics and instrumentation for particular observations, methods of calibration, and changes in placement of sensors all need to be captured in a data digital library to support field deployments. Temperature is among the most problematic measures. While some computer science and engineering researchers interviewed said roughly, *"temperature is temperature,"* biologists gave much more nuanced descriptions of how temperature was measured: *"There are hundreds of ways to measure temperature. 'The temperature is 98' is low-value compared to, 'the temperature of the surface, measured by the infrared thermopile, model number XYZ, is 98.' That means it is measuring a proxy for a temperature, rather than being in contact with a probe, and it is measuring from a distance. The accuracy is plus or minus .05 of a degree. I [also] want to know that it was taken outside versus inside a controlled environment, how long it had been in place, and the last time it was calibrated, which might tell me whether it has drifted.."*

Documentation about deployment activities also can yield important contextual information about data, such as "trip lists" of equipment and supplies, procedures for setting up and calibrating equipment, team members who participated, and types of expertise required. Such information is valuable in maintaining consistency from one deployment to the next, in training new team members, and in improving the effectiveness and efficiency of individual deployments. As CENS has matured and expanded, the need to track deployment information has increased. At last count, more than 140 graduate students were affiliated with the Center. An "oral culture" about how to conduct deployments no longer suffices for maintaining continuity between deployments, and is particularly problematic in identifying information about past deployments that is needed to interpret resulting data.

In order to preserve and dynamically handle contextual information about deployments, CENS is developing the Deployment Center (CENSDC), a resource manager that will serve pre-deployment planning and post-deployment knowledge transfer by providing web interfaces to a searchable database of deployment information. Storing metadata descriptions of past, current, and future deployments in a centralized location, CENSDC will allow automation of data labeling procedures and reuse of sensor hardware specifications. Teams will have access to CENSDC before, during, and after deployments.

5.4 Integrating Data from Multiple Sources

The ability to integrate and reconcile data from multiple sensors in a network is among the fundamental technical challenges in embedded networked sensing research, and one that is an active research area in CENS. Some sensors periodically send data to nearby servers or to satellites. Others store data locally until a person retrieves them. The choice of sensor technology and data transfer method depends upon many considerations, including power sources, battery life, network access, and type of observation. Time stamps can be synchronized by in-network processing algorithms, but their accuracy is influenced by factors such as battery degradation rates. Many other factors complicate data integration, such as the varying intervals for data capture depending upon the sensor and the experiment. Scientists devote a substantial amount of effort to synchronizing these data and to performing other types of data cleaning before analysis: *"[A colleague] usually takes it from there and futzes with it in MATLAB, because really synching all of the sensors is a chore… he does an excellent job … but it's still a concern..because I've received data sets that I'm sure are not synched properly."* The differences in time scales between sensor data collection and hand collection, noted in the data life cycle discussion above, is similarly problematic for data integration and analysis. While digital libraries cannot solve these problems, they can capture as much contextual information as possible about the data to assist in its later interpretation.

5.5 Data Analysis

The increasing volume of data is rendering obsolete many of the traditional statistical tools of these scientists: *"When I started thinking about how this was different, [and could] let me ask different questions, I also knew that we were going to get data in a magnitude that I just could not analyze with all the normal tools that I use."* Not surprisingly, given the many disciplines and scientific applications in CENS, a wide range of statistical and analytical tools are popular, and the uses of those tools varies: *"Some people want to see a whole week's worth of data averaged, give me a number. Some people want it on a daily basis. Some people want it on a monthly basis. Some people want to see day-by-day, hour-by-hour, minute-by-minute. They want to see the pattern. It varies depending on the question that you are asking, and the data analysis might be vastly different."* Digital libraries do not need to incorporate statistical and visualization tools for this diverse audience, but they should be capable of importing and exporting data in an array of formats compatible with most tools in use by this community. As for calibration and sensor specifications, outlined in section 5.2, the statistical tools used post-capture should be documented in a well-defined ontology that would allow later data reuse and evaluation.

5.6 Data Preservation

As CENS has matured, the need for data preservation has increased, as expected. Early deployments *"were spitting out numbers. At that time it was more important that things were working at all, than were spitting out accurate data… If the data* *has been quality controlled and error checked, it is more valuable and something that we would want to preserve in perpetuity as opposed to a goofy data set that we end up dumping."* Today's CENS data sets are no longer considered "goofy." Scientists are concerned about how to assure data quality from the earliest stages of the life cycle so that it can be trusted and interpreted at the end of that cycle and into the future.

CENS supports the vision of an interoperable data framework capable of providing transparent data access, exchange, and reuse of heterogeneous resources. This approach requires that sensor data be labeled appropriately from the early stages of data acquisition to the final steps of data storage and publication. Tools will assist scientists in annotating their data and will progressively build knowledge bases for automatic annotation of captured data. The data lifecycle culminates with the presentation of results in a publication and its deposition in a scholarly repository. Digital data libraries, such as SensorBase and CENSDC, need to preserve data, expose them in an intelligible manner and, at the same time, inter-operate with each other. However, to preserve the integrity and value of the data lifecycle, a greater level of interoperability between the data libraries and the scholarly repository is needed. We intend to achieve such level of connectivity between all diverse digital resources - from sensor data to bibliographic data - by evaluating our data framework as a testbed for the Open Archives Initiative for Object Reuse and Exchange (OAI-ORE) in which CENS is currently participating [24, 25].

6. CONCLUSIONS

Digital libraries can best serve cyberinfrastructure requirements if they can accommodate data from its earliest generative stages. The volumes of data being produced by embedded sensor networks and other scientific technologies are transforming the field research methods of the environmental sciences. To this community, gigabytes of data per day is a deluge, and far more than they can capture and manage usefully. Data that accumulate in ad-hoc computer files on individual and communal servers cannot easily be leveraged for purposes such as longitudinal and comparative analyses.

We identified six sets of requirements for digital library architecture to serve the data lifecycles of these scientific and technical communities: the ability to obtain and maintain data in the field, to verify data in the field, to document data context, to integrate data from multiple sources, to analyze, and to preserve the data. These requirements intersect with other research within CENS to improve the integrity of data. Data integrity begins at the earliest stages in the cycle. Unless scientists and other subsequent users of data from dynamic sensor deployments can trust the integrity of the data through each stage of processing, those data will be of minimal value.

In presenting the technical requirements, we introduced three digital resources that we will be tightly coupled in an interoperable framework: deployment information (CENSDC,) sensor data (SensorBase), and publications (OAI-compliant bibliographic repository). Each of these digital libraries will help to document the others. Ultimately it will be possible to search CENSDC for a deployment and then follow links to the resulting data and publications, to search SensorBase and follow links to deployments and publications, and to search the bibliographic database and follow links from papers to datasets and to the deployments from which they originated. Such a grand level of interoperability between digital objects will not only improve the reuse and long-term preservation of sensor data, but also augment the quality and extent of scholarly communication of the disciplines by leveraging the intrinsic value of digital objects "beyond the borders of hosting repositories". We expect the results of our digital library research and development with dynamic deployments of embedded sensor networks to have implications far beyond the domain of the environmental sciences.

7. 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. SensorBase research in CENS is led by Mark Hansen and Nathan Yau. Support for CENS bibliographic database development is provided by Christina Patterson and Margo Reveil of UCLA Academic Technology Services.

8. REFERENCES

[1] Hey, T. and A. Trefethen, The Data Deluge: An e-Science Perspective, in Grid Computing – Making the Global Infrastructure a Reality. 2003, Wiley. Retrieved from http://www.rcuk.ac.uk/escience/documents/report_datadeluge.pdf on 20 January 2005.

[2] Price, D.J.d.S., Little Science, Big Science. 1963, New York: Columbia University Press

[3] Elson, J. and D. Estrin, Sensor networks: A bridge to the physical world, in Wireless sensor networks, C.S. Raghavendra, K.M. Sivalingam, and T.F. Znati, Editors. 2004, Kluwer Academic: Boston.

[4] Pottie, G.J. and W.J. Kaiser, Principles of embedded networked systems design. 2006, Cambridge, England: Cambridge University Press.

[5] Borgman, C.L., et al., Social Aspects of Digital Libraries. Final Report to the National Science Foundation; Computer, Information Science, and Engineering Directorate; Division of Information, Robotics, and Intelligent Systems; Information Technology and Organizations Program. 1996. Retrieved from http://is.gseis.ucla.edu/research/dl/index.html on 28 September 2006.

[6] Arzberger, P., et al., An International Framework to Promote Access to Data. Science, 2004. 303(5665): p. 1777-1778.

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

[8] Hilgartner, S. and S.I. Brandt-Rauf, Data access, ownership and control: Toward empirical studies of access practices. Knowledge, 1994. 15: p. 355-372.

[9] Borgman, C.L., J.C. Wallis, and N. Enyedy, Little Science Confronts the Data Deluge: Habitat Ecology, Embedded Sensor Networks, and Digital Libraries. International Journal on Digital Libraries, in press.

[10] Borgman, C.L., J.C. Wallis, and N. Enyedy, Building digital libraries for scientific data: An exploratory study of data practices in habitat ecology. 10th European Conference on Digital Libraries, Alicante, Spain, 2006. Berlin: Springer. LINCS 4172: 170-183.

[11] Shankar, K., Scientific data archiving: the state of the art in information, data, and metadata management. 2003. Retrieved from http://cens.ucla.edu/Education/index.html on 19 January 2005.

[12] Ecological Metadata Language. 2004. Retrieved from http://knb.ecoinformatics.org/software/eml/ on 25 November 2004.

[13] Knowledge Network for Biocomplexity. 2004. Retrieved from http://knb.ecoinformatics.org/index.jsp on 25 November 2004.

[14] Botts, M. Sensor Modeling Language (SensorML) Status. 2006. Retrieved from http://stromboli.nsstc.uah.edu/SensorML/status.html on 20 November 2006.

[15] Bishop, A.P., N. Van House, and B.P. Buttenfield, eds. Digital library use: Social practice in design and evaluation. 2003, MIT Press: Cambridge, MA.

[16] Bowker, G.C., Memory Practices in the Sciences. 2005, Cambridge, MA: MIT Press.

[17] U.S. Long Term Ecological Research Network. 2006. Retrieved from http://lternet.edu/ on 5 June 2006.

[18] Consortium of Universities for Advancement of Hydrologic Science. 2006. Retrieved from http://www.cuahsi.org on 15 November 2006.

[19] Collaborative Large-Scale Engineering Analysis Network for Environmental Research. 2006. Retrieved from http://cleaner.ncsa.uiuc.edu/home/ on 16 August 2006.

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

[21] National Ecological Observatory Network. 2006. Retrieved from http://neoninc.org/ on 3 October 2006.

[22] Real-time Observatories, Applications, and Data Management Network. 2007. Retrieved from Http://roadnet.ucsd.edu on 3 April 2007.

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

[24] Open Archives Initiative. 2007. Retrieved from http://www.openarchives.org/ore/ on 4 February 2007.

[25] Pepe, A., C.L. Borgman, M. Mayernik, and J.C. Wallis, Knitting a Fabric of Sensor Data Resources. International Conference on Information Processing in Sensor Networks, Cambridge, Massachusetts, 2007.