Digital Libraries for Scientific Data Discovery and Reuse: From Vision to Practical Reality

Jillian C. Wallis
Dept. of Info. Studies
University of California, Los Angeles
+1(310)206-0029
jwallisi@ucla.edu

Matthew S. Mayernik
Dept. of Info. Studies
University of California, Los Angeles
+1(310)206-0029
mattmayernik@ucla.edu

Christine L. Borgman
Dept. of Info. Studies
University of California, Los Angeles
+1(310)825-6164
borgman@gseis.ucla.edu

Alberto Pepe
Dept. of Info. Studies
University of California, Los Angeles
+1(310) 206-0029
apepe@ucla.edu

ABSTRACT
Science and technology research is becoming not only more distributed and collaborative, but more highly instrumented. Digital libraries provide a means to capture, manage, and access the data deluge that results from these research enterprises. We have conducted research on data practices and participated in developing data management services for the Center for Embedded Networked Sensing since its founding in 2002 as a National Science Foundation Science and Technology Center. Over the course of eight years, our digital library strategy has shifted dramatically in response to changing technologies, practices, and policies. We report on the development of several DL systems and on the lessons learned, which include the difficulty of anticipating data requirements from nascent technologies, building systems for highly diverse work practices and data types, the need to bind together multiple single-purpose systems, the lack of incentives to manage and share data, the complementary nature of research and development in understanding practices, and sustainability.

Categories and Subject Descriptors
H.3.7 [Information Storage and Retrieval]: Digital Libraries – *collection, standards, user issues.*

General Terms
Management, Design, Human Factors, Standardization.

Keywords
Data deluge, cyberinfrastructure, eScience, distributed research, collaborative research

1. INTRODUCTION
As the "fourth paradigm" [1] of data-driven science becomes a reality, scientists and funding agencies alike are presuming that scientific data are being captured and curated for reuse by others. Capture and reuse are occurring in "big science" fields such as physics, astronomy, and genomics. The smaller and more artisanal sciences – those that rely on collecting physical samples and whose methods and instruments are less standardized – also are experiencing a data deluge, but these data are less likely to be captured and curated for reuse. These small sciences also have become more instrumented through the use of technologies such as wireless sensing systems [2]. The increase in volume and variety of data in these research areas suggests an increased demand for digital libraries to manage and to share those data.

We have been documenting and facilitating the data practices of a distributed, collaborative, and interdisciplinary research center, the Center for Embedded Networked Sensing (CENS), since its inception in 2002. CENS is a National Science Foundation (NSF) Science and Technology Center based at UCLA, with four other partner institutions in California [3]. The mission of CENS is to develop sensing systems for scientific and social applications through collaborations between scientists, computer scientists, engineers, and experts in other domain areas. CENS initially received five years of NSF funding, from 2002 to 2007; funding was renewed for another five years, from 2007 to 2012. Over 300 faculty members, students, and research staff are now associated with CENS. Technology research partners in CENS include computer scientists, electrical engineers, and environmental engineers, while application scientists include seismologists, terrestrial ecologists, and aquatic biologists. Other members of the Center come from urban planning, design and media arts, and information studies.

CENS began in the early days of eScience [4], predating the report of the NSF Blue Ribbon Panel on Cyberinfrastructure [5]. CENS' evolution is a mirror for science, technology, and policy issues that have arisen in eResearch [6]. Of the many lens through which CENS and cyberinfrastructure can be viewed, our perspective in this paper is the role of digital libraries for data, and how those roles have evolved over the course of eight years of interdisciplinary research on embedded sensor networks [7; 8; 9; 10; 11; 12; 13; 14; 15]. Not the least of our challenges in digital library research is defining "data" and "sharing." This brief paper is the story of our experiences.

2. THE GRAND VISION
When CENS began in 2002, one of the founding co-investigators (Borgman) foresaw the need to develop digital library services for the Center in parallel with its growth. At the beginning, both the technology and science applications were under development; no scientific data were "streaming" from the networks yet. We had a rare opportunity to conduct formative evaluation of CENS' data management requirements throughout the life span of the Center. In 2002, we hired Kalpana Shankar as a post-doctoral fellow, who had just completed a dissertation on scientific archiving practices [16], to assist in developing a data management plan. The product of her year with CENS was a 40-page white paper [17].

JCDL 2010, July, 2010, Gold Coast, Queensland, Australia.

In that paper, which looks remarkably prescient nearly seven years later, she laid out "The Case for an Integrated Information and Data Management Plan" as follows:

- The integrity and accessibility of data in CENS is crucial to its research and educational activities and to the multiple research groups within CENS.
- An increased emphasis by funding agencies on data sharing and the re-use of expensive data and resources will mandate that large data providers implement plans for managing and disseminating data and other products.
- Appropriate data management and preservation strategies will make data available for longitudinal and multidisciplinary research, as well as for multiple audiences and purposes.
- Data and information must be preserved for risk management. Risks include the threats of data loss, the need to protect the intellectual work of CENS investigators, and the continued usability of data that may be needed to corroborate scientific claims.

Following Dr. Shankar's analysis of the CENS data requirements, we developed grant proposals and investigated an array of alternative mechanisms to build such a CENS data repository, to little avail. The reasons for not building that repository were several. One was that we were a bit ahead of our time – in 2003, few recognized the impending data deluge, much less the need to plan for how to manage it. Our arguments fell on deaf ears.

Another reason was the lack of actual data to demonstrate the need for a data repository. Much of the first three years of the Center was devoted to developing the sensing technology. Battery life was a much larger technical barrier than anticipated, for example. The data that were produced tended to be small and experimental, and more useful to the technology researchers – for whom the presence of data was an indicator that the technology was working – than to scientists.

Seismology was the most technologically mature discipline represented in CENS, and they already had a disciplinary repository for their data [18]. The only environmental science partner in CENS with a public data repository was the James San Jacinto Reserve [19]. The James Reserve was already streaming sensor data such as micro-meteorological records from weather stations to networked hubs [20]. These data are downloaded regularly from the hubs and stored in a MySQL database.

We considered whether the James Reserve system could serve as a basis for a larger CENS repository. We found that their observational data could serve as useful baseline data for experiments by other investigators, but that those data could not easily be integrated with the diverse experimental and laboratory observations being collected by CENS scientists [4]. The James Reserve data were more easily compared to static sensing deployments, which was the initial scientific model for CENS, than for periodic campaigns of data collection, the direction in which CENS research was then heading [15].

In lieu of a comprehensive CENS data repository, we considered a federated approach. At least three XML schema existed that were relevant for CENS data: Ecological Metadata Language [21], Sensor Modeling Language [22], and Open GIS [23]. In principle, they could be integrated into a larger schema or used to federate multiple XML repositories across CENS. This approach was not feasible at the time. One reason was the state of development of these standards. They were nascent, and we were not in a position to participate in the development community. A second reason was that few, if any, CENS investigators were adopting these standards for their own data. Scientific and technological research at CENS was already pushing boundaries; no one had sufficient resources or incentives to work at the edge of digital library development at the same time. In the long run, however, XML-based models appear to be the most feasible method of federating sensor network data [24].

In retrospect, our plans to build a comprehensive CENS data archive appear grandiose and naïve. We thought we could hire a data archivist, build XML-based repositories, and the data would come. But little data were being generated in the early days of CENS, and what data did exist was not of adequate scientific or technical quality to share, or if they were "good data," they were not marked up in a sharable manner. Where discipline-based metadata or markup formats existed, the CENS researchers were not using them. We had hoped to understand data and metadata requirements in advance of substantial data production so that we could be prepared to manage them in digital libraries. However, "the data" was too much of an abstraction to our research community to be able to identify needs effectively in advance.

3. CENS DATA

As a result of our interviews and field research, we were able to classify CENS data into four categories [11], which are shown in Figure 1. While this figure is specific to field deployments in ecology, the model is a useful framework for comparing types of data. In these types of deployments, 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 water and soil samples.

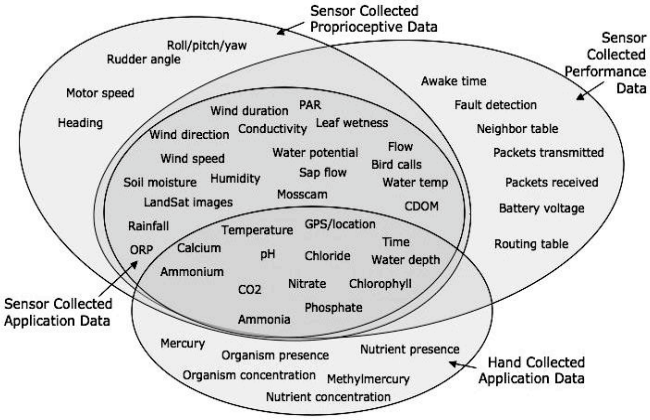

Figure 1: CENS data organized by collection method and use.

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.

4. SENSORBASE AS A SOLUTION

By 2005, three years into the Center's progress, it was clear that the "top down" model proposed by Dr. Shankar was not feasible for such a loosely coupled set of research projects and diverse array of data types. Multiple research teams controlled their own sets of sensors and associated data. No common data format existed that transcended the array of disciplines, research problems, and technologies involved in CENS. The statistics group, formed in 2005, proposed a complementary approach, which was to stream data from sensors directly into a database application. Their efforts resulted in SensorBase, which was a "bottom up" approach to creating simple data structures. SensorBase provided sufficient flexibility that any researcher at CENS, no matter what the domain or variables, could develop an appropriate data structure [25].

SensorBase incorporates RSS technology to support blogging of sensor data ("sensor logging," or "slogging") from the sensors to the database. The idea behind slogging is to allow researchers to post their sensor data online with the same ease by which they post content to blogs and other web 2.0 platforms. Each slogging project takes the form of a customized online database; new data are integrated into the structure on a regular basis. Each stream, in theory, is sufficiently structured that users can subscribe to streams and get new data pushed at them. They can also construct queries to request the delivery, for example, of all temperature data streams that had a 5-degree change in a 5-minute time period. Another feature is "Data RSS (dRSS)" that allows data streams to produce email alert messages, for example if the temperature on a sensor rises above a certain level.

In the alpha release of SensorBase, we (as members of the data practices research group) set up accounts, created test projects, and explored projects from other users. Many other users set up test projects; we found that few "real" projects were established. The group that most embraced SensorBase was the James Reserve. They were having problems with their own servers, and after a very close call with a forest fire that nearly destroyed the research station, saw the benefit of having a backup 100 miles away in Los Angeles. When they next upgraded their sensing equipment, they changed their storage mechanism so that their data streams directly from the sensors into SensorBase [25].

Our work with the SensorBase team informed our larger agenda of studying data practices as input to the design of data curation and sharing mechanisms. SensorBase is more of a database tool for data collection and analysis than it is a digital library for managing, curating, and serving data to users. It enables researchers to transfer sensor data from the field directly into a shared database. Users can define database structures, variable names, data types, and variable descriptions. They can write free text notes on projects, although not on data tables, as any individual project may have multiple data tables. This level of description proved to be sufficient for internal use by one-person projects and small teams. SensorBase's flexibility as a tool runs counter to digital library notions of metadata, structure, and consistency, however. As a result, the degree of data description provided rarely appears to be sufficient for prospective users to interpret the data for secondary analysis.

By focusing on datasets rather than on projects, considerable duplication of effort occurs in creating SensorBase records from one project deployment to the next. Data tables can be copied and edited, but not project descriptions. SensorBase is oriented toward projects whose sensor networks have live online connections so that data can be streamed to the database and to users through RSS feeds. Many CENS projects, especially those that conduct data collection campaigns – ranging from a few hours to a few days – do not have this capability. Such projects will gather data in the field onto laptops and then transfer them to SensorBase or other data management tools upon returning from the field. The latter type of projects thus could not take advantage of the streaming data features of SensorBase.

SensorBase is itself a research project, not an endeavor to build a production-grade repository. Similarly, our research on data curation, usability, and access contributes to the requirements definitions, but we are not in a position to build or to manage production-level repository services. The tension between research and infrastructure became ever more evident through the evolution of the SensorBase project – a tension now recognized as a characteristic of eScience and cyberinfrastructure endeavors [26].

5. CENSDC AS A SOLUTION

As CENS grew from a few dozen participants to over 100, oral culture no longer sufficed to transfer knowledge within and between research teams. The need for consistent data capture and management was increasingly evident. Whether or not the data, per se, were being collected and curated, it became more necessary to capture contextual knowledge about how to conduct this new kind of science and how to use these new technologies effectively. New teams would go out on field deployments, only to find they lacked some critical expertise or equipment. Many of these sensor network deployments were at remote sites, 100 miles or more away from equipment stores and other project members.

We, the data practices team, undertook development of the CENS Deployment Center (CENSDC) as a means to capture, manage, and reuse critical information about deployment activities. The CENSDC is a web-based system that allows researchers to create deployment plans and to gather post-deployment feedback. Its capabilities are based on our prior studies of field activities. CENSDC includes data fields for deployment dates and locations, for lists of people, equipment, and planned tasks (including who is responsible for each task), and general notes. A post-deployment "debriefing" questionnaire asks questions about lessons learned, problems encountered, and how they were solved or might be solved for the next deployment. Thus CENSDC enables researchers to plan deployments and to collect contextual information that can be leveraged to describe the resulting data.

The process of implementing the CENSDC informed our data practices research in several respects. First, we found that the culture of ad hoc planning is very strong in some of the field-based sciences, particularly the environmental and ecological sciences, which were our initial target users. Researchers in these disciplines conduct their field activities in a flexible way, adjusting activities to unpredictable weather, varied flora and

fauna, and unreliable equipment. Thus, spending time creating a formal plan was not a high priority for some of our targeted users. However, most teams appreciated the assistance of a graduate student from the data practices team to work with them in the field and to help document their processes. After a series of initial requests to participate in field deployments, some teams started taking the initiative to invite our students to join them.

A second lesson was that deployment planning was conducted with a variety of tools that are not easily integrated. Email remains the main planning tool for most teams, with plans circulating via mailing lists. (CENS has about 100 dedicated distribution lists and about a third of them are heavily used for deployment planning). Planning also relies on basic computer applications such as Word documents and Excel spreadsheets. Some research groups in CENS now use collaborative web tools, such as Google Docs and Spreadsheets, to track equipment lists and other deployment information. CENSDC development predated the wide availability of such shared online tools. While these tools serve some of the CENSDC functions, they are team-specific. An important goal of CENSDC is to share knowledge across the CENS community, as much duplication exists in the use of sensor technologies, other kinds of equipment, and field deployment activities across research teams and disciplines.

Yet a third lesson in data practices from CENSDC is the diversity of deployment activities. In constructing a typology of deployments, we found that they vary in duration, spatial extent, equipment complexity, number of people involved, focus, and outcome. Some deployments involved installing sensors for months at a time in static locations, and some deployments involved campaigns of three to five days with movable sensor systems. Some research groups repeated the same kinds of deployments at multiple sites or at the same site, while other groups never perform the same activities on consecutive deployments.

We continue to work on synergies to be gained by integrating the CENSDC and SensorBase systems. SensorBase can capture the sensor data per se, while CENSDC can capture the description of the deployment where the data were collected. If these objects can be linked effectively, each becomes more valuable: one can discover data in SensorBase, then obtain descriptions of the scientific activities from which it arose; conversely, one can discover a deployment with a set of scientific activities and then obtain the resulting dataset. Further, the data and deployment information can be linked to resulting publications, embodying the value chain of those scholarly activities [6]. We have developed the conceptual framework for instantiating these relationships using the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) protocol, and are working on the technical implementation [24].

6. DATA DISCOVERY LIBRARY AS A SOLUTION

The National Science Foundation, which is the primary funder for CENS, quite reasonably expects our scholarly products to be disseminated as widely as possible. Until recently, dissemination took the form of publication lists in quarterly and annual reports to the Foundation. In 2007, the data practices team launched the effort to construct a CENS publication repository as part of the University of California eScholarship repository [13]. This has been a very successful effort in making CENS publications more visible; they are downloaded frequently from the repository.

At our most recent annual NSF site visit to CENS (spring, 2009), we were asked specifically to make our data more widely available. This request was music to the ears of the data practices team, and we began work to consolidate our efforts in data management and data policy toward an action plan. We explored a number of avenues. One obvious path was to contribute CENS data to the UC eScholarship repository. However, they are having great difficulty handling the present diversity of CENS scholarly products and are not in a position to accept our data, unfortunately.

Drawing on our research and experience with constructing digital libraries for CENS data, our new strategy is to construct a metadata repository that enables CENS data to be discovered by potential users. We chose to focus on a Data Discovery Library for several reasons. One reason is the diversity of scholarly products that might be considered "data." Among the data resources collected by CENS researchers are images, audio files, physical samples, and numeric data in both digital and analog form. Another reason is the distribution of these resources around community, lab, and individual computer systems. Some CENS research data and documentation are available online through lab websites, but large portions of these resources reside in protected computer systems, personal laptops, file cabinets, and even in refrigerators. Collecting and integrating all of the Center's data into a single system would be prohibitively expensive and time consuming, even if investigators were willing to release them. We are encouraged to share our data, but no mandate exists to provide them. Such a mandate would require a clear definition of data, and clear policies for what is to be released, in what form, and when.

Lacking such mandates, definitions, and policies, we are designing a metadata repository that will enable potential data users to discover what CENS data exist, to determine whether those data may be useful, and to learn how to acquire data of interest. A record in the Data Discovery Library may link directly to a dataset or it may provide contact information for individuals who can provide access to the data. We are investigating policies for contributing to the Data Discovery Library, such as an embargo for one year from the date of collection or one year from the date of initial funding.

The Data Discovery Library is expected to serve internal needs of CENS as well as dissemination of our scholarly products. The system will assist CENS researchers in keeping track of their own data resources and will provide the Center's administrators with additional documentation of our research output. It should also promote the sustainability of CENS data and other scholarly products beyond the NSF funding of the Center, which ends in 2012. Thus, we are focusing on designing the system to be lightweight, i.e., easy to use with minimal assistance. Additionally, we are using open source software tools for the back end database and web display.

The fields in our prototype are drawn from the Dublin Core metadata set [27], as shown in Table 1. The Dublin Core metadata set was chosen for its flexibility and simplicity in providing descriptive fields for resource discovery. Discipline-specific metadata schemas were considered, such as the Ecological Markup Language and SensorML, but these were deemed to be too inflexible for the diversity of research and data types found in the center.

Table 1. Metadata Fields used for Prototype Testing

Data Description fields	Dublin Core element*
1. CENS project name	title
2. CENS research group	publisher
3. dates (of collection)	date
4. place	coverage
5. people	-
- contact person	creator
- other particpating researchers	contributor
6. data type	type
7. data description	-
- research question (why collected)	description
- what collected (variables)	description
- data collection process and equipment	description
- size, format	format
8. related publications (eScholarship URL)	relation
9. related deployment info (CENSDC URL)	relation
10. keywords	subject
11. location of the data (URL)	identifier
12. permissions	rights
13. funding source	source

*the Dublin Core element "language" is not used.

Testing and implementation of the Data Discovery Library began in February, 2010. We are leveraging the existing NSF reporting cycle to acquire the initial metadata about CENS datasets. To fulfill NSF reporting requirements, CENS' staff annually solicit metadata about a variety of scholarly products such as publications, presentations, and new grants. This information is useful not only for the NSF, but is a means for CENS to document its return on investment to our many other partners and funders. Moreover, the collection and preservation of CENS' scholarly and scientific output in a structured format enables us to construct social and bibliographic networks of collaboration [28]. This year, each research group was also asked to report their datasets using a new automated Annual Report System.

A few weeks before the annual reports were due, this solicitation text was sent to CENS researchers explaining the new dataset reporting module: "NSF has asked that we report on data sets that have been collected as a part of CENS research. This includes data sets created or contributed to as part of the research being reported on during the reporting cycle. By reporting the existence of a data set you are not giving up your rights of ownership. The following metadata for these data sets should be reported in the online submission system. We recommend that you assemble this information along with the report narrative. The metadata elements are: title, collection start and end date, location, people who contributed to data collection, data type (image, numerical, etc.), research question, variables, data collection process, data format, permissions, funding source(s), keywords, location of data set, related publications, and related CENS deployment center record. These metadata descriptions will be posted to the CENS website to facilitate data discovery."

As of this writing (mid-April, 2010), the annual reporting process is coming to a close and CENS' staff are aggregating the various report pieces from each research group. Of the roughly 65 reports submitted to date, which is estimated to be a 90-95% response rate from all of the CENS research groups, only 11 datasets were submitted. Four of these 11 datasets were submitted by the data practices group, which includes the authors of this paper. The other seven datasets consist of four from the Participatory Sensing group (which uses mobile phones as sensors), and one each from Contaminant Transport and Management, Multiscale Actuated Sensing, and our statistics partners in the Statistics and Data Practices group. No datasets were reported by other science groups such as Seismic, Aquatic, or Terrestrial sensing, or by the other computer science and engineering groups in this first round. Based on our previous research, it is likely that at least one dataset fitting the reporting criteria could be associated with each of 70 or so individual research reports.

This low response rate can be attributed to a variety of factors. First, the CENS solicitation was ambiguous about whether reporting datasets was mandatory. The solicitation language was carefully written to reflect the facts that NSF has specifically requested that CENS report datasets, but that datasets are not a mandatory section of the NSF official reports. Second, the reporting process changed in several ways this year, creating some confusion. This is the first year in which CENS used an online reporting platform, where each report section is entered into a webform. In prior years annual reports were submitted manually, i.e., documents in a template sent by email. The template included only the sections mandated by the NSF (description of research, people, publications, etc). Third, the new Annual Report System impeded dataset submission in two crucial ways: 1) the new system did not require the majority of fields, resulting in gaps in the dataset, publication, and general report metadata, and 2) the user interface was difficult to use. We anticipate that these problems can be reduced significantly in the second version of the Annual Report System. The system was implemented somewhat prematurely, lacking some planned interface features, due to the hard deadlines for NSF annual reports.

A fourth reason for the low response rate, and one of particular interest for our data practices research, is the diffuse responsibility for datasets. Multiple research groups may be involved in the production of any given dataset. Conversely, any given deployment may result in multiple datasets used by different researchers and teams. Other data are produced in laboratories or in simulation studies. Once collected, different individuals may analyze, manage, and report these data in publications. It is often unclear which individual has ultimate responsibility for any given dataset or who should take responsibility for including it in an annual report. The sections of annual reports often are delegated to graduate students who are the leads on specific projects. Faculty team leaders review the reports and synthesize results across research areas of CENS. The determination of who has ultimate responsibility for a given dataset is among our research questions.

Despite the low response rate for datasets in the first iteration of the Data Discovery Library, the reported datasets will be listed in the CENS annual report to the NSF. This is the first year that CENS will be able to report datasets in an annual report, which is a significant step forward in increasing the availability of CENS datasets for secondary uses. The metadata from these datasets are not yet exposed on the CENS website, but will be within a few months.

As a means to promote an ethic of sharing data, the Data Discovery Library will be OAI-compliant. One of our research and development projects [24] is to make the system generate OAI-ORE aggregations automatically. ORE aggregations will expose links between scientific datasets and publications for discovery, reuse, and replication. We are assessing how to display the dataset metadata. One possibility is a two-level metadata display with a short record and full record, following the examples of the National Evolutionary Synthesis Center and the UNC Metadata Research Center's Dryad dataset repository [29; 30].

Our next iteration of the Data Discovery Library will focus on increasing our advocacy for contributing datasets and improving the usability of the Annual Report system. We will make individual inquiries about why CENS researchers do and do not report datasets in their Annual Reports. We will also present the Data Discovery Library in the CENS weekly seminar series, and follow up with individual instruction. Usability improvements will focus on simplifying the data entry process, for example by auto-populating repeated fields such as people's names, research sites, dates, and equipment. We will present initial evaluation results from these next steps at the JCDL 2010 meeting.

7. LESSONS LEARNED

Our participation in CENS since its inception in 2002 has provided the rare opportunity to study data practices throughout the evolution of the science, technology, and collaborative activities of a major research center. We foresaw the need to capture and manage CENS data for use and reuse by others. However, our initial vision for how that would be done was far too monolithic, and presumed far more consistency in activities and in data types than was the case. Over the eight years of the Center's existence, research activities have evolved, becoming more diverse while also becoming more mature. We have studied the data practices and management activities of CENS researchers via interviews, field studies, and participation in the construction of multiple digital libraries and other types of data-handling systems.

These are the essential lessons we have learned about digital libraries and data from our eight years of data practices research and development in CENS:

An early lesson was the difficulty in determining requirements for a data digital library prior to the time that "real data" were available from the sensor networks. The science and technology partners got bogged down initially in discussions wherein the computer scientists and engineers asked the scientists, "what shall we build for you?," and the scientists responded by asking, "what can you build for us?" The resulting data, and how to manage them, were far from their immediate concerns. Once that cycle was broken, collaborative projects became operational, data began to flow, and our conversations became more concrete.

A second lesson was that the diversity of work practices and of data confound attempts to build both unitary "grand" systems and smaller federated systems. Not only did the types of data vary widely, so did the means by which they were produced. Some teams installed static sensor networks that would stream the same data continuously. Others collected data on short-term campaigns, bringing datasets home from the field with them. Some teams repeated their campaign deployments at multiple sites with common instruments and variables. Other teams started afresh with equipment, variables, and research questions for each campaign. No common solution fit these loosely coupled practices for data collection. The teams also varied greatly in how consistently they managed their data, once acquired.

A third lesson is that building multiple systems for different kinds of resources, such as data sets and deployment information, can be effective at the level of the individual components. However, binding the systems together is a significant and ongoing challenge. We continue to address the technical and political constraints on linking records of data, publications, and deployment records, for example.

A fourth lesson is that without external pressure, data management – in the form of digital libraries for capture, curation, and access – is not high on the priority list of science and technology researchers. Our data practices team has had the good will, and considerable funding, from CENS to study and develop digital library services. As eScience and cyberinfrastructure matured, and as CENS data production reached a level that researchers became concerned about maintaining their own data, interest in our work grew. Now that external pressure is increasing, the Center's management is eager to take advantage of our knowledge of data requirements. In turn, we have welcomed the opportunity to give back to the Center.

The fifth lesson is that our social-scientific studies of data practices and our involvement in technical development of digital library services for CENS have complemented each other in essential ways. If we had constrained our roles to social scientific observation, we would not have encountered the technical challenges of modeling and integrating data. If we had attempted to build systems without studying practices in depth, we would have missed the mark widely, investing in digital libraries that were unlikely to have been used effectively.

Lastly is the lesson of sustainability. CENS has been a tremendously successful scientific and technical research center in terms of research productivity, collaboration, and interdisciplinary interactions. Making the scholarly products of the Center available – the publications, data, and other resources – is only one part of the sustainability challenge. Capturing and maintaining the knowledge gained and the socio-technical network developed is a far greater concern for CENS, its partners, its funders, and the larger scholarly community.

As CENS has evolved, so has the larger social and political framework for data sharing. Now that we have external pressure to share our data, we have more mechanisms to establish policies and systems. Our approach to sharing data has come full circle from building a data repository to building a metadata repository that will enable CENS data to be discovered. Once discovered, prospective users can obtain data from wherever they may be located, and from whoever has the rights and ability to release them. CENS scientific and technological activities have evolved at Internet speed over the course of eight years. The focus shifted

from static scientific deployments to campaigns, and from single-purpose sensing technologies to cell phones as mobile sensing devices. Concurrently, the Internet has moved from basic web services to "web 2.0" and cloud computing. Digital library technologies and services have not evolved as rapidly as the applications they serve. We've worked intensively to keep pace with these moving targets over the last eight years, and the pace shows no signs of slowing. We are not much closer to defining either "data" or "sharing" than we were in 2002, but that is largely because these concepts have been shifting under our feet during this entire process.

8. ACKNOWLEDGMENTS

The research reported here is supported by a series of grants from the National Science Foundation and by gifts from the Microsoft Technical Computing Initiative and Microsoft External Research. NSF grants include: *Center for Embedded Networked Sensing, Science and Technology Center*: Cooperative Agreement #CCR-0120778, Deborah L. Estrin, UCLA, Principal Investigator; *CENS Education Infrastructure (CENSEI):* #ESI- 0352572, William A. Sandoval, PI; Christine L. Borgman, co-PI; *Towards a Virtual Organization for Data Cyberinfrastructure*, #OCI-0750529, Christine L. Borgman, UCLA, PI; Geoffrey C. Bowker, Santa Clara University, Co-PI; Thomas Finholt, University of Michigan, Co-PI; *Monitoring, Modeling & Memory: Dynamics of Data and Knowledge in Scientific Cyberinfrastructures:* #0827322, Paul N. Edwards, University of Michigan, PI; Christine L. Borgman, UCLA, Co-PI; Geoffrey C. Bowker, University of Pittsburgh, Co-PI; Steven Jackson, University of Michigan, Co-PI; David R. Ribes, Georgetown University, Co-PI; and Susan Leigh Star, University of Pittsburgh. We also thank Kalpana Shankar, now of Indiana University, for comments on this paper, and our Microsoft partners, Catherine van Ingen and Catherine C. Marshall, for their thoughtful commentary on our research designs and on our interpretation of findings.

9. REFERENCES

[1] Hey, T., Tansley, S. & Tolle, K. (Eds.). (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA: Microsoft. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/ on 16 December 2009.

[2] Embedded, Everywhere: A Research Agenda for Networked Systems of Embedded Computers. (2001). Washington, D.C.: National Academy Press. Retrieved from http://www.nap.edu/ on 11 March 2005.

[3] *Center for Embedded Networked Sensing*. (2009). Retrieved from http://research.cens.ucla.edu on 14 April 2009.

[4] Hey, T. & Trefethen, A. (2005). Cyberinfrastructure and e-Science. *Science,* 308: 818-821.

[5] Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messina, P., Messerschmitt, D. G., Ostriker, J. P. & Wright, M. H. (2003). *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon panel on Cyberinfrastructure*. National Science Foundation. Retrieved from http://www.nsf.gov/cise/sci/reports/atkins.pdf on 18 September 2006.

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

[7] Borgman, C. L. (2004). The Interaction of Community and Individual Practices in the Design of a Digital Library. *International Symposium on Digital Libraries and Knowledge Communities in Networked Information Society*, University of Tsukuba, Tsukuba, Ibaraki, Japan., University of Tsukuba. Retrieved from http://www.kc.tsukuba.ac.jp/dlkc/e-proceedings/papers/dlkc04pp9.pdf on 10 April 2006.

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

[9] 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,* 7(1-2): 17-3029 September 2007.

[10] Borgman, C. L., Wallis, J. C., Enyedy, N. & Mayernik, M. S. (2006). Capturing habitat ecology in reusable forms: A case study with embedded networked sensor technology. *Society for the Social Studies of Science*, Vancouver, BC.

[11] Borgman, C. L., Wallis, J. C., Mayernik, M. S. & Pepe, A. (2007). *Drowning in data: Digital library architecture to support scientific use of embedded sensor networks*. Vancouver, British Columbia, Canada, Association for Computing Machinery: 269-277. Retrieved from http://doi.acm.org/10.1145/1255175.1255228 on June 17-23, 2007 Accessed.

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

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

[14] Wallis, J. C., Borgman, C. L., Mayernik, M. S. & Pepe, A. (2008). Moving archival practices upstream: An exploration of the life cycle of ecological sensing data in collaborative field research. *International Journal of Digital Curation,* 3(1). Retrieved from http://www.ijdc.net/ijdc/issue/current on 24 November 2008.

[15] Wallis, J. C., Borgman, C. L., Mayernik, M. S., 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. 380-391.

[16] Shankar, K. (2002). *Scientists, Records, and the Practical Politics of Infrastructure*. PhD Dissertation, Department of Information Studies: University of California, Los Angeles.

[17] Shankar, K. (2003). *Scientific data archiving: the state of the art in information, data, and metadata management*. Retrieved from http://works.bepress.com/borgman/234 on 30 January 2010.

[18] *Incorporated Research Institutions for Seismology*. (2010). Retrieved from http://www.iris.edu/hq/ on 1 February 2010.

[19] *UC James San Jacinto Reserve Data Management System*. (2010). Retrieved from http://dms.jamesreserve.edu/ on 1 February 2010.

[20] Szewczyk, R., Osterweil, E., Polastre, J., Hamilton, M., Mainwaring, A. & Estrin, D. (2004). Habitat monitoring with sensor networks. *Communications of the ACM,* 47(6): 34-40.

[21] *Ecological Metadata Language*. (2010). Retrieved from http://knb.ecoinformatics.org/software/eml/ on 1 February 2010.

[22] *Sensor Modeling Language*. (2010). Retrieved from http://vast.uah.edu/SensorML/ on 1 February 2010.

[23] *Open Geospatial Consortium*. (2010). Retrieved from http://www.opengeospatial.org/ on 1 February 2010.

[24] Pepe, A., Mayernik, M., Borgman, C. L. & Van de Sompel, H. (2010). From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web. *Journal of the American Society for Information Science and Technology,* 61(3): 567–582. Retrieved from http://www3.interscience.wiley.com/journal/123214737/abstract on 1 February 2010.

[25] Chang, K., Yau, N., Hansen, M. & Estrin, D. (2006). SensorBase.org - A Centralized Repository to Slog Sensor Network Data. *Proceedings of the International Conf. on Distributed Networks(DCOSS)/EAWMS*.

[26] Edwards, P. N., Jackson, S. J., Bowker, G. C. & Knobel, C. P. (2007). *Understanding Infrastructure: Dynamics, Tensions, and Design*. National Science Foundation: University of Michigan. Retrieved from http://hdl.handle.net/2027.42/49353 on 26 July 2007.

[27] *The Dublin Core Metadata Initiative Terms*. (2009). Retrieved from http://dublincore.org/documents/dcmi-terms/ on 14 April 2009.

[28] Pepe, A. & Rodriguez, M. A. (2010, forthcoming). Collaboration in sensor network research: an in-depth longitudinal analysis of assortative mixing patterns. *Scientometrics*. Retrieved from http://www.springerlink.com/content/v1w5695932tg52g2/ on 1 February 2010.

[29] *Dryad*. (2010). Retrieved from http://datadryad.org/ on 12 April 2010.

[30] Greenberg, J., White, H. C., Carrier, S. & Scherle, R. (2009). A Metadata Best Practice for a Scientific Data Repository. *Journal of Library Metadata,* 9(3): 194-212.