Who's got the data? Interdependencies in Science and Technology Collaborations

Christine L. Borgman*, Jillian C. Wallis*, Matthew S. Mayernik**

*Department of Information Studies and Center for Embedded Networked Sensing
University of California, Los Angeles

borgman@gseis.ucla.edu, jwallisi@ucla.edu,

**National Center for Atmospheric Research

mayernik@ucar.edu

Status and citation: This is the revised and accepted version, prior to publisher's copy editing. Please quote the final version:

Borgman, C. L., Wallis, J.C., Mayernik, M. S. (In press). Who's Got the Data? Interdependencies in Science and Technology Collaborations. Journal of Computer Supported Cooperative Work. DOI: 10.1007/s10606-012-9169-z

This article is available "Online First" on SpringerLink

http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s10606-012-9169-z

Table of Contents

ABSTRACT

Science and technology always have been interdependent, but never more so than with today's highly instrumented data collection practices. We report on a long-term study of collaboration between environmental scientists (biology, ecology, marine sciences), computer scientists, and engineering research teams as part of a five-university distributed science and technology research center devoted to embedded networked sensing. The science and technology teams go into the field with mutual interests in gathering scientific data. "Data" are constituted very differently between the research teams. What are data to the science teams may be context to the technology teams, and vice versa. Interdependencies between the teams determine the ability to collect, use, and manage data in both the short and long terms. Four types of data were identified, which are managed separately, limiting both reusability of data and replication of research. Decisions on what data to curate, for whom, for what purposes, and for how long, should consider the interdependencies between scientific and technical processes, the complexities of data collection, and the disposition of the resulting data.

KEY WORDS

Cyberinfrastructure; data curation; data practices; eScience; scientific collaboration, scientific software development; technology research; sensor networks; environmental sciences.

1. INTRODUCTION

Today's scientific instrumentation yields data of unprecedented volume and granularity. Some of these instruments are constructed for scientific innovation; others for technological innovation. The best case is when innovations of both kinds occur, which in turn depends on effective collaborations between scientific and technical researchers. Data are the "glue" of a collaboration, hence one lens through which to study the effectiveness of such collaborations is to assess how they produce and use data. The research reported here is drawn from a series of studies conducted in the Center for Embedded Networked Sensing, a National Science Foundation Science and Technology Center. We address the following research questions:

1. What are the "data" in science and technology research collaborations?
2. How do concepts of "data" vary by purpose of research activity?
3. What roles do data serve within and between science and technology collaborations?
4. What can be learned about collaborative work by following the data?

We first examine concepts of data in scientific and technical research practice. We then introduce our research setting, the Center for Embedded Networked Sensing (CENS), a science and technology collaboratory, and detail the data practices of the CENS science and technology researchers, respectively. Our discussion identifies ways in which those data practices are interdependent, implications of those interdependencies for collaborative research projects, and implications for curation of the data that those

projects produce. By studying how data are created, conceived, handled, managed, and curated in multi-disciplinary collaborations, we can inform science policy and practice.

2. SCIENCE, TECHNOLOGY, AND DATA

Data are both visible and invisible in scientific collaborations. The empirical studies reported here address the roles of data in collaborations between science researchers and technology researchers. They work together to develop new instrumentation and to deploy it for field research. The data produced by these joint projects sometimes is shared and sometimes is separate. One investigator's data may be context to another investigator, and vice versa. By focusing on the data, we can identify the collaborative tensions, the short and long term goals of the research teams, and the successful (and unsuccessful) practices in the capture, management, and use of data. Data curation – adding value through documentation, standardization, migration to new formats – is essential for long-term use and reuse of data. Public policy for data management plans and data sharing, in turn, depends upon the proper care and curation of data from scientific and technological research. Thus studies of data and data practices have implications for social policy as well as for cooperative work.

2.1. Science, Technology, and Sensor Networks

Scientists learn how to use tools, instruments, and materials to accomplish their research as part of membership in the community. They sometimes build their own tools, adapting

them to the task or question at hand. In larger projects, scientific and tool-building activities usually are done by different people, resulting not only in collaboration but in interdependencies (Fry, 2006; Shrum, Genuth & Chompalov, 2007). Technologies for data-intensive science include hardware, such as sensor networks, laboratory instruments, and telescopes, and software, such as workflow systems, analytic tools, and visualization tools.

The deployment of satellites for remote sensing, starting in the 1970s and 1980s, transformed many areas of the environmental sciences (Kwa, 2005); satellite sensing became an essential component of scientific infrastructure. Changes in weather and in crops could be seen from above, enabling geospatial and temporal comparisons never before possible. Embedded sensor networks, which are the inverse of remote sensing with satellites, are transforming scientific practice in much the same way. Many small and discrete sensors can be placed in situ to study local conditions, providing data at far greater spatial and temporal detail than is possible with hand sampling of soil, water, or plants. As many ecological phenomena happen at a scale smaller than a mile square, embedded sensor networks are far superior to remote sensing for studying local phenomena.

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 employ embedded sensor networks. In the U.S. alone, public regulatory agencies monitor several hundred million

individual sensors on streams, lakes, and rivers. The use of embedded sensor networks for scientific data collection emerged as a means to accelerate inductive and experimental data collection in research specialties heretofore characterized by flexible, hand-crafted field research (Embedded, Everywhere: A Research Agenda for Networked Systems of Embedded Computers, 2001).

Sensor networks are now used widely to collect observational data for environmental research. They fall into two general categories: static sensor networks and dynamic deployments. Static networks, such as the LTER (U.S. Long Term Ecological Research Network, 2010), GEON, which originally was an acronym for Geosciences Network (GEON, 2010), and NEON (National Ecological Observatory Network, 2010) consist of sensors placed in appropriate positions to report data continuously on local conditions. The 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).

While static sensor network deployments are effective for large-scale data collection, participants must agree on common data structures, semantics, services, ontologies, and preservation policies. Once agreements and technologies are in place, static networks can stream data directly into data repositories. Sensor equipment must be robust enough to be left in the field for months or years. Security is a constant concern, as sensors left

unattended can be vandalized or can be damaged by animals, farm equipment, or weather.

Dynamic deployments, in contrast, are suited to experimental field research. "Human in the loop" sensor networks allow investigators to adjust monitoring conditions in real time. Teams can conduct data collection "campaigns" in which they deploy a sensor network in the field for a few hours or a few days. The dynamic nature of mobile equipment and reconfigurable networks allows sensors to be moved to maintain coverage of interesting phenomena or to adjust resolution to capture phenomena that were previously unobservable. Teams may return to the same site, or a similar site, repeatedly, each time with slightly different equipment or research questions. Dynamic deployments provide opportunities for science and technology researchers to collaborate in field research, especially in refining instrument design and data collection methods. Fragile, research-grade equipment can be deployed in campaigns, as teams are on site to protect, adjust, and maintain the devices. The inherent variability in the science conducted with dynamic sensor deployments results in data products that also are highly variable. Data variability is a constant challenge for digital data repositories (Mayernik, 2011; Mayernik, Wallis & Borgman, in review).

2.2 Data in Science and Technology

Data often are boundary objects, both bridging and demarcating the lines between communities (Star & Griesemer, 1989). A focus on data can reveal much about

communities and relationships (see for ex. Collins, 1998; Aronova, Baker, & Oreskes, 2010). "Data" is a difficult concept to define, as data may take many forms, both physical and digital. Among the most widely cited definitions is this one, from a National Academy of Sciences report: "*Data* are facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors." (A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases, 1999, p. 15). A more current working definition, from an internal Academy document, better reflects the variety of data types that may arise in collaborative work:

> The term "data" as used in this document is meant to be broadly inclusive. In addition to digital manifestations of literature (including text, sound, still images, moving images, models, games, or simulations), it refers as well to forms of data and databases that generally require the assistance of computational machinery and software in order to be useful, such as various types of laboratory data including spectrographic, genomic sequencing, and electron microscopy data; observational data, such as remote sensing, geospatial, and socioeconomic data; and other forms of data either generated or compiled, by humans or machines. (Uhlir & Cohen, 2011).

The above notion of "data" transcends the sciences and other domains of scholarship, acknowledging the many forms that data can take. Sources of data also vary widely. In the physical and life sciences, most data are gathered or produced by researchers, such as

by observations, experiments, or models. In the social sciences, researchers may gather or produce their own data, or they may obtain data from other sources such as public records of economic activity (Borgman, 2007; 2012).

Data also can be categorized by type or origin. The typology presented in an influential U.S. policy report (Long-Lived Digital Data Collections, 2005) and incorporated in National Science Foundation strategy (Cyberinfrastructure Vision for 21st Century Discovery, 2007) is now widely accepted. *Observational* data include weather measurements, which are associated with specific places and times, and attitude surveys, which also might be associated with specific places and times (e.g., elections or natural disasters), or involve multiple places and times (e.g., cross-sectional, longitudinal studies). *Computational* data result from executing a computer model or simulation, whether for physics or for cultural virtual reality. Replicating the model or simulation in the future may require extensive documentation of the hardware, software, and input data. In some cases, only the output of the model might be preserved. *Experimental* data include results from laboratory studies such as measurements of chemical reactions or from field experiments such as controlled behavioral studies. Whether sufficient data and documentation to reproduce the experiment are kept varies by the cost and reproducibility of the experiment. It should be noted that records of government, business, and public and private life also yield useful data for scientific, social scientific, and humanistic research.

The term "dataset" is sometimes conflated with the notion of "data." However,

definitions of "dataset" in the scientific literature have at least four common themes – grouping, content, relatedness, and purpose – each of which has multiple categories (Renear, Sacchi & Wickett, 2010). While "dataset" may be useful to refer to a collection of data for the purposes of citation, the term does little to clarify what is meant by data.

Data collections are larger than datasets, and can be categorized into three levels: *Research collections*, in support of small communities, which may not conform to broader standards; *Resource data collections*, which may serve larger communities and either set or abide by community standards; and *Reference collections*, which support large segments of the scholarly community and drive standards processes (Long-Lived Digital Data Collections, 2005). A collection may start as a small research collection and take on a greater role over time; the Protein Data Bank is the canonical example of this transition (Protein Data Bank, 2006; Berman et al., 2000; Bourne, 2005).

Efforts to define and categorize data, datasets, and data collections risk obscuring the complexity of work practices around data. These complexities include reward structures, authority structures, formalization of knowledge, interdependencies among groups, trust mechanisms, and the "transitional nature" of data collections (Cragin & Shankar, 2006). Data are deeply embedded in tacit knowledge and in local practices, which make them difficult to extract from their context (Kanfer et al., 2000). The challenge is to capture data in a sufficiently rich form that they can be interpreted, while making them "mobile" enough to manage in information systems. Cole (2008) refers to this process as "differentiation," acknowledging the lack of coherence and the need for scaffolding to

maintain the integrity of data. All too often, as he notes, information systems and policies simply take data as a given or else treat them as a commodity. These approaches either assume or impose coherence, in Cole's terms.

2.3. Computer Supported Cooperative Work in Science and Technology

Cyberinfrastructure is the point of intersection between the interests of those who study computer-supported cooperative work and those who study scientific and technical collaboration (Jirotka, Procter, Rodden & Bowker, 2006; Lee, Ribes, Bietz, Jirotka & Karasti, 2010; Turner, Bowker, Gasser & Zacklad, 2006). Ribes and Lee (2010) identified seven themes at this intersection. Their second theme, "integration of heterogeneity," focuses on interdisciplinary relations that must be fostered to develop "cyberinfrastructure and the novel science that it is hoped CI will thereafter bring about." Among the relationships they identify are those between "the computer and information scientists that develop the technological systems and on the other hand the domain scientists that collaborate with them to develop technology, but who are ultimately interested in the final stable and functioning scientific resources" (Ribes & Lee, 2010, 234). Similarly, Faniel and Jacobsen (2010) identified how the multiple contexts of data production in these collaborations influence the ability of researchers to evaluate and trust data for reuse. Context, curation, and dissemination of data are concerns that arise consistently in cyberinfrastructure collaborations (Karasti, Baker & Halkola, 2006; Karasti, Baker & Millerand, 2010).

Collaborations between researchers in science and in technology require a delicate balance between autonomy and interdependence. Shrum et al. (2007, p. 123-) distinguish between three types of collaborations involving scientific instruments: (1) Use of standardized and familiar instrumentation, where the science contribution is based on how the instruments are deployed; (2) adaptation of extant instrumentation to improve the science; and (3) design and construction of "an unprecedented instrument." For each type, the interdependencies varied temporally. For instance, the third type of collaboration required a high level of initial coordination, but once the equipment was in place collaborators were free to work at their own pace. While Shrum et al. were studying physicists, the analogy to sensor networks applies. Static sensor network deployments fall into the first category, while field deployments in CENS fall into all three categories, thus increasing the complexity of the data interactions between collaborators.

The type of instrument-based collaboration also influences the choice, use, and interpretation of data (Collins, 1975; Shrum et al., 2007). Some teams integrate their data collection and others collect data independently. Shrum et al. found that data sharing was most effective when standard protocols were in place. Collins, in a study of physicists, found that the aftermath of data collection is a most interesting time to study data practices, for that is the collaborative stage where participants explore the meaning of their data and phenomena.

2.3.1. Scientific Data Practices

In ecology, science takes the form of "naturalistic realism" in which models of reality are established and then compared to the real world (Giere, 1999; Maurer, 2004). The practice of ecological science is largely inductive, beginning with accumulation of observations in the field, with the intent to discover patterns. In the search for repeatable patterns, parameter estimation often is more useful than formal hypothesis testing. When patterns are well established, deductive methods may be applied in which data are gathered to test hypotheses (Maurer, 2004). Ecology researchers are studying complex systems that do not lend themselves as well to consistent methods and measures as do the physical sciences (Aronova, Baker & Oreskes, 2010).

Field-based research in ecology takes place in unpredictable real-world settings, making technology design particularly daunting. Data practices in ecological research can be associated with the four salient features identified by Bowen and Roth (2007): (1) research design in ecology has a highly emergent character; (2) tools and methods are developed in situ, often from locally available materials, and are highly context-specific; (3) studies are not easily replicable because of the dynamic nature of ecological systems; and (4) social interactions between members of the community are highly important. While all of these features exist to some extent in other scientific disciplines, they are central to the practices of field ecology.

Small science areas such as ecology are in the early stages of collaborating with computer

scientists and engineers to build research instruments. Traditionally, scientists in these fields – working alone or in small groups – have taken samples and sensor readings by hand, a process that is time- and labor-intensive. New technologies such as networked embedded sensors enable ecologists and environmental scientists to study the context of phenomena at much finer spatial and temporal scales than was previously possible (Arzberger et al., 2004a; b; Hamilton et al., 2007; Szewczyk et al., 2004). Although sensors do not replace the need for hand collection of biological samples, sensors can capture data on physical conditions such as ambient temperature, wind speed and direction, and chemical concentrations in water and soil.

Case studies of scientific collaboration reveal many kinds of data that may mean different things to different participants (Kanfer et al., 2000; Lawrence, 2006; Lee, Dourish & Mark, 2006; Ribes & Finholt, 2007). Ribes and Finholt (2007), for example, identify competing interests of environmental engineers and hydrologists, despite their common interests in water. Environmental engineers collect data such as pollution, contamination, sewage, and potability as indicators of water quality. Hydrologists gather data on drainage and erosion as indicators of water quantity. Tensions between the short and long term value of data are illustrated in a study of the Long Term Ecological Research (LTER) centers (Karasti et al., 2006). In the short term, participants focused on issues such as technology solutions, data volume, and metadata, whereas long-term concerns addressed scientific inquiry, data sharing, and stewardship.

2.3.2. Computer Science and Engineering Data Practices

Computer science research is characterized by a focus on system development and theory building over conducting empirical experiments (Basili & Zelkowitz, 2007). The most notable exception is computer science research on information retrieval, which can be highly experimental (Voorhees, 2007; Voorhees & Harman, 2005). Computer software, rather than experimental data, is the primary non-publication research product. De Souza, Froehlich, and Dourish (2005) show how in open-source development projects, software artifacts are "not merely the objects of software development processes, but are also the means by which those processes are enacted and regulated" (pg. 205). Thus, the structure and state of software can constrain or enable the ways that the development process proceeds, similar to the role that data play in experimental scientific research.

2.3.3. Science and Technology Interdependence

Scientists often rely on professional software engineers to construct tools for data collection and analysis. Here the collaboration challenges lie in the lack of clear software specifications for scientific instrumentation, in comparison to industrial projects (Easterbrook & Johns, 2009; Segal, 2005; 2009). When scientists collaborate with technology researchers, tensions often arise between the needs for research-grade and production-grade technologies. Scientists in the application domain need technologies to use in their own research, thus they have to define features and architecture sufficiently for the systems to yield the requisite research data. Conversely, computer scientists are engaged in their own research and desire as much flexibility as possible in pursuing their own questions (Lawrence, 2006). In contrast, when scientists write their own software, the resulting code is often more functional than elegant, as the optimization of code tends

to be secondary to its utility to the science and its portability to other settings (Carver et al., 2006).

Some big science areas of the physical sciences such as high-energy physics and astronomy, have a long history of scientists and technologists working closely together to design and build instruments, and to write software. Endeavors such as linear colliders and space telescopes take many years to design, build, and deploy. Requirements definitions are rigorous, yet must adapt to changes in technology and in the scientific questions to be asked (Latour, 1987; Traweek, 1992; 2004). These kinds of collaborations are more novel in the field-based life sciences.

In sum, new data collection technologies such as embedded networked sensors offer great research opportunities to small life science areas such as field ecology and marine biology. Taking advantage of these technologies requires collaborations with computer science and engineering researchers. Conversely, computer science and engineering researchers need partners in the domain sciences if they are to design, develop, and deploy their research in real world settings. These collaborations between scientists and technology developers take both groups out of their comfort zone: technologists must test new equipment in highly unpredictable field settings, scientists must rely on technologists to ensure that field excursions are successful, and everyone must be able to assess the trustworthiness of the data.

2.4. Center for Embedded Networked Sensing

The Center for Embedded Networked Sensing (CENS), the site of the research reported here, is a National Science Foundation Science and Technology Center established in 2002 and funded to 2012 by two five-year awards (http://research.cens.ucla.edu/). CENS supports multi-disciplinary collaborations among faculty, students, and staff of five partner universities (UCLA, USC, CalTech, UC-Merced, UC-Riverside). The Center's goals are to develop and implement innovative wireless sensor networks. Most projects involve researchers in sensing technologies and researchers in scientific application domains. About 80% of CENS collaborators (which now number about 300 faculty, students, staff, and post-doctoral researchers) are in computer science or engineering. This group cuts a wide swath across environmental, electrical, and structural engineering, and computer science areas such as robotics, systems theory, networks, and actuators. Their research focuses on dynamic deployments of networked sensing technologies, which allow for more flexible capture of local phenomena than observatory networks do. The remaining 20% of CENS collaborators are in application domains such as environmental sciences, biology, seismology, and geology; a few are in the arts, architecture, or medicine, principally concerned with cell phone-based mobile applications. These participants' research focuses on science or other domain problems that can be investigated using sensor networks.

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, and whether data were being captured at all outweighed considerations of data quality and usefulness. Once the basic technical problems were resolved, the CENS research program became more science-driven, while continuing to explore core computer science and engineering problems in wireless sensing networks. The initial framework for CENS was based on static networks. Early scientific results revealed the difficulty of specifying field requirements in advance well enough to operate systems remotely. Most CENS' research since 2005 involves dynamic deployments where investigators adjust monitoring conditions in real time (Batalin et al., 2004; Chen et al., 2003; Estrin, Michener & Bonito, 2003; Hamilton et al., 2007; Mayernik et al., in review; Pon et al., 2005; Rahimi, Kaiser, Sukhatme & Estrin, 2005).

The period from 2005 to 2009 was the peak of scientific deployments in CENS. By late 2009, funding for NIMS, the largest technology platform in CENS, was ending (NIMS: Networked Infomechanical Systems, 2006; Batalin et al., 2004; Harmon et al., 2007; Pon et al., 2005; Rahimi et al., 2005; Sutton, 2003). Researchers were transitioning to other projects, inside and outside of CENS, and many turned to non-science applications, such as participatory sensing using mobile telephone platforms (Mun et al., 2009). Figure 1 depicts many of the sensor technologies developed or deployed by CENS.

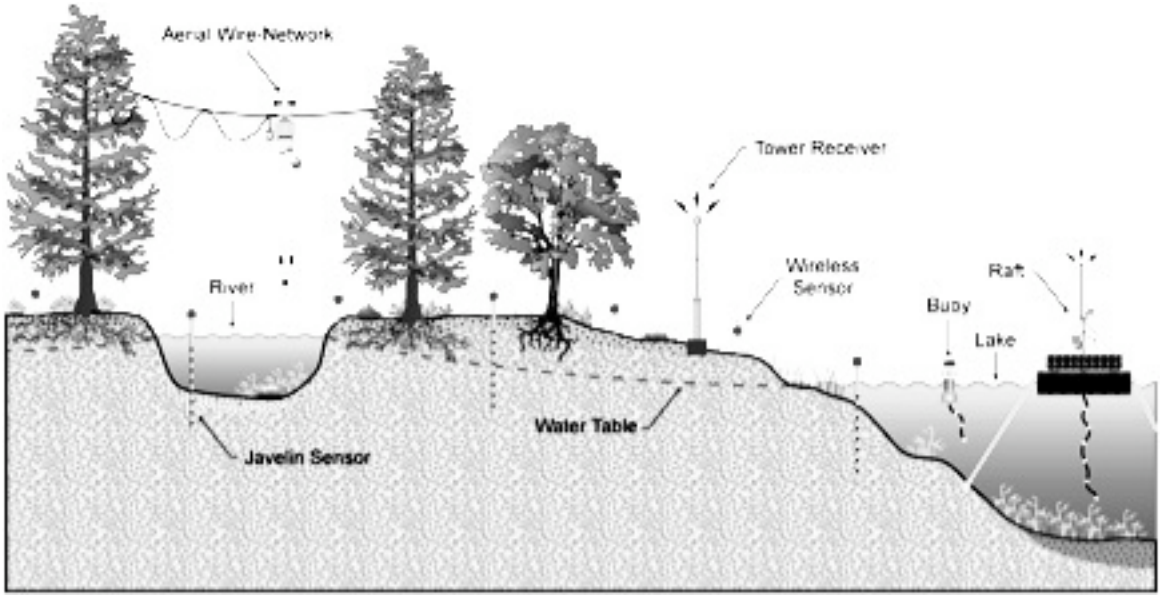

Figure 1: Heterogeneous sensor deployment. Graphic by Jason Fisher

As a founding co-investigator of CENS, the lead author of this paper has conducted data practices research in the Center since its inception. Despite the minimal amount of research data produced in the first three years of CENS' existence, our presence as participant observers enabled us to follow the trajectory of the science and technology research from its early days.

Our early research at the Center revealed the division of labor between the science researchers and the technology researchers. The research of technical teams addresses the design, development, and deployment of sensor technology and networks. The research of science teams addresses patterns of physical and biological phenomena. Technology researchers are interested in science questions only to the extent necessary to design for "real world" problems. Scientists are interested in technology issues only to the extent

necessary to calibrate, trust, and interpret the resulting sensor data (Borgman, Wallis, Enyedy & Mayernik, 2006). In other studies, we found that data on the same variables are gathered by multiple means, data exist in many states and in many places, and that publication practices often drive data collection practices. Scientists, computer scientists, and engineering researchers may make use of common datasets but interpret the data differently (Borgman, Wallis & Enyedy, 2007). Metadata creation is a distributed process within collaborations, as different members have different pieces of the overall picture. Frictions can arise at several stress points: (a) conflicting standards, (b) temporal rhythms, (c) data sharing practices, and (d) the availability of support (Edwards, Mayernik, Batcheller, Bowker & Borgman, 2011; Mayernik, 2011; Mayernik, Batcheller & Borgman, 2011). Pepe, in conducting a multi-year social network analysis of CENS, found that as the Center progressed, CENS members were more likely to publish papers with CENS members in academic departments other than their own, reflecting an increase in the interdisciplinarity of their collaborations (Pepe, 2010; Pepe & Rodriguez, 2010).

In our effort to track the life cycle of CENS data from experiment design to publication, we participated in technology-only deployments, science-only deployments, and joint science and technology deployments (Wallis, Borgman, Mayernik & Pepe, 2008; Wallis, Pepe, Mayernik & Borgman, 2008). Our prior research in CENS highlights the role of collaboration in research processes and phases during which the integrity of the captured data could be compromised. The present article explores these collaborative behaviors in depth.

3. METHODS

This article is based on field observations of five CENS environmental science projects conducted between 2006 and 2009, interviews conducted in 2006 with participants in these projects, and experiences from being part of the community throughout the life of the Center, from the proposal writing stage to the present. Interpretations are based on these data and on later discussions with the project participants, both formal and informal.

The period of 2005-2009 was ideal for capturing the perspectives of scientists and technology researchers about their data, their research, and their collaborations. Science and technology field deployments were at their peak; interviews took place in the fourth year (2006) of CENS' decade-long existence (2002-2012) as an NSF Science and Technology Center. CENS activities were maturing – the technology had begun to stabilize, research activities were less constrained by the battery life in the sensors, the networks were beginning to produce data of scientific and technological research value, and collaborators had learned more about the capabilities, strengths, and weaknesses of their partners.

We briefly described the initial interview results in a conference paper (Borgman, Wallis, Mayernik & Pepe, 2007). In this article we present our full analysis of the interviews and field deployments of the five CENS projects included in our study, focusing on science

and technology data practices and their interdependence. Results reported here also serve as a baseline for later rounds of interviews, document analysis, and field research that are currently in progress.

3.1. Interviews

At the time of these intensive interviews (2006), CENS was 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. For each of five CENS environmental sensing projects, we interviewed a complementary set of science and technology participants, including faculty, post-doctoral fellows, graduate students, and research staff. A pilot ethnographic study in 2005 consisted of in-depth interviews with two participants, each two to three hours over two to three sessions (Borgman, Wallis & Enyedy, 2006). From these pilot interviews, we identified the most compelling topics in how researchers talked about their data and the problems they encountered in managing their data. The questions used in the pilot interviews were modified to cover these areas at greater depth. The areas and sample questions are described below.

In the full study, we interviewed 22 participants involved in these five projects. Sessions lasted 45 minutes to two hours in length, averaging about 60 minutes, and were audio-recorded. Most interviews took place in the subjects' offices or laboratories. Transcripts of the audio-recordings were returned to subjects for review, comment, correction, or deletion. A few subjects made corrections; none asked to remove any comments from the record.

As shown in Table 1, the 22 participants consist of 11 faculty, doctoral students, and research staff in the science application domains of CENS and 11 in the technology areas. Given the interdisciplinary nature of CENS, some judgments had to be made about drawing the sample to pair science and technology collaborators on projects. Those in fields such as biology, oceanography, or public health are easily classified as science participants. Similarly, those in computer science and electrical engineering are easily classified as technology researchers. Participants in environmental engineering and statistics (4 of the 22) were classified based on their role in the project rather than by their academic affiliation. Two of these four participants are classified as science partners, as their roles were more concerned with scientific applications, and two were classified as technology, as they were more focused on the design and use of the technology. In reporting our results, we draw particular attention to the activities of these boundary spanners in the deployments.

Interviews		Pilot	Terrestrial	Contam	Aquatic	**Total**
Scientists	Faculty		3	1	2	**6**
	Staff	1	1		1	**3**
	Students	1		1	1	**3**
Technology researchers	Faculty		3	1	1	**5**
	Staff		3		1	**4**
	Students		2	1		**3**
Total		**2**	**12**	**4**	**6**	**24**

Table 1: Interview participants and their distribution

Our interview questions clustered 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?

3.2. Field Deployments

Two members of our research team participated in 12 CENS sensor deployments for environmental science projects, both observing and taking part in deployment activities, encompassing approximately 22 days of participant observation over three years (2006-2009). The number of CENS researchers participating in each of these deployments ranged from two to ten. The length of our participant observations in the environmental science deployments ranged from single-day excursions to a four-day stay with CENS researchers at a remote field site. Observers took ethnographic field notes and digital photographs about the nature of deployments, field-based scientific research practices, and the role of information systems in these instrumented field-based research projects. Research teams put us to work as lab assistants, recorders, haulers, or anything else within our perceived skill sets. We participated in equipment installation tasks, data collection, and numerous other field activities. Our participant observations were supplemented by informal interviews and discussions with CENS researchers before, during, and after deployments regarding their data collection and collaboration practices, using the interview questions listed above or similar versions thereof. Our willingness to participate in team activities made us welcome partners, and we often are invited to join deployment excursions.

3.3. Data Analysis

The interviews were audio-recorded, transcribed, and complemented by the interviewers' memos on noteworthy phenomena. Transcription of interviews totaled 312 pages. Field notes varied in depth and format by observer; several dozen pages of notes also were including in the analysis. Photographs taken at field sites were used as memory tools and aids in description, but were not coded, per se. We developed a full coding process using NVIVO, which was used to test and refine themes in coding of subsequent interviews and field notes. Initial codes were based on our research and interview questions; other codes were developed iteratively by examining the data to identify themes of interest. With each coding scheme refinement, the remaining corpus was searched for confirming or contradictory evidence.

We submit sections of our papers that describe research deployments or other activities to the participants to review for accuracy. If we send them full papers, we mark the sections where we are particularly concerned about accuracy, which increases the likelihood they will respond quickly. Our science and technology colleagues in CENS have been extremely helpful in clarifying our understanding of their work. Most corrections are to our errors in describing instruments, findings, or scientific concepts.

4. RESULTS

The results are divided into several sections. First, we briefly present a scenario of a CENS field deployment in which both scientists and technology researchers (computer science and engineering) participated. Second and third, we present science and technology perspectives on data as seen in the interviews and deployments. Fourth, we draw out the interdependencies between the science and technology partners for data, context, and interpretation. The discussion section develops the comparisons more fully.

Interview quotations from members of science teams are prefixed with an S (e.g., [S16]) and those from members of computer science or engineering teams are prefixed with a T (e.g., [T5]). Comments from field deployments are paraphrased, as these were not formal interviews for which we have transcriptions. People, projects, and equipment to which these individuals refer are anonymized with bracketed descriptions, e.g., [Prof. X] or [System Z].

4.1. Science and Technology Joint Deployment Scenario

While science and technology teams each conduct independent field studies, the joint deployments provide the richest sources for studying the interaction between groups. Each team has to be more articulate about its goals so they can explain them to their partners who have different academic training and research methods. Conflicts in methods and interpretations can arise, which also offer insights to data practices.

This scenario is drawn from a series of similar field deployments to study harmful algal blooms (HAB) that occurred over the course of the summer of 2006. A harmful algal bloom is a situation in which a particular algae suddenly becomes dominant in the water. HAB create toxic conditions that kill fish and other animals such as sea lions by consuming the available dissolved oxygen that fish need, or by releasing domoic acid, a harmful neurotoxin that affects large mammals. HAB can occur in fresh water and in oceans. The environmental conditions under which HAB occur are poorly understood, partly because they are difficult to predict and partly because few means exist to capture HAB in sufficient density to understand the underlying processes. While rare, they can cause severe damage, potentially killing tens of thousands of fish in a day. The deployment scenario described here takes place at a lake known for summer blooms that kill off all the fish stocked in the lake.

From the marine biologists' perspective, HAB are a good application for sensor networks as they can collect a much larger number of observations, and study more variables than is possible with hand-sampling techniques. Scientists can adjust their data collection to local conditions by moving sensors. From the computer science and engineering perspective, HAB are a good application for sensor networks because they test the ability of physical and biological sensors to collect large numbers of variables. HAB are of particular interest in robotics, for the rapidly changing science data can be used to actuate or trigger sensing systems on robotic boats, buoys, helicopters, cameras, and other autonomous vehicles that can follow the phenomena in greater detail than static sensor networks.

The science team has conducted research on this lake for several years. They have access to baseline data collected by the Department of Fish and Game in addition to their own data from earlier visits. To design their field study, the team assembles background information about this lake such as peak months for algae, a topology of the lakebed, phytoplankton and zooplankton species they are likely to see, and nutrient presence and concentration. Prior to going into the field, they calibrated their sensors in their laboratory using known solutions.

The technology team, in contrast, had no prior experience at this lake. Their research site is wherever the scientists are. They planned to field test their equipment based on the science team's requirements as a means to conduct their own research on algorithms for robotic guidance, for network health, for sensor fault detection, and for the design of

sensor technology interfaces. The computer science and engineering researchers relied on discussions with the science team to guide their choices of equipment, specific sensors, and the time, place, and length of deployment of each. For this HAB deployment, the technology researchers' main goal was to test a new vision algorithm for navigating the robotic boat. The team prepared and tested their equipment in their campus laboratories or local water bodies, at a university fountain and a faculty member's pool, prior to the deployment.

Participation in this four-day deployment differed from day to day. Most of the students and research staff arrived the first day to set up equipment; faculty investigators arrived on the second day; others came and went, staying for hours or days. The scene was often chaotic. Of the 20 or so people involved, eight to ten were associated with the electrical engineering team that built the sensing system, six to eight with the marine biology team, four or five were part of the robotics team from computer science and engineering, and two were from statistics. Participants came from at least three universities. The numbers are approximate due to the overlapping roles and responsibilities of many individuals.

Sensing equipment was deployed as soon as everyone arrived, using a tethered buoy network along this narrow lake, a robotic boat that could respond to data collected by the buoys, and a fixed robotic system to transect across the lake. The team documented GPS coordinates (latitude and longitude) of the sensors, times of placement, and serial numbers of each sensor as it was dropped at its sensing location. In addition to the sensing equipment, the marine biologists brought equipment to set up a wet lab for

processing samples on-site. Once the equipment was collecting data, minor changes were made to sensor placements to capture more "interesting" phenomena, which typically means that more sensor density is required in areas with greater change of some physical variable, such as light, temperature, or water flow. Data were collected by sensors and by hand-collecting samples of water at different depths. Corroboration in the field of these data and sensor data led to more changes in the topography of the sensor network. When data collection was finished, everyone pulled their equipment from the lake and from the lodge where the wet lab was established, packed up, and headed to their respective institutions.

After the deployment, the scientists analyzed the water samples for nutrient concentrations and for organism identification and concentrations. The technology researchers adjusted the sensor data to reflect calibration and cleaned them to remove outliers, equipment artifacts, and other identifiable errors. The sensor-collected data elements of interest to the science team were provided to those teams. Once the science team received the sensor data from the technology team, those sensor data were compared to the in-lab and in-field calibration curves and to other trusted data sources. Water sample data and sensor data were integrated for analysis. The technology researchers used the sensor data for simulations and to generate algorithms for automatic sampling strategies of interest to the marine biologists. The output of these simulations and results of subsequent testing became the basis for technical research publications. After data analyses were complete and papers are published, the scientific data are burned

to DVDs and shelved with other data. Observational data used by the technology researchers were maintained on servers, but simulated data usually were discarded.

As is evident from the harmful algal blooms scenario, science and technology researchers work side by side in planning and conducting environmental field research, but go into the field with different research questions and methods. The teams' data needs and uses are both complementary and competing. Each team is dependent on the other for some aspects of the research. We report on science perspectives on the data from these deployments, then technology perspectives, and then interdependencies between them.

4.2. Science Perspectives on Data

The scientists are seeking patterns in their data. Their research questions usually address correlations between phenomena and trends in the environment. In the case of the harmful algal blooms scenario, for example, the scientific goal is to predict, and ultimately to prevent, such blooms. We participated in two lake deployments that fit the scenario described above, where we observed the scientists placing a network of static sensors and taking hand samples of the water.

4.2.1. Sensor Data for Scientific Variables

Among the many scientific phenomena that can be observed with these embedded sensor networks are chemical and physical variables such as soil moisture, wind direction, sap

flow, bird calls, temperature, sounds, and images, and the presence of substances such as dissolved oxygen, ammonia, or indicators for the presence of mercury.

The ability to geo-locate and time-stamp observations is among the scientific advantages of sensor networks and is now trivial from a technical standpoint. Geographic coordinates may be insufficient for scientific purposes, however. Satellite ground positioning systems (GPS) provide latitude and longitude coordinates, i.e., two-dimensional positions. These scientists may need four or more dimensions to interpret data from the sensors. Altitude above or below sea level and altitude relative to the ground or water level are also needed. For example, it matters scientifically whether the sensor is collecting observations on the ground, 1m above the ground, or 10m above the ground – or 1m or 10m below the water surface. CENS scientists who compare patterns of plant growth at different altitudes and light levels need to know if the sensor is above or below a leaf or other object of interest, or obstructed by a rock or a tree. Capturing these additional dimensions is often beyond the technical capabilities of the sensor network, and must be recorded by hand.

Some of the sensor data "is generic, so it's pretty much applicable to anybody's study: plants, animals, climate, soils, insects –you name it" [S12]. Generic data, such as weather and environmental variables, usually provide context for experimental variables. Scientists "don't necessarily study their weather at that [Site A], but maybe they're studying some animal and they could look at how the weather pattern has changed and see there might be a correlation with this animal's behavior. Having that environmental

data is just another sort of useful piece of information that a researcher could need. Other data I hope to eventually provide our researchers … is geographic data, like GIS, different layers of the [Site A], vegetation layers, and elevation stuff" [S17]. These "generic data" typically come from static or autonomous networks. For these CENS researchers, these data usually serve as baselines for comparing experimental variables. In other types of research, such as climate monitoring, these data may be of primary interest.

4.2.2. Hand-Sampling in the Field

Hand-sampling is the scientific activity that has no counterpart in technology research. Many of the CENS science teams collect physical samples of water, soil, or plants to accompany the sensing data. One team does "old traditional … methods of stream sampling. So I go out to site, take my water sample, do my little algal counts, take my algae samples and bring it all back to the lab and analyze it" [S1]. Another team that gathers hand samples analyses them to "get data that relate to what kind of classes of algae and cyanobacteria are present, in what sorts of abundances and things like that." [S10].

To take water samples, science researchers must judge the density and type of phenomena to determine which screens and dilutions to use. The principal investigator, an experienced marine biologist, is able to estimate parameters such as chlorophyll concentration by visual examination of the lake. He said he thought the lake was "about a 10," meaning 10 micrograms per liter, which turned out to be fairly accurate. When the

weather warms up the biomass concentration will increase by another order of magnitude, and HAB growth is off the charts. Chlorophyll concentration provides a rough estimate of plankton biomass.

We observed the marine biology team perform a multi-step process to determine what they were finding in the way of organisms and concentrations and then to act on those results to determine their next round of sampling. They would take three samples at each site, observe the activity in one of those samples, and kill the organisms in the other two samples using formaldehyde and liquid nitrogen to preserve them for more detailed examination later in the campus lab. Their goal is to assess their samples at different levels of concentration. Based on the chlorophyll estimate, they used 200-micrometer pore screens to separate out the zooplankton and 20-micrometer pore screens to separate out the phytoplankton, which are an order of magnitude smaller. They only brought enough bottles to collect six samples at each of the three sites in this lake. Thus they needed to process the samples right away – filtering, petri dishing, freezing, etc, so that the bottles could be reused. They need multiple bottles for each sample because so much liquid is used for each process and to add redundancy for contamination.

A graduate student participant proceeded to kill and preserve samples while the PI examined the third sample of zooplankton under the microscope. The PI and two of his graduate students all viewed the samples, then speculated about the types of limnological samples they had, pointing out various characteristics of the two predominant species. The 200-micrometer pore sample was packed with daphnia or daphnids, a very common

zooplankton that even we, the social scientists, were able to identify; they were very active and the phytoplankton in their digestive tracts were visible as a green line. In this sample, the biologists also found a much rarer type of zooplankton, a volvox with offspring visible within the super-structure that would need to be more closely identified in the lab. The 20-micrometer pore sample was given the same treatment by the PI and his graduate students. This sample had some green balls and some clear objects that looked to us like snowflakes but were clearly more meaningful to the marine biologists. These phytoplankton are apparently unusual, making them difficult to identify in the field, but the biologists narrowed the options to a couple of candidate families. The samples are marked with a collection time and location for future identification. Based on what was known about the zooplankton and phytoplankton families identified, the biologists decided that in the next steps, they would do a surface sample first and then profiles of samples at multiple depths for each location that samples are collected.

4.2.3. Sensor placement

Sensor placements are determined by research questions, technical capabilities of the sensor network, and local conditions such as variations in the depth of the lake. Sensors may descend no deeper than a half meter above the lakebed or else they will churn it up, clouding the instruments, which limits how much of the lake can be measured. During the day, the sensors placed at a half meter and one meter depth below the lake surface were of the most interest because the light in the water is strongest at the former and barely reaches the latter, making these the critical depths for tracking organism activity. At night, deeper sensors become more important as the organisms migrate down through the

water column.

Based on the results of hand samples, the science team adjusted their sensing to take three samples spaced by eight hours, instead of four samples by six hours, because they now think that will be sufficient for their research questions. Because no algae were presently blooming, they could not observe algae movement in the lake. Measures from this deployment would serve as a baseline for later data collection.

While the sensors and hand samples are measuring some of the same variables, the duplication of effort is necessary to minimize risk of sample contamination and to serve as a "ground truthing" mechanism. "Ground truthing" refers to the use of known measurement methods to test the validity of new measurement methods. In this case, contamination data collected via established hand-sampling techniques provided a "ground truth" that could be used to validate the contamination data that came from the sensors.

4.2.4. Derivation and Analysis of Data

Chemical and physical sensors yield measurements such as voltage; they do not measure the phenomena of interest directly. These voltages need to be converted into data by integrating the calibration coefficients captured before data collection, or, in some cases, the voltages are used as indicators and need to be interpreted as data through models of scientific phenomena and relationships. Our scientists offered a number of examples of how measurements are derived from sensor data through filters and calculations: "We're interested in something which an instrument isn't going to tell us, (such as) we've

reached a certain level … Dew point temperature is the temperature at which air gets saturated and begins to condense. … You've got to calculate it from other measurements. So there are a lot of things like that that you want the machine to tell you when something interesting happens, and you've got to give it a model of how you're going to calculate what that is, so what kind of output do we want" [S3]. For the [Site B] arsenic project, "there's no arsenic sensor, so we were looking at, let's see, iron, ammonium, chloride, nitrate, just geochemical parameters, pH, oxidation reduction potential. So it's just things to help us understand the chemistry" [S11].

Physical samples of water and soil also require processing and manipulation to yield useful data. Water samples are diluted to expected detection levels, for example. Soil samples must be centrifuged. Some hand samples are analyzed in the field with portable wet labs. Other samples are brought back to the campus laboratory for testing that requires specialized equipment or extensive processing. Some physical samples must be cultured for 24 hours or longer to yield useful data. DNA testing of samples for the presence or absence of living organisms requires expensive equipment and supplies. In another example, environmental scientists described to us how the EPA-defined protocol for measuring mercury in soil samples would take from 9 months to a year to complete.

4.2.5. Interpretation and Trust in Data

Scientists' ability to interpret their data depends heavily on their trust in their instruments and in the accuracy of the data that the instruments yield. As noted in the above scenario, they calibrate their equipment in the laboratory and again at the field site. Most of the

science researchers (faculty, post-doctoral fellows, students) we interviewed reported that they needed to know as much about the instruments as possible to be able to interpret their data. An ecologist provided the richest quote on this point:

> Oh, I think you need to know everything that you can about the instrument. Yeah, so there's hundreds of different ways of measuring temperature. If you just say, "The temperature is," then that's really low-value compared to, "The temperature of the surface measured by the infrared thermopile, model number XYZ, is…" That means that I know that it's measuring a proxy for a temperature. Rather than being in contact with a probe and it's measuring it from a distance, I know that its accuracy is plus or minus .05 of a degree based on the instrument itself. I want to know that it was taken outside versus inside in a controlled environment. I'd like to know how long it had been in place and the last time since it's been calibrated. That might tell me whether it's drifted. You know, these are all of these pieces of the metadata that will hopefully automatically get associated with the fact that it's a temperature point. [S12]

Reconciling timestamps of individual sensors is a continuing problem in research that relies on sensor networks. In the early days of CENS, sensors often rebooted themselves after an electrical fault or battery decay, which would restart their clock. The inability to synchronize timestamps on data from individual sensors wreaked havoc on scientific data collection. Most of the battery and rebooting problems had been resolved by the time of these interviews, but the scientists remained sensitive to the issues of synchronization in

interpreting their data. This sensitivity was noted by a faculty scientist who described how his student had reliability issues related to timestamps from sensor data:

> Well "data" are the instrument feeds, … A lot of what [my student] is trying to do now is to get good reliable data out of these [instruments]. …The problem is it's really been difficult to keep those stations running reliably….[My student] wanted the time all the same on each one, and the real time, and not just set to zero randomly. We've just got a lot of issues of data matching. [S3]

4.3. Technology perspectives on data

The research questions that computer science and engineering teams study in the field address the design, health, and efficacy of the sensing technologies. "In robotics people are most interested in the performance characteristics of the robot. So in our case we report in robotics journal papers that we can control the robot – meaning the robot based on its sensing can control itself to accurately go from one location to the other" [T18]. Computer science research on networking is "looking at the network itself. So instead of looking at 'is this sensor bad,' it would say 'is this node not communicating very well,' 'are the batteries low,' 'does it have a tree that's blocking its communication' kind of thing" [T6]. These types of research are highly iterative. Teams in the field are constantly debugging, whether to make robots go where they want them to go, to identify faults in networks in real time so they can be corrected, or to compensate for sensor limitations by collecting more hand-samples.

A continuing theme in the deployments, in the weekly CENS seminars, and in other venues such as NSF site visits and annual research reviews, was the amount of science that the technology teams had to learn to do their own research. While the need to learn sufficient domain knowledge to design instruments – whether in biology, chemistry, ecology, or public health – seemed to come as a surprise to many of them, they also appeared to welcome the challenge. Most were proud of how much science they had learned by working in CENS. The science teams learned more technology, though many were technophiles, eager to tweak technology in the field, or to construct their own devices when necessary. On balance, the science teams appear to have brought more technical knowledge to these collaborations than the technology teams brought science knowledge.

4.3.1. Sensor Data for Technology Variables

Computer science and engineering researchers' initial concern was getting any data of scientific relevance at all: "at that point we weren't looking at the quality of the data. We were just looking at the quantity of the data" [T6]. Another subject simply said, "a temperature sensor is a temperature sensor" [T5].

Four types of data were identified in the interviews with technology researchers and in our participation in field deployments: First are observations of physical and chemical phenomena, sounds, and images of scientific phenomena; these data are provided to the scientific partners. Second are observations of natural phenomena used to actuate or to

guide the robotic sensors to a place in the environment. The sensor system could move autonomously toward brighter sunlit patches of earth or toward areas of water with higher concentrations of algae, for example. These first two categories of data are the same observations but are used for different purposes by the science and technology research teams. The third category is performance data by and about the sensors, such as the time that sensors are awake or asleep, the faults they detect, battery voltage, and network routing tables [T6]. A fourth category of data identified in the interviews and in field studies is proprioceptive data collected by the sensors – data to guide robotic devices such as motor speed, heading, roll/pitch/yaw, and rudder angle [T18].

These four types of data played different roles in the technology research. Scientific field data are of interest to the technology researchers as a basis for their guidance or network algorithms. In robotics,

> one way you convince people that that's the case is you take your 10 runs, here's the commanded location, here's where the robot actually ended up, and look, it's close, right? And in a robotics experiment that would be called data. That's purely from a robotics point of view. And in fact, the robot actually needn't be doing anything. It needn't be sampling the water at all because most of the sampling is based on being able to go someplace accurately, so you'd abstract away the task the robot is actually doing, and only really focus on how well it can navigate to get someplace. [T18]

An engineering staff member said that "Every piece of information that comes back is data, because I care about both the environmental data, which is what we're collecting, but I also care about the system data, if you will, the sensor health data, those kinds of things, network status" [T4]. A computer science researcher studying the sensor networks distinguished between two types of data: (1) data from the sensors, and (2) system metrics. [T6].

Another robotics and engineering researcher was able to distinguish among three categories of data associated with his system:

> One would be data that our instruments acquire that is to be supplied to the user, that the [System A] will not observe, that it has archived or transmitted. But the [System A] will not observe, it's not of interest to the [System A], other than it's of interest to collect. Another class of data is collected by the [System A] and it also pertains to an environmental measurement, and it might be used to guide [System A], that is, to collect more samples or to reveal a sensing task that then must be executed. Like for example, [team members A, B, C] are pursuing our systems where an imager examines a forest floor and finds patches that are bright, and are regions where [System A] should be sent to collect high resolution data. …. There's this third class of data which is really about the [System A] itself, which might be data associated with fault detection, with degradation, overall performance, energy usage and so on. Some of that

> itself is acted on, and other data is archived and simply used to drive future design advances. [T15]

As this last quote illustrates, the data types are used by the technology researchers in combination to evaluate system functionalities, facilitate additional data collection, and isolate problems. The validity of the sensor data, while useful in evaluating some sensor system functionalities, is irrelevant to questions about the effectiveness of the communication systems or the accuracy of a robot's mobility.

4.3.2. Sensor Placement

Compared to the scientists, technology researchers were flexible about sensor placement. Technology researchers could gather their data if sensors were deployable in ways that the scientists might use them. Given the high failure rate of sensors in the field, scientists were very cautious in trusting the technology. Carrying sensor platforms on a long car ride often rendered them useless, for example. Technology researchers needed to take the equipment out of their labs to experience the pitfalls faced by their science partners. Technology-only deployments took equipment to the locations of scientific research and ran them in ways that would mimic deployments. These activities allowed the technology teams to perform more rapid hardware and software iteration cycles, making changes on the fly, because they did not need to support scientific practice concurrently.

4.3.3. Derivation and Analysis of Data

As noted above, the technology teams often "abstract away" the science data to focus on their technical findings. They are concerned, for example, with whether they can guide a

robot to the correct location for sampling, or whether they can model the fault patterns in a network. Issues of calibration, time-stamping, and error detection are paramount to these teams.

The role of software "code" for the technologists is roughly parallel to the role of data for the science teams. The centrality of software in the data collection and interpretation process was more apparent in the deployments than in our interviews. We observed several deployments of a fixed robotic system for use in forest canopies, rivers, and lakes. To interpret the data from these experiments, it is necessary to know which of the many versions of code and hardware were used to generate them. Questions about versions and states of CENS' data are closely linked with questions about versions and states of software. The developers use version control systems to maintain records of the states of their code.

A river deployment of the fixed robotic system offered insights to how parameters are adjusted on site in response to field conditions. Most of the sampling was determined on site; the science team arrived with a general idea of their sampling plan (including desired sampling locations and methods), but knew from past experience that these plans would have to be adjusted based on the state of the field site. The technical team arrived without a sampling plan of their own beyond "piggy-backing" on the scientists' activities. A science graduate student conducted the first three sampling runs. In the first run, they set the horizontal and vertical intervals for the sensors, which determined the grid size in the water that would be observed, and set the dwell time (the length of time the sensors

captured data at a given location) to 60 seconds. In the second run, the intervals were held constant but the dwell time was halved. The third run had the same observation parameters, but started at a different point below the water, thus increasing sampling density overall. The principal investigator for the project then examined the student's three transect profiles. For the fourth and final run, he changed several parameters. The horizontal and vertical intervals were both changed, the dwell time was much longer, and a sonar sensor also was attached to the system.

4.3.4. Interpretation and Trust in Data

The technology researchers are very conscious of the fact that data from scientific sensors are a means and not an end for their own research. Thus the sensors "are the payload the robot is carrying. And so if you think like a roboticist, sure, the robot is being built to do something, but your focus is on getting the robot to accurately do motion or avoid obstacles" [T18]. Roboticists abstract away the "real world" task. Like real estate agents, their concern is location, location, location. As another researcher put it, "the [System A] isn't really concerned with any scientific objective" [T15].

In several deployments, we noted how technology researchers use scientific data to adjust their systems in real time. In most cases this is a manual process of assessing targeting results and making decisions, as described above with the varying parameters on the robotic system. In the longer term, a research goal is to create algorithms and systems that will create sampling paths adaptively based on one or two initial runs, a process known as adaptive sampling.

The scientists were in charge of the science and the technology researchers were in charge of the technology. The technology researchers had a much better idea of what data could be trusted because they had access to information about sensor and network health that influence the interpretation of sensor readings. The technology researchers were interested in the presence or absence of data, while the scientists were interested in the data as evidence of phenomena.

4.4. Science and technology interdependencies

Observations of physical, chemical, and biological phenomena are of interest to all of the researchers, but these data serve different research purposes for each team. Robotics and network researchers may be able to use the real-time data in the field, but most of the data require cleaning to be of full value, especially to the science teams. The technology teams devote considerable effort to reconciling timestamps from multiple sensors before passing those data to the science teams. For scientists to assess trends or to compare phenomena between places or over time, they need accurate records of exactly when and where an observation was taken. These interdependencies fall into several categories, which we illustrate with examples from our interviews and field deployments.

4.4.1. Data Collection Interdependencies

Science and technology teams go into the field together to gather scientific data and to test and evaluate hardware and software. The science teams also collect water samples to

calibrate, verify, and validate the sensor data. Sensor data can be gathered at much higher spatial and temporal granularity than can physical samples, hence the scientific teams gather far more sensor data points than water samples.

In the four-day lake deployment described in the scenario, one of the primary robotic instruments failed to function. The science team had hoped to corroborate the data collected at the buoy right next to the robotic system node. As a result of losing the robotic system, the joint teams were unable to accomplish their three goals for the week: 1) corroborate biological and buoy data, 2) test the system interfaces with the biologists, and 3) test a new algorithm to give the robotic boat location points that it could use to fill in data collection gaps between buoys.

While the failed technology undermined the planned goals of the deployment, another exchange between the teams improved both short- and long-term data collection. The technology team was dropping the sensor kit to the lowest depth in the lake and then making sampling stops on the way up to the surface. The biologists pointed out that they would get cleaner data, literally, if the sampled on the way *down* rather than *up*. By starting at the bottom, they were churning the lakebed, which changed the characteristics of the sample. This lesson was applied in other CENS water deployments thereafter.

4.4.2. Interdependent Knowledge of Instrumentation

In our interviews and deployment observations, science and technology researchers said they needed to know as much about the technology as possible to interpret the resulting

data. One scientist, in particular, felt that he needed to know much more about the instruments than he was being told by his technology partners:

> To be frank, for a lot of our engineers, if it works, that's it. … And to me, 'No, it doesn't work. You haven't told me anything about that data point.' How do I know that that difference between that data point and that data point is not due to an engineering issue, rather than due to a microclimate variation? Why are you confident that that measurement and that measurement are the same? It's like, 'Well, it's still working.' That's not the answer I want to hear. That's the answer I get. [S12]

Similarly, technologists rely on scientists to help them interpret the output from technical systems. In the following exchange, a technologist explains how calibration information produced by scientists is critical to understanding instrument output:

Q: What do you need to know about the instrument to interpret the data?

T5: I just need calibration data, like in chlorophyll, what quantity of chlorophyll produces what voltage output per se, or temperature or wind speed or whatever.

Q: Do you find that that varies over time, those calibration equations?

T5: It depends on the instrument. Some instruments are more stable than others.

Q: How do you get the calibration data?

T5: The biology technician gets an algae sample which she characterizes in the lab and dilutes to different dilutions and we do direct readings.

4.4.3. Data Cleaning and Handoff

Until data are verified and corrected, they are difficult to interpret. Research-grade equipment is often unreliable, wreaking havoc on data of interest to both scientists and technology researchers. Sensors failed in predictable and unpredictable ways, whether suddenly, gradually, or erratically. Clocks drifted, and sensors sometimes would reboot themselves, resetting the clock each time. While the worst of these problems had been resolved by the time of these interviews and field deployments, measurement reliability remained a concern and a research issue. The accuracy of space and time measurements depends on GPS technology, on the ability to triangulate signals between sensors, and on clocks in sensors.

The technology teams would clean data by adjusting measurements to observed or predicted calibration curves and by reconciling differences between instruments. As a result, scientists might have to wait some days for their data, and trust the skills and judgment of the people between them and their data. One scientist was concerned about both of these matters:

> Even when my student, who is the one I really interact with, when I've asked him for something I know that then he has to wait until [Prof. B's student] has time to do it. What would be great would be if he could just… when the data came in, he could just use it. As it is now, he can't just use it. He has to wait. And [Prof. B's

> student] is like the gatekeeper. He's the only person who knows how to do that right now and can do that [S20].

The graduate student of this scientist expressed a similar concern about the bottleneck that occurs, but also added a concern about the quality of the work:

> [Prof. B's student] usually takes it from there and futzes with it in MATLAB, because really synching all of the sensors is a chore. And that's to put it lightly. And he does an excellent job with it, but it's still a concern of mine, actually. Because I've received data sets that I'm sure are not synched properly [S1].

In addition to synchronization problems was the concern as to whether outliers were sensor artifacts or the result of actual phenomena. By increasing spatio-temporal resolution, the scientists were navigating uncharted territory. The models the scientists used to interpret their data were calibrated for lower-resolution data collection, but what was actually happening may not be so smooth. Disagreements arose over whether individual data points were "real" or not. Similarly, the determination of what were the "interesting" features of data involved an iterative exchange between the scientists – who understand the phenomena being studied – and the technologists, who understood the instrument and statistical methods used to derive data:

> Q: This idea of having the static nodes kind of identify something that's potentially interesting for the actuated sensor robot, this idea of "interesting" seems to be kind of a thorny issue in and of itself.

T18: Yes it's a big issue.

Q: Is that something that you use the biologists for, to tell you what could potentially be interesting?

T18: A lot, absolutely.

Q: Okay. But also using statistical methods…

T18: Exactly. So one way to do it is to say, well you look for patterns, and the other way is to do a combination of providing statistical tools and then getting some insight from people who are interested in measurements, biologists. And often in some cases, they're able to articulate exactly what is interesting, and in some cases they're sort of in the mode where they say, well what can your network discover, and then I can decide whether it's interesting or not. So there's some room for both.

Such exchanges exemplified the tensions that arose when one set of partners had difficulty understanding or assessing the veracity of data produced by the other set. The technology researcher learned what phenomena or patterns his partners in biology would value, and combined that knowledge with his own expertise in robotics and statistics.

5. DISCUSSION

The science and technology teams in the Center for Embedded Networked Sensing go into the field together with mutual interests in gathering more scientific data, at higher

sampling rates and finer granularity, at more locations, and with greater ability to adapt to field conditions than is possible with manual methods. The technology teams endeavor to learn enough about the science domain to collect data that are appropriate and accurate to scientific standards. The science teams endeavor to understand the technology well enough to assess the scientific accuracy of the data and to assist their technology partners in improving the scientific efficacy of their systems. Beyond these commonalities, their goals diverge. The scientists need the technology to gather observations of greater volume, variety, granularity, and complexity. Their scientific interest is to find patterns in their data. The technologists need domain problems of sufficient complexity to test their algorithms and code for actuation, guidance systems, fault detection, and networking. Each team's data is context to the other. Despite these interdependencies, the data are managed in ways that limit the ability to recombine them for later reuse.

We addressed four research questions spanning computer-supported collaborative work in science and technology, exploring concepts of data, interdependencies of practice, and tensions within and between teams, as discussed below.

5.1. What are the "data" in science and technology research collaborations?

We identified four categories of data that are gathered or generated from these sensor network deployments in the environmental sciences, as illustrated in Figure 2. The data types shown in the figure are examples of the many indicators that can be obtained from sensors and from physical sampling.

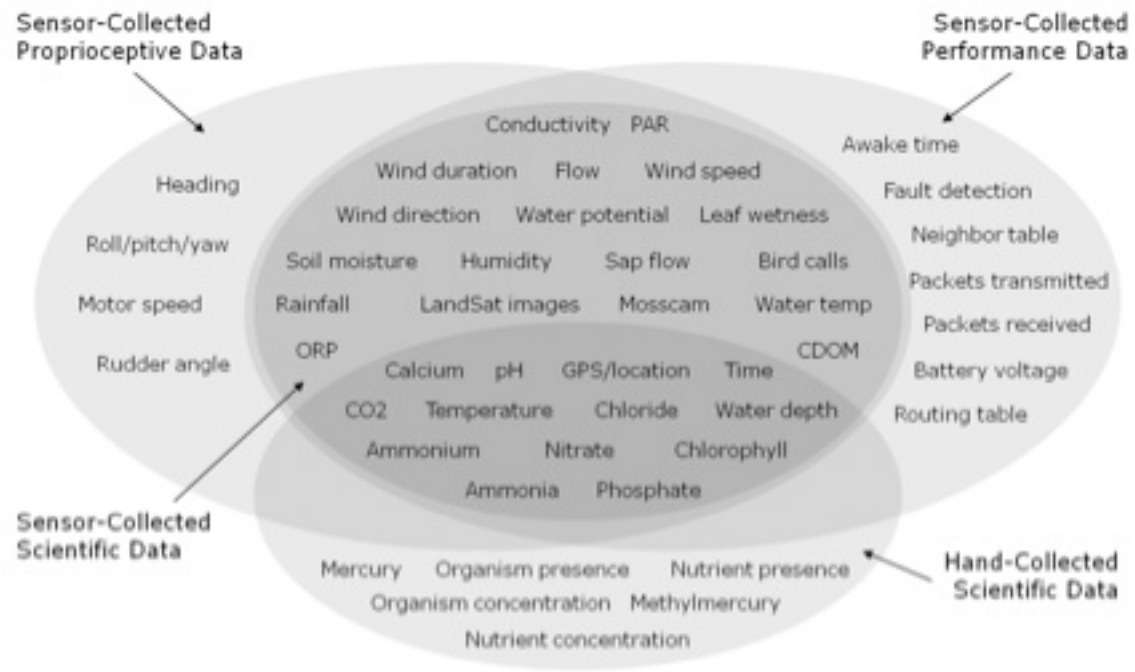

Figure 2: CENS Data Types from Joint Science-Technology Research Deployments. Based on figure that first appeared in Borgman, Wallis & Enyedy (2007).

In the center of the figure are the sensor-collected scientific data; these are the objects that first came to mind as "data" in our interviews with scientists and technologists alike. In the bottom set are the hand-collected scientific data. These are the physical samples of water, plants, and organisms that scientific teams collect. Some of these data are used to calibrate or corroborate the sensor data; some are used independently. At the top right are the sensor-collected performance data. Performance data are used to reconcile temporal and other indicators on the sensors so that observations can be matched and compared accurately. Technology teams use these data to improve the technology, to study network processing, and to assess the characteristics and quality of other types of data. Performance data are of central interest to the technology teams and of indirect value to

the science teams as context. The set at the top left of Figure 2 is sensor-collected proprioceptive data. These are used to guide mobile sensors on cables, boats, submarines, and other devices. The technology teams also use these data to refine their code and to clean and verify the sensor-collected science data. Once the scientific data are released to the science teams, the technology teams may keep or discard the sensor-collected performance and proprioceptive data.

5.2. How do concepts of "data" vary by purposes of research activity?

The four sets of data illustrated in Figure 2 serve two distinct purposes. To the scientists, data from the sensors and from the samples are evidence. They are seeking patterns in their observations of the natural world, making comparisons over time, location, and conditions. In the harmful algal blooms example, marine biologists seek to identify the conditions under which such blooms occur, and ultimately to prevent them. To do so, they need observations before, during, and after a bloom.

To the technologists, the data from sensor networks are a means to test and to improve their software code, algorithms, and instrument configurations. Their goals are to design and construct instruments that can help to reveal such patterns in the natural world. They are concerned with accurate and consistent measurements and with the ability to locate and respond to targets. While the science problem motivates their research problem, they can abstract away the science to accomplish their goals. If their software code can guide their robotic sensor to a target object quickly and accurately, while avoiding obstacles, their research is successful. The performance and proprioceptive data they generate in a

field study are indicators to be used for refining the code and the instruments. The code and instruments are the scholarly products of their research, rather than the data as evidence of scientific phenomena.

In sum, what is data to one team is context to the other. Two types of information gathered in field deployments – the sensor-collected scientific data and the physical samples – are *data* to the science teams, and the other two are context. Conversely, the information that is context to the science teams – the performance and proprioceptive data – are data to the technology teams, while the other types of data are context to them. The technology researchers are able to abstract away the scientific domain information in using those data to guide their sensing systems. The physical samples may also be viewed as context for them, but they have little direct involvement with those data.

5.3. What roles do data serve within and between science and technology collaborations?

Data remain useful as evidence of phenomena to the scientists over the long term. In contrast, the technology researchers rarely return to the sensor data, as these data serve as indicators of the effectiveness of their technology rather than as evidence, per se. The science teams keep their observations in some form indefinitely. They may or may not keep the physical samples and may or may not keep multiple states of the derived data, but in all cases the scientists claimed to keep the final, cleaned datasets. The science teams maintained their data for their own future uses, whether as points of comparison across time and sites or for calibration purposes. Most scientists were willing to share

their data, at least in principle, although few contributed their data to repositories or otherwise posted them. Our findings on sharing these data are reported elsewhere (Borgman, Wallis & Enyedy, 2007) and are a continuing line of inquiry (Borgman, 2012).

The technology teams were mixed in their choices of what data to maintain for the long term. Those who worked mostly with simulated data saw little reason to keep them. Those who were concerned with patterns of data flows in the network, network health, and modeling of sensor data had more reasons to maintain data from the field deployments, at least for awhile. These data did not appear to have the long-term value accorded to scientific observations, even though they inform the interpretation of those observations.

The research methods employed by CENS' joint science and technology teams reflect all four salient features of ecology research identified by Bowen and Roth (2007). In all the deployments observed, research design was highly emergent. Teams went into the field with stated goals, but finalized their data collection plans on site. While the tools were rarely developed in situ, they often were adapted to the local context with locally available materials. Keeping track of which sensors were used in what location was essential to later interpretation of the observations. Some sensors were "off the shelf" commercial grade, others were research-grade technologies developed by CENS researchers, and some were fashioned in the field out of sensor parts, aluminum foil, zip ties, duct tape, and other objects. Ecology studies remain difficult to replicate due to the

dynamic nature of field conditions, but the ability to obtain consistent data from instruments makes these studies more replicable. Social interactions between members of the community also have great significance, as Bowen and Roth found. Data collection is iterative, with team members consulting each other to determine subsequent activities. We observed stochastic data collection activities within and between the science and technology teams. The choice of action depended directly upon the assessment of the prior action, or on several prior actions.

While all four types of data shown in Figure 2 are obtained concurrently in deployments, their control is independent. The science teams control only their own physical samples and the derivations thereof. The technology teams control the other three sets of data. Performance and proprioceptive data are used to clean the scientific data from the sensors before releasing the cleaned data to the science teams. Once that process is complete, the performance and proprioceptive data may or may not be kept by the technology teams. Thus, an open question is whether curating the cleaned scientific sensor data and the sample derivations are sufficient for scientific replication and reuse. These data are the evidence reported in scientific publications. Subsequent reproducibility of the research is likely limited to the cleaned data. Rolling back the sensor-collected data to prior states will rarely be possible, due to the divergence of the datasets and to differential practices for retention of the four types of data.

In our decade of studying data in CENS, we continue to struggle with questions of what *are* the data and which of those data should be curated. Our ambitious plans to construct

a comprehensive data repository that would grow along with the Center were undermined by the diversity of the data, the lack of standards for data description, and the lack of infrastructure to support data curation (Wallis, Mayernik, Borgman & Pepe, 2010). The core of the challenge, as is evident from the findings reported here, is that "the data" are of at least four different types, each with its own considerations for curation. The physical samples are the only ones that have evidentiary value independent of the other three types. Even those samples are more valuable when corroborated with scientific data from the sensor networks. These two categories of data are observations, in the usual scientific sense (Long-Lived Digital Data Collections, 2005). The other two categories, sensor-collected performance and proprioceptive data, are indicators that are used to interpret the scientific data and to improve the code, algorithms, and physical instruments.

5.4. What can be learned about collaborative work by following the data?

The interdependencies of the science and technology teams resulted in at least three tensions in collecting and using data. The first was particularly evident in the early stages of CENS: the technology teams were focused on whether *any* data were flowing from the sensors, while the science teams were concerned with the quality of the data. These differences were most apparent around seemingly simple measurements such as temperature. While a technology team member might casually remark that "temperature is temperature," the scientists held such measurements to far higher standards. They wanted to know precisely the conditions and instrumentation being used to assess temperature. At one research site, the scientists were running parallel tests of three temperature sensors for a full year to determine their accuracy.

The second tension related to the different research rhythms of the collaborators. Scientists were willing to work with technology teams to develop useful technology for their own purposes. Once stabilized, they wanted to use the sensor instruments and networks for longer periods of time, ranging from months to years, to assure comparable observations. The technologists, however, usually were more interested in developing the next new technology than in "hardening" the prior ones. These differences in the rhythms of collaboration, where partners operate on different time scales, often plague joint ventures (Jackson, Ribes & Buyuktur, 2010).

A third tension arose around the mediation of access to scientific data by the technology researchers. Sensor data flows into the computers of those who control the sensors, which usually are the technologists. Once the data were collected in the field, the technology teams would reconcile time stamps and location information for the observations before releasing the dataset to the science teams. This task usually was delegated to a graduate student. Thus the science team could not acquire the data immediately; they might have to wait some days or weeks while the data were being cleaned. Sometimes two people might be between the data acquisition and the investigator – a person in the field who managed the sensors and another who cleaned the data. Scientific investigators were understandably concerned about the accuracy of the synchronization processes and the data handling. As evidenced in several of the interview quotations reported above, scientists were not always sure they were getting the full story or the full data from their partners. After a particularly devastating data loss, when the failure of sensors in a

deployment on a distant continent was evident only upon returning home, much more attention was paid to assessing sensor and network health in real time. In the latter period of our field observations, especially after more statisticians were involved in deployments, data could be viewed in real time. Trust in the data appeared to improve accordingly.

The interdependencies between the science and technology teams also reflected all three types identified by Shrum et al. (2007). In some cases, deployments used commercial grade sensors designed for other applications to support scientific inquiry. In other cases, the technology itself was adapted in support of science. In many cases, especially with the robotic instruments we observed, the technology teams were designing and deploying novel instruments in support of the science teams' goals. To the extent that standard protocols for data collection existed, they were the science teams' practices for gathering and analyzing physical samples. The sensor-collected data varied by instrument, and each team made its own choices for what, how, and where the data were to be kept. While Collins (1975) found that collaboration was richest after data collection, when participants explore the meaning of data and phenomena, we found the opposite. The science and technology teams took their data to their respective labs to analyze independently. This difference is likely due to Collins' focus on intra-disciplinary collaborations between physicists, whereas CENS collaborations are highly interdisciplinary.

6. CONCLUSIONS

Science and technology have been interdependent since Galileo's time. Today, few individuals have Galileo's talents as both instrument builder and scientific observer. The requisite expertise is spread across teams of scientists and technology researchers, each learning enough about the other's domain to address common problems together. As teams grow in size and technology increases in complexity, the interdependencies multiply. However, those dependencies may be deeply buried in the minutiae of complex systems and research methods. Teasing out those relationships may require many years of study, as we report here.

From our nuanced descriptions of research activity, we draw five conclusions for cooperative work and for science policy. These are based on the interdependencies of scientific and technology research practices. First is that "data" is a complex notion, and one that is not well understood even by the parties creating and using them. Data, like beauty, exist in the eye of the beholder. Researchers give only partial accounts of their data when asked. What one team or individual considers to be data may not be recognized as such by another. Concepts of data vary considerably by research activity and by individual. One consequence is that data cannot be managed and shared as a "black box." The box must be pried open and its contents examined, under a microscope if necessary, to determine what data exist in a collaboration, how they are used, and how they are managed or should be managed.

Our second conclusion is that science and technology researchers depend upon each other's data for interpretation of their own data. What are data to one researcher are context to another. The relationship is reflexive. Scientists use the sensor network data produced by the technology teams to interpret and to corroborate evidence gathered from physical samples of water, soil, sand, or other matter. The scientists thus depend upon the technology researchers to obtain useful data. Conversely, the technology researchers need the scientific data to validate their algorithms and to interpret their results. Without the scientists, the technology researchers lack real world problems to study.

Third, data curation practices are not consonant with the interdependence of the teams on each other's data. Of the four categories of data we identified, only one was deemed worthy of preservation by all parties: scientific data from the sensor networks. These data are evidence to the scientists and are context to the technology teams. The other three categories of data are managed separately, if they are kept at all. The scientists keep records of data derived from hand-collected samples, whether or not they keep they physical materials. The technology researchers use sensor-collected proprioceptive and performance data such as system health, guidance accuracy, speed, and direction for adjustments in the field, but often discard them after the deployment or after the paper reporting their findings was written. Simulated data used to develop the technologies were not of long-term interest. Technology researchers do keep and maintain their software, however, and may submit them to a code repository for public reuse.

Our fourth conclusion is that interdependent field research of the type described here tends not to produce data that are easily reusable, particularly for outside researchers. Neither the scientific nor the technological research data can be interpreted without the other, yet these datasets are quickly separated, never to be reconciled again. As a result, none of these data remain useful beyond the teams that generated them. Even those teams may not be able to reuse them, given the difficulty of obtaining the context data necessary for interpretation, which often are held by others. Reconstructing such data sets for the purpose of reuse would require extensive communication with the researchers who collaborated to collect those data. A corollary of this conclusion is that interdependent research studies of this type are difficult, if not impossible, to replicate.

Lastly, successful collaboration depends upon each party working in good faith to capture the best possible data to the standards of their partners. The scientists need to trust their technology partners' ability to obtain scientifically accurate data. The technology researchers depend upon scientists' ability to document accurately their use of the technology. When these elements were not in place, as they sometimes were not, the collaboration weakened. Technology researchers did not always respect scientific standards for instrumentation. Scientists sometimes moved or adjusted sensors without recording those changes, which made the sensor records difficult to interpret.

Overall, we found that following the data provides a rare set of insights into science and technology research collaborations, with significant implications for data curation and scientific policy. Researchers are often unaware, or minimally aware, of what their

partners consider to be valid and reliable data. While such tensions are not new, they take on new implications in an age of data management, data sharing, and pressure to make research results more replicable. For data to be sharable and reusable, they must be interpretable. We find that the context necessary for interpretation is quickly lost in these partnerships. Few of these data are made available in public repositories or otherwise released. Technological advances in scientific data collection do not necessarily lead to advances in the management of those data. Until partners begin to recognize the value in each other's data for their own purposes, and for the collective good, the situation is unlikely to change. Much more research is needed on how data are created, conceived, handled, managed, and curated in multi-disciplinary collaborations to inform science policy and practice.

ACKNOWLEDGEMENTS

Research reported here is supported in part by grants from the National Science Foundation (NSF): (1) The *Center for Embedded Networked Sensing* (CENS) is funded by NSF Cooperative Agreement #CCR-0120778, Deborah L. Estrin, UCLA, Principal Investigator; (2) CENS Education Infrastructure (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. (3) *Towards a Virtual Organization for Data Cyberinfrastructure*, #OCI-0750529, C.L. Borgman, UCLA, PI; G. Bowker, Santa Clara University, Co-PI; Thomas Finholt, University of Michigan, Co-PI; (4) *Monitoring, Modeling & Memory: Dynamics*

of Data and Knowledge in Scientific Cyberinfrastructures: #0827322, P.N. Edwards, UM, PI; Co-PIs C.L. Borgman, UCLA; G. Bowker, SCU; T. Finholt, UM; S. Jackson, UM; D. Ribes, Georgetown; S.L. Star, SCU.

We also are grateful to Microsoft Technical Computing and External Research for gifts in support of this research program. The authors would also like to thank Archer Batcheller, David Fearon, George Mood, Alberto Pepe, Katie Shilton, Elizabeth Rolando, and Laura Wynholds for their thoughtful comments on prior drafts of this paper.

REFERENCES

Aronova, E., Baker, K. S. & Oreskes, N. (2010). Big Science and Big Data in Biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) Network, 1957–Present. Historical Studies in the Natural Sciences, vol. 40, no. 2, pp. 183-224.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G. C., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. F. & Wouters, P. (2004a). An International Framework to Promote Access to Data. Science, vol. 303, no. 5665, pp. 1777-1778.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G. C., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. F. & Wouters, P. (2004b). Promoting Access to Public Research Data for Scientific, Economic, and Social Development. Data Science Journal, vol. 3, pp. 135-152.

Basili, V. R. & Zelkowitz, M. V. (2007). Empirical studies to build a science of computer science. Communications of the ACM, vol. 50, no. 11, pp. 33-37.

Batalin, M. A., Rahimi, M., Yu, Y., Liu, D., Kansal, A., Sukhatme, G. S., Kaiser, W. J., Hansen, M., Pottie, G. J., Srivastava, M. & Estrin, D. (2004). Call and Response: Experiments in Sampling the Environment. Proceedings of the 2nd international conference on Embedded networked sensor systems, Los Angeles, New York, NY: ACM Press. pp. 25-38. Retrieved from http://cres.usc.edu/pubdb_html/files_upload/420.pdf on 28 September 2006.

Berman, H. M., Westbrook, J., Feng, J., Gilliland, G., Bhat, T. N., Wessig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Research, vol. 28, pp. 235-242.

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

Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology. Retrieved from http://dx.doi.org/10.1002/asi.22634 on 15 April 2012.

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. pp. 170-183.

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, vol. 7, nos. 1-2, pp. 17-30.

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.

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. Joint Conference on Digital Libraries, Vancouver, British Columbia, Canada, Association for Computing Machinery. pp. 269-277. Retrieved from http://doi.acm.org/10.1145/1255175.1255228 on 1 February 2010.

Bourne, P. (2005). Will a biological database be different from a biological journal? PLoS Computational Biology, vol. 1, no. 3, pp. e34. Retrieved from http://dx.doi.org/10.1371/journal.pcbi.0010034 on 28 September 2006.

Bowen, G. M. & Roth, W.-M. (2007). The practice of field ecology: Insights for science education. Research in Science Education, vol. 37, no. 2, pp. 171-187.

Carver, J., Hochstein, L., Kendall, R., Nakamura, T., Zelkowitz, M., Basili, V. & Post, D. (2006). Observations about software development for high-end computing. Cyberinfrastructure Technology Watch Quarterly, vol. 2, no. 4a, pp. 33-38. Retrieved from http://www.ctwatch.org/quarterly/articles/2006/11/observations-about-software-development-for-high-end-computing/ on 20 April 2010.

Chen, J. C., Elson, J., Wang, H., Maniezzo, D., Hudson, R. E., Yao, K. & Estrin, D. (2003). Coherent Acoustic Array Processing and Localization on Wireless Sensor Networks. Proceedings of the IEEE, vol. 91, no. 8, pp. 1154-1162.

Cole, F. T. H. (2008). Taking "Data" (as a Topic): The Working Policies of Indifference, Purification and Differentiation. 19th Australasian Conference on Information Systems, Christchurch, NZ. pp. 240-249.

Collins, H. M. (1975). The seven sexes: A study in the sociology of a phenomenon, or the replication of experiments in physics. Sociology, vol. 9, pp. 205-24.

Cragin, M. H. & Shankar, K. (2006). Scientific data collections and distributed collective practice. Journal of Computer Supported Cooperative Work, vol. 15, pp. 185-204.

Cyberinfrastructure Vision for 21st Century Discovery (2007). National Science Foundation. Retrieved from http://www.nsf.gov/pubs/2007/nsf0728/ on 17 July 2007.

de Souza, C., Froehlich, J. & Dourish, P. (2005). Seeking the source: software source code as a social and technical artifact. Proceedings of the 2005 international ACM SIGGROUP Conference, Sanibel Island, Florida, Association for Computing Machinery. pp. 197-206. Retrieved from http://doi.acm.org/10.1145/1099203.1099239 on 20 April 2010.

Easterbrook, S. M. & Johns, T. C. (2009). Engineering the software for understanding climate change. Computing in Science & Engineering: pp. 64-74.

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C. & Borgman, C. L. (2011). Science Friction: Data, Metadata, and Collaboration. Social Studies of Science, vol. 41, no. 5, pp. 667-690.

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.

Estrin, D., Michener, W. K. & Bonito, G. (2003). Environmental cyberinfrastructure needs for distributed sensor networks: A report from a National Science Foundation sponsored workshop. Scripps Institute of Oceanography. Retrieved from http://www.lternet.edu/sensor_report/ on 12 May 2006.

Faniel, I. M. & Jacobsen, T. E. (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. Journal of Computer-Supported Cooperative Work, vol. 19, nos. 3-4, pp. 355-375. Retrieved from http://dx.doi.org/10.1007/s10606-010-9117-8 on 15 April 2012.

Fry, J. (2006). Scholarly research and information practices: A domain analytic approach. Information Processing and Management, vol. 2006, no. 42, pp. 299-316.

GEON. (2010). Retrieved from http://www.geongrid.org/ on 20 August 2010.

Giere, R. N. (1999). Science without Laws. Chicago: University of Chicago Press.

Hamilton, M. P., Graham, E. A., Rundel, P. W., Allen, M. F., Kaiser, W., Hansen, M. H. & Estrin, D. L. (2007). New Approaches in Embedded Networked Sensing for Terrestrial Ecological Observatories. Environmental Engineering Science, vol. 24, no. 2.

Harmon, T. C., Ambrose, R. F., Gilbert, R. M., Fisher, J. C., Stealey, M. & Kaiser, W. J. (2007). High-Resolution River Hydraulic and Water Quality Characterization Using Rapidly Deployable Networked Infomechanical Systems (NIMS RD). Environmental Engineering Science, vol. 24, no. 2, pp. 151-159.

Jackson, S. J., Ribes, D. & Buyuktur, A. (2010). Exploring Collaborative Rhythm: Temporal Flow and Alignment in Collaborative Scientific Work. iConference 2010, Urbana-Champagne, IL.

Jirotka, M., Procter, R., Rodden, T. & Bowker, G. C. (2006). Special Issue: Collaboration in e-Research. Journal of Computer Supported Cooperative Work, vol. 15, pp. 251-255.

Kanfer, A. G., Haythornthwaite, C., Bruce, B. C., Bowker, G. C., Burbules, N. C., Porac, J. F. & Wade, J. (2000). Modeling distributed knowledge processes in next generation multidisciplinary alliances. Information Systems Frontiers, vol. 2, nos. 3-4, pp. 317-331.

Karasti, H., Baker, K. S. & Halkola, E. (2006). Enriching the notion of data curation in e-Science: Data managing and information infrastructuring in the Long Term Ecological Research (LTER) Network. Journal of Computer-Supported Cooperative Work, vol. 15, no. 4, pp. 321-358.

Karasti, H., Baker, K. S. & Millerand, F. (2010). Infrastructure Time: Long-term Matters in Collaborative Development. Computer Supported Cooperative Work, vol. 19, nos. 3-4, pp. 377-415.

Kwa, C. (2005). Local ecologies and global science: Discourses and strategies of the International Geosphere-Biosphere Programme. Social Studies of Science, vol. 35, no. 6, pp. 923-950.

Latour, B. (1987). Science in Action: How to Follow Scientists and Engineers through Society. Cambridge, MA: Harvard University Press.

Lawrence, K. A. (2006). Walking the Tightrope: The Balancing Acts of a Large e-Research Project. Journal of Computer Supported Cooperative Work, vol. 15, pp. 385-411.

Lee, C. P., Dourish, P. & Mark, G. (2006). The human infrastructure of cyberinfrastructure. Proceedings of the Conference on Computer-Supported Cooperative Work, Banff, Alberta, Association for Computing Machinery. pp. 483-492.

Lee, C. P., Ribes, D., Bietz, M., Jirotka, M. & Karasti, H. (2010). Supporting Scientific Collaboration Through Cyberinfrastructure and e-Science: Special issue. Journal of Computer Supported Cooperative Work, vol. 19, nos. 3-4.

Long-Lived Digital Data Collections. (2005). National Science Board. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/ on 18 April 2009.

Maurer, B. A. (2004). Models of Scientific Inquiry and Statistical Practice: Implications for the structure of scientific knowledge. In Taper, M. L. & Lele, S. R. (Eds.). The Nature of Scientific Evidence: Statistical, philosophical, and empirical considerations. Chicago, London, The University of Chicago Press, pp. 17-50.

Mayernik, M. S. (2011). Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators. PhD Dissertation. Information Studies. UCLA. Los Angeles. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2042653 on 15 April 2012.

Mayernik, M. S., Batcheller, A. L. & Borgman, C. L. (2011). How Institutional Factors Influence the Creation of Scientific Metadata. iConference, Seattle, Association for Computing Machinery.

Mayernik, M. S., Wallis, J. C. & Borgman, C. L. (in review). Unearthing the infrastructure: Humans and sensors in environmental and ecological field research.

Mun, M., Reddy, S., Shilton, K., Yau, N., Burke, J., Estrin, D., Hansen, M., Howard, E., West, R. & Boda, P. (2009). PEIR, the Personal Environmental Impact Report, as a Platform for Participatory Sensing Systems Research. Proceedings of the International Conference on Mobile Systems, Applications, and Services. Krakow, Poland. pp. 1-14.

National Ecological Observatory Network. (2010). Retrieved from http://www.neoninc.org/ on 20 August 2010.

NIMS: Networked Infomechanical Systems. (2006). Retrieved from http://www.cens.ucla.edu/portal/nims on 3 October 2006.

Pepe, A. (2010). Structure and Evolution of Scientific Collaboration Networks in a Modern Research Collaboratory. Doctoral. Information Studies. UCLA. Los Angeles. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1616935 on 15 June 2011.

Pepe, A. & Rodriguez, M. A. (2010). Collaboration in sensor network research: an in-depth longitudinal analysis of assortative mixing patterns. Scientometrics, vol. 84, no. 3, pp. 687-701. Retrieved from http://www.springerlink.com/content/v1w5695932tg52g2/ on 1 February 2010.

Pon, R., Maxim Batalin, M., Gordon, J., Rahimi, M. H., Kaiser, W., Sukhatme, G. S., Srivastava, M. & Estrin, D. (2005). Networked Infomechanical Systems: A Mobile Wireless Sensor Network Platform. IEEE/ACM Fourth International Conference on Information Processing in Sensor Networks (IPSN-SPOTS). pp. 376-381. Retrieved from http://cres.usc.edu/pubdb_html/files_upload/450.pdf on 30 April 2006.

Protein Data Bank. (2006). Retrieved from http://www.rcsb.org/pdb/ on 4 October 2006.

A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases. (1999). Washington, DC: National Academy Press.

Rahimi, M. H., Kaiser, W., Sukhatme, G. S. & Estrin, D. (2005). Adaptive sampling for environmental field estimation using robotic sensors IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3692-3698. Retrieved from http://research.cens.ucla.edu/pls/portal/url/item/10532382ACC965BBE0406180528D77CF on 4 October 2006.

Renear, A. H., Sacchi, S. & Wickett, K. M. (2010). Definitions of Dataset in the Scientific and Technical Literature. American Society for Information Science and Technology, Pittsburgh, Information Today. pp. 1-4. Retrieved from http://portal.acm.org/citation.cfm?id=1920447 on 21 June 2011.

Ribes, D. & Finholt, T. A. (2007). Tensions across the scales: Planning infrastructure for the long-term. Proceedings of the 2007 International ACM SIGGROUP Conference on Supporting Group Work, Sanibel Island, Florida, USA, Sanibel Island, Florida, Association for Computing Machinery. pp. 229-238.

Ribes, D. & Lee, C. P. (2010). Sociotechnical Studies of Cyberinfrastructure and e-Research: Current Themes and Future Trajectories. Journal of Computer Supported Cooperative Work, vol. 19, nos. 3-4, pp. 231–244.

Segal, J. (2005). When software engineers met research scientists: A case study. Empirical Software Engineering, vol. 10, pp. 517-536.

Segal, J. (2009). Software Development Cultures and Cooperation Problems: A Field Study of the Early Stages of Development of Software for a Scientific Community. Computer Supported Cooperative Work: pp. 1-26.

Shrum, W., Genuth, J. & Chompalov, I. (2007). Structures of Scientific Collaboration. Cambridge, MA: MIT Press.

Star, S. L. & Griesemer, J. (1989). Institutional ecology, "translations," and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-1939. Social Studies of Science, vol. 19, no. 3, pp. 387-420.

Sutton, C. (2003). UCLA Develops Mobile Sensing System for Enriched Monitoring of the Environment. UCLA. Retrieved from http://www.engineer.ucla.edu/stories/2003/nims.htm on 4 October 2006.

Szewczyk, R., Osterweil, E., Polastre, J., Hamilton, M., Mainwaring, A. & Estrin, D. (2004). Habitat monitoring with sensor networks. Communications of the ACM, vol. 47, no. 6, pp. 34-40.

Traweek, S. (1992). Beamtimes and Lifetimes: The World of High Energy Physicists (1st Harvard University Press pbk. ed.). Cambridge, Mass.: Harvard University Press.

Traweek, S. (2004). Generating high energy physics in Japan. In Kaiser, D. (Ed.). Pedagogy and Practice in Physics. Chicago, University of Chicago Press.

Turner, W., Bowker, G. C., Gasser, L. & Zacklad, M. (2006). Information Infrastructures for Distributed Collective Practices. Journal of Computer Supported Cooperative Work, vol. 15, pp. 93-110.

U.S. Long Term Ecological Research Network. (2010). Retrieved from http://lternet.edu/ on 20 August 2010.

Uhlir, P. F. & Cohen, D. (2011). Personal communication. Board on Research Data and Information, Policy and Global Affairs Division, National Academy of Sciences. 18 March 2011.

Voorhees, E. M. (2007). TREC: Continuing information retrieval's tradition of experimentation. Communications of the ACM, vol. 50, no. 11, pp. 51-54.

Voorhees, E. M. & Harman, D. K. (Eds.). (2005). TREC: Experiment and Evaluation in Information Retrieval. Cambridge, MA: MIT Press.

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, vol. 3, no. 1. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/67 on 24 November 2008.

Wallis, J. C., Mayernik, M. S., Borgman, C. L. & Pepe, A. (2010). Digital Libraries for Scientific Data Discovery and Reuse: From Vision to Practical Reality. Joint Conference on Digital Libraries, Gold Coast, Queensland, Australia, Association for Computing Machinery.

Wallis, J. C., Pepe, A., Mayernik, M. S. & Borgman, C. L. (2008). An exploration of the life cycle of eScience collaboratory data. iConference 2008, Los Angeles, CA.