UNIVERSITY OF CALIFORNIA

Los Angeles

The Distribution of Data Management Responsibility within Scientific Research Groups

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Information Studies

by

Jillian Claire Wallis

2012

ABSTRACT OF THE DISSERTATION

The Distribution of Data Management Responsibility

within Scientific Research Groups

Jillian Claire Wallis

Doctorate of Philosophy in Information Studies

University of California, Los Angeles, 2012

Professor Christine L. Borgman, Chair

Scientific data often are expensive to produce or impossible to reproduce. Those data may be of great future value for reuse, recombination, and replication by other researchers. However, the potential value of these data can only be achieved if the data producers manage them properly. Visions of data management and the role of the data producer have been constructed by data curators and funders from the top-down, but we have little understanding of what data management looks like on the ground. What do data producers see as their data management responsibilities? The exploratory research reported in this dissertation provides a rich description of data management tasks performed by members of six research groups and members' perception of data management responsibilities. Groups were selected from the Center for Embedded Networked Sensing (CENS), an NSF-funded Science and Technology Research Center, where researchers are already experiencing the data deluge. Document analysis, semi-structured interviews, and field observations were coded and analyzed for emergent themes and used to construct models of data management practices. Significant findings include: (i) these six research groups acquired a diverse array of data (ii) a generalized data life cycle can be applied to practices of these groups, (iii) researchers actively managed their data throughout the data life cycle to support their own use, and (iv) data management tasks were distributed between the members of a research group, and are tied to data handling tasks such as collection, processing, and analysis. The data management tasks performed by researchers are categorized into four core functions: selection for quality, verification for validity, storage for accessibility, and documentation for interpretability. A set of roles and responsibilities were identified for the data producers collaborating on each research project. These findings suggest that including author contribution statements in publications would assist future users of those data in determining who to contact for questions about their creation and context. This study reveals how, when, and why science and technology researchers manage their data and makes recommendations for data management within research groups that will make data more usable and sharable.

The dissertation of Jillian Claire Wallis is approved.

Jonathan Furner

Christopher Kelty

Deborah L. Estrin

Christine L. Borgman, Chair

University of California, Los Angeles

2012

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

To establish a baseline understanding, the following terms are defined as they have been used throughout this dissertation.

Data. Data are a class of information used by researchers to measure and understand phenomena. The term ‘data’ is used to describe any things that the researchers studied classified as data, including water samples, numerical measures, timeseries, and aggregate or derivative calculations.

Data management. Data management is the active management of data for future use. Data management is a process that includes policies, methods, practices, and infrastructure that add value to the data and support their discovery and reuse.

Task. The term ‘task’ is used to designate work performed that supports data management.

Function. The function of a given task is the intended outcome of that task.

Responsible. A person is held responsible for something they are in charge, or appointed to look after. For instance, individuals are held responsible for completing tasks or fulfilling functions. Responsible here is synonymous with “accountable.”

Responsibility. A task or function for which one is responsible.

Research. Research is the systematic investigation or inquiry aimed at contributing to knowledge of a theory, topic, etc. The research under study within this dissertation ranges from study of natural phenomena, to testing new methods, to evaluating and refining technology.

Research project. A research project is a specific investigation or inquiry.

Researcher. An individual, i.e., student, faculty, or staff, who performs systematic investigation or inquiry.

Research lab. A research lab is the group comprised of a faculty member, and their associated student researchers, research faculty, and staff. The faculty member is the head of the lab, and also known as the primary or principal investigator (PI).

Research group. The research group consists of two or more researchers collaborating to perform a specific research project. The research group may be members from a single research lab, or a combination of subsets from multiple research labs, as is common in inter-disciplinary collaboration.

Distribution. Distribution is the action of dividing among a number of individuals. Specifically, this dissertation is concerned with the division of research and data management labor among members of a research group.

Life cycle. A life cycle is a course from a beginning, through development and productivity, to decay or ending. A life cycle is made up of discrete stages of import to some object, where either the object is affected by external processes, or the object affects the world.

ACKNOWLEDGEMENTS

The dissertation genre is odd, because only one person's name appears on work which has been the product of so many people. I would like to take this opportunity to acknowledge the contributions of those people who have helped this document come into being.

Prof. Christine Borgman has served as my advisor on two degrees, boss for eight years, mentor, and now dissertation committee chair. Her activity, teaching style, research ethic, and career are an inspiration, but even more important is her patience and nurturing. She has provided me with many opportunities to develop myself, and I can only hope that I will make her proud as I embark on my own academic career.

My dissertation committee members, Prof. Deborah L. Estrin, Prof. Jonathan Furner, and Prof. Christopher Kelty, provided excellent constructive criticism and pushed this dissertation research in directions I could not have come up with on my own. I wish I had taken more advantage of them during the writing process.

The Borgman Research Group has seen a rotating group of colleagues during my tenure, all of whom have contributed to a generative research atmosphere. The group has included, in alphabetical order: David Fearon, Ariel Hernandez, Rachel Mandell, Matt Mayernik, Stasa Milojevic, Alberto Pepe, Lizzy Rolando, Ashley Sands, Kalpana Shankar, Katie Shilton, and Laura Wynholds. Every one of these people has brought something special to the group, from Matt's quiet brilliance to Rachel's infectious laughter, from Katie's class to Alberto's enterprising spirit, and from Lizzy's enthusiasm to Laura's frankness. These people have provided a tremendous amount of support during the last nine years, including snacks, feedback, inspiration, and commiseration. I hope we remain close colleagues and friends in the future.

The Borgman Group has also had the fortune to collaborate with other faculty at UCLA, including Prof. Noel Enyedy and Prof. Bill Sandoval from education, and Prof. Sharon Traweek from gender studies and history. The UCLA community has so many amazing researchers, and getting time to work with them is such an honor.

We have also collaborated with researchers beyond UCLA. The Modeling, Monitoring, and Memory research collaboration has, over the last four years, pushed my research in new directions and exposed me to iSchool faculty that I would not have otherwise had a chance with which to work. This collaboration has included, in alphabetical order: Archer Batchellor, Chris Borgman, Ayse Buyuktur, Geof Bowker, Scout Calvert, Paul Edwards, Matt Mayernik, Tom Finholt, Steve Jackson, David Ribes, Lizzy Rolando, and the late Susan Leigh Star. We came together to identify common themes across our study of science-based cyberinfrastructure projects. Although our formal NSF funding ends this year, this group is evolving to become something more, a community engaged broadly in the study of knowledge infrastructures. I look forward to future conference calls and workshops with this lively community.

The members of the Center for Embedded Networked Sensing (CENS) have been my research community and my community of research. This vibrant and active center has provided a rich substrate for the study of data practices across collaborators from a variety of disciplines. Faculty, student, and staff members have been generous of their time and providing access – agreeing to be interviewed and followed around as they work.

My family, friends, and cats should be acknowledged for their commitment to my educational journey and support in many ways over this extended period of my life. It is trying to have a permanent student in your family – they are always broke, mentally exhausted, and with all the freedom in their schedule somehow do not have time to vacuum – so it is commendable that my student-years have been tolerated this long. The sweets will be worth the effort. My partner, Bryan C. Tysinger, specifically, should be acknowledged for his kind work in proofreading this document and his steady navigation through all the ups and downs that went into its creation.

And I would finally like to acknowledge the Information Studies Department. I may have only slept there occasionally, but it has been my home for the last nine years. The IS department students, faculty, and staff are so open, welcoming, and generous with their time and attention.

The research that lead up to and including this dissertation has been generously supported by the National Science Foundation, Microsoft Technical Computing and External Research, and the Department of Information Studies. NSF grants include: (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.

VITA

2001 B.A. Studio Art, with minor in Philosophy, Scripps College, Claremont, California.

2004-2005 Graduate Student Researcher, Center for Embedded Networked Sensing (CENS), University of California, Los Angeles (UCLA)

2005 Masters of Library and Information Science, Graduate School of Education and Information Studies, UCLA.

2005-2008 Data Archivist, CENS, UCLA.

2010, 2011 Special Reader, IS 270: Information Technology, Department of Information Studies, UCLA.

2008-2012 Graduate Student Researcher, CENS, UCLA.

PUBLICATIONS

Wallis, J. C., Rolando, E. J. & Borgman, C. L. (in progress). If We Share Data, Will Anyone Use Them? Data sharing and reuse in the long tail of science and technology.

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

Borgman, C. L., **Wallis, J. C.** & Mayernik, M. S. (forthcoming). Who's got the data? Interdependencies in Science and Technology Collaborations. *Journal of Computer Supported Collaborative Work.*

Wallis, J. C., Wynholds, L. A., Borgman, C. L., Sands, A. Traweek, S. (2012). Data, Data Use, and Scientific Inquiry: Two case studies of data practices. *Proceedings of the Joint Conference on Digital Libraries*, Washington, DC. http://dl.acm.org/citation.cfm?id=2232822.

Wallis, J. C. & Borgman, C. L. (2012). Who is responsible for data? An exploratory study of data authorship, ownership, and responsibility. *Proceedings of the American Society for Information Science and Technology*, 48(1). http://onlinelibrary.wiley.com/doi/10.1002/meet.2011.14504801188/abstract.

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. *Proceedings of the 10th Annual Joint Conference on Digital libraries*, 333-340.

Borgman, C. L., Bowker, G. C., Finholt, T. A. & **Wallis, J. C.** (2009). Towards a Virtual Organization for Data Cyberinfrastructure. *Proceedings of the Joint Conference on Digital Libraries*, Austin, TX. http://dl.acm.org/citation.cfm?id=1555459.

Wallis, J. C., Borgman, C. L., Mayernik, M. S. & Pepe, A. (2008). Moving archival practices upstream: An exploration of the life cycle of ecological sensing data in collaborative field research. *International Journal of Digital Curation*, 3(1): 114-126.

Wallis, J. C., Pepe, A., Mayernik, M. S. & Borgman, C. L. (2008). An exploration of the life cycle of eScience collaboratory data. *iConference*, Los Angeles, CA. http://hdl.handle.net/2142/15122.

Mayernik, M. S., **Wallis, J. C.**, Pepe, A. & Borgman, C. L. (2008). Whose data do you trust? Integrity issues in the preservation of scientific data. *iConference*, Los Angeles, CA. http://hdl.handle.net/2142/15119.

Pepe, A., Borgman, C. L., **Wallis, J. C.** & Mayernik, M. S. (2007). Knitting a fabric of sensor data and literature. *Proceedings of Information Processing in Sensor Networks*, Cambridge, MA, Association for Computing Machinery/IEEE. http://works.bepress.com/albertopepe/5/.

Borgman, C. L., **Wallis, J. C.** & Enyedy, N. (2007). Little Science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, 7(1-2): 17-3029.

Wallis, J. C., Borgman, C. L., Mayernik, M. S., Pepe, A., Ramanathan, N. & Hansen, M. (2007). Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries. *Proceedings of the 11th European Conference on Digital Libraries*, Budapest, Hungary. http://escholarship.org/uc/item/4xx221vv.

Borgman, C. L., **Wallis, J. C.**, Mayernik, M. S. & Pepe, A. (2007). Drowning in data: Digital library architecture to support scientific use of embedded sensor networks. Vancouver, British Columbia, Canada, Association for Computing Machinery: 269-277.

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

CHAPTER 1: INTRODUCTION

The research presented in this dissertation is the culmination of eight years of research performed at the Center for Embedded Networked Sensing (CENS). CENS is an NSF-funded Science and technology center established in 2002, and ends in 2012. CENS research has been focused on the development of embedded networked sensing systems to support a few scientific application disciplines. The CENS community is quite diverse, being made up of roughly 300 researchers across 5 research institutions. Research at CENS includes technology areas such as robotics, computer science, and electrical engineering, as well as scientific applications including marine biology, ecology, environmental engineering, and seismology.

Over the last eight years I have been part of the data practices research team within the center, performing data practices studies and developing systems to support collaborative research. The overarching goal of our research was to study the ecology of research data, data practices, information infrastructure, and data policies. We have been particularly concerned with how to effectively share data for reuse, as they are a valuable research investment. During a round of interviews in 2009-2010, we asked researchers from CENS who would be the author of their data. At the time we were thinking about creating data citations, where there would need to be a data “author.” We were trying to identify who would be the person to deposit data in repositories, as this person would need to have the authority to release the data. For the majority of researchers interviewed, “data author” was not a designation with which they were comfortable. In some interviews, the researchers would offer other ways to identify themselves, such as “owner” or “collector.” In the following quote, the participant identified himself as “responsible” for the data because of his role as lead investigator of the project.

> 1:29:28 CB: Okay. When would you consider yourself to be the author of a dataset?
> 1:29:43 P21: When it is our deployment. And I have paid for it. [chuckle]
> 1:29:50 CB: Okay. Good. What are the criteria that determine authorship?
> 1:29:59 P21: For the dataset?
> 1:29:59 CB: For the dataset.
> 1:30:00 P21: Let's see, I think that if you're the lead investigator or the leader of the deployment that was, whether the equipment and the students are funded, I'd buy it and they're kind of responsible for that data and sort of lead around the data (Interview 21, 2010).

Recent data sharing efforts are developing incentives, infrastructure, and policy to encourage data sharing, but they all rely on someone to be responsible in some way. Many researchers will only share data if they are given credit (Wallis, Rolando & Borgman, in progress), and data citation initiatives, such as DataCite, are developing standard data citations that are both human and machine-readable. In order to create a citation, someone will need to be cited as the person responsible for the creation and/or maintenance of the data. Many researchers are sharing their data on personal websites, but links decay rapidly. A more sustainable solution for data hosting is the development of domain-specific data repositories, such as GenBank. Data repositories will only gain traction if there is someone to deposit their data in a data repository. The NSF and NIH data management plan requirements are also predicated on the idea that there is someone accountable for the creation and long-term usability of research data. According to the NSF, grantees are expected to encourage and facilitate data sharing (NSF, 2011a), and the data management plan should describe how the grantee intends to fulfill this requirement (NSF, 2011b). Who will be cited, deposit data, and be responsible for data management?

Data management, data curation, and other terms such as data stewardship are all in a state of flux. How they overlap and relate is currently up for grabs. Unfortunately these terms are being used, such as the NSF Data Management Plan requirement, without a clear definition of data management. Throughout this dissertation, I am really speaking more towards data curation, especially as defined by the Center for Informatics Research in Science and Scholarship's Data Curation Education Program and Lord and MacDonald.

> "Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to scholarly and educational activities across the sciences, social sciences, and the humanities. Data curation activities enable data discovery and retrieval, maintain data quality, add value, and provide for re-use over time" (CIRSS).

> "The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. … Archiving is a curation activity which ensures that data are properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity. … Preservation is an activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology" (Lord & Macdonald, 2003).

In these two definitions, data curation is defined by a set of functions that make up data curation, such as enabling data discovery and supporting data reuse. Data management tends to imply only data storage and documentation functions, whereas data curation encompasses a broader set of activities the support data reuse more generally. Although data curation is a better fit, I use the term data management within this dissertation, because this term is subsumed under data curation and at this point in time is more meaningful to the scientific community being studied.

Data management has many stakeholders, such as the data producer who created the data, data users who reuse the data, research funders who want to maximize investment, and data managers or curators who preserve the data for reuse. In response to the mandates from funding agencies regarding data management and sharing, the would-be data curators have developed strategies to encourage data management and sharing, identifying stakeholders and setting out responsibilities. Unfortunately, the responsibilities set out for data producers in these strategy reports are no more specific than what the NSF mandate requires.

Within the data curation literature, the data producer and the context of data creation are mentioned as important to data curation. Yet models of data curation developed within the data curation community treat the science that leads to the production of data as a black box. Similarly, the data producer's contribution to curation is invisible, as data curation models only capture the work performed by the data curators. At the same time, there is a lack of data management best practices. Researchers are unsure of what data management entails, and are looking to their peers for guidance. Even the NSF is unwilling to commit to a list of required data management activities. To support the diversity of data practices exhibited between disciplines, each directorate provides different guidance for what is appropriate to include in a data management plan, but they provide no best practices. Best practices may not be what researchers need, but some functional goals would provide a better framework of what data management entails.

There are many visions of what data management looks like from the top-down view of research funders and data curators. But what does data management look like from the bottom-up? This question is formalized into the following three research questions:

1. What tasks are performed by researchers to manage their data?
2. How are data management tasks distributed among members of the research group?
3. For what data management tasks do researchers perceive they are responsible?

In answering the first question, the invisible data management work performed by data producers is made visible. In answering the second question, the data management roles within the data producer category are identified. And in answering the third question, the roles the data producers apply to themselves are uncovered. By uniting what data producers think they should be doing, with what they are doing, and with what the data curators strategize the data producers should be doing some middle ground can be reached between top-down and bottom-up approaches.

The exploratory research reported in this dissertation provides a rich description of data management tasks and perceived data management responsibilities from a selection of six research groups across four universities, and six disciplines, all affiliated with the CENS research community. The research into embedded networked sensing systems performed at this center has allowed higher resolution data collection in both time and space. This resolution has revealed the previously unobservable, and generated a higher volume of data than the majority of researchers were prepared to handle with existing tools and practices. The incorporation of embedded networked sensing technology is viewed by these researchers as the future of their own disciplines and others'. Study of data practices within these groups provides a snapshot of evolving data management practices that other researchers will experience in the near future. The CENS community is also marked by a strong culture of sharing, collaborating on open source code development, sharing research platforms and tools, and well as sharing data.

Research data are tied to the publication in which they are described. Using a publication from each research group as the sampling frame, two to four authors from each publication were selected as participants. A grounded theory approach, relying on mixed methods, was used to construct and test theories about how researchers were managing their data and their perceived data management responsibilities. Document analysis of the research publication and other documents provided by the participants were used to inform participant selection, the interview, and provide data management tasks as reported to their peers. Semi-structured interviews were used to capture a description of the research project, the data reported in the research, the sequence of events, who handled the data when, what data management tasks were performed, and who was perceived to be responsible for what by whom. Field observations from my prior research with the same community were used to corroborate interviews results. Data were coded and analyzed for emergent themes, and used to construct the following models: a generalized data life cycle, a typology of data management tasks, a model of when data management tasks are performed during the data life cycle, and a model of the distribution of data management tasks among members of the research group.

The goal of this dissertation is to make existing data management practices on the part of data producers visible to both data producers and data curators. With visibility, effective dialog between data producers and curators can be encouraged, more specific data management policy can be shaped by research funders, and the construction of data management infrastructure can be informed. Ultimately, the motivation was to encourage data management for data sharing.

CHAPTER 2: REVIEW OF LITERATURE

Scientific research data are becoming first class objects, with value beyond the publications in which they are interpreted. The value of data lies in using them for research beyond that for which they were collected by sharing them amongst the research community. Data-driven research, the so-called 4th paradigm, relies on the availability of research data (Hey, T., Tansley & Tolle, 2009). Data sharing, or the sharing of research data with others for reuse, relies on the long-term management of data so that they are accessible and usable. Data management is a time intensive process, to maintain the integrity of the data and apply contextual information for interpretation. The following sections provide a picture of the diversity of scientific research data; studies of data management practices; studies of science and technology collaborations; stakeholders in the data management effort; research contribution as encoded in publication bylines; life cycle studies of information resources; and prior data practices research from the CENS community that motivated this research.

Diversity of Scientific Research Data

The policy communities promoting data management all too often ignore the contentious nature of the concept, drawing instead upon practical definitions such as this one promulgated by NASA: “A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen” (Reference Model for an Open Archival Information System, 2002, 1-9). Another widely referenced definition is, that, “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, 15).

Data also can be categorized by type or origin. The typology presented in an influential U.S. policy report (CODATA-CENDI Forum on the National Science Board Report on Long-Lived Digital Data Collections, 2005) and incorporated in National Science Foundation strategy (Cyberinfrastructure Vision for 21st Century Discovery, 2007) is now widely accepted. These four categories of data are observational, computational, experimental, and records. 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 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. Records of government, business, and public and private life also yield useful data for scientific, social scientific, and humanistic research.

The same report distinguishes between three levels of data collections: 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 (CODATA-CENDI Forum on the National Science Board Report on 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, a registry of biological macromolecular structures, is the canonical example of this transition (Protein Data Bank, 2006; Berman et al., 2000; Bourne, 2005).

Cragin and Shankar (2006) take the three types of collections above as a point of departure, showing how this simple taxonomy obscures the complexities 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. Case studies of distributed, collaborative projects in science reveal many kinds of data, which mean different things to different participants (Kanfer et al., 2000; Lawrence, 2006; Lee, C. P., 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 that reflect their concerns with water quality. Hydrologists tend to gather data on features such as drainage and erosion, reflecting their concern with the quantity of water (Ribes & Finholt, 2007: 113). Tensions between short and long term concerns about the data are illustrated in a study of the Long Term Ecological Research (LTER) centers (Karasti, Baker & Halkola, 2006). In the short term, participants were concerned about matters such as technology solutions, data volume, and metadata, whereas long-term concerns were driven more by scientific inquiry, data sharing, and stewardship.

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.

Data Management Practices

Data management is a term very much in the public eye as a result of the recent NSF Data management Plan requirement, released in January of 2011. The “data management plan” is described in the NSF guidelines (NSF, 2011b) as a supplement that, “should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results.” The dissemination and sharing of research results policy, section 4 of the Award and Administration guide, addresses the dissemination and sharing of research data. The beginning of section 4.b. (NSF, 2011a) is as follows:

> “Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing.”

Using the description from the NSF guidelines, data management includes all activities that would support the sharing of data. To complicate this very broad concept of data management, data management is also a term of art for business and computer science, which applies to the management of databases, rather than supporting data sharing.

Data curation is a term more specific to the preservation of scientific data than the NSF understanding of data management, and yet more inclusive than the business and computer science definitions. Lord and MacDonald (2003) provided the following working definition of data curation. Data curation is, "the activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use." According to the authors, data curation includes archiving – "a curation activity which ensures that data are properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity," and preservation, "an activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology." Preservation is one of the four core archival functions, including: appraisal and acquisition; arrangement, processing, and description; preservation; and public programming (SAA). The data curation activities described by Lord and MacDonald would support the NSF Data Management requirements. Although data curation would be my preferred term, data management will be used for the purposes of this dissertation because of the current emphasis on this term.

Enabling reuse of scientific data can be of tremendous future value as such data are often expensive to produce or impossible to reproduce. Data associated with specific times and places, such as ecological observations, are irreplaceable. They are valuable to multiple communities of scientists, to students, and to nonscientists such as public policy makers. Research on scientific data practices has concentrated on big science such as physics (Galison, 1997; Traweek, 1992) or on large collaborations in areas such as biodiversity (Bowker, 2000a; b; c). Equally important in understanding scientific data practices is to study small teams that produce observations of long-term, multi-disciplinary, and international value, such as those in the environmental sciences.

The emergence of technology such as wireless sensing systems has contributed to an increase in the volume of data generated by small research teams. Scientists can perform much more comprehensive spatial and temporal in situ sensing of environments than is possible with manual field methods. The "data deluge" resulting from these new forms of instrumentation is among the main drivers of e-Science and cyberinfrastructure (Hey & Trefethen, 2003). Data produced at these rates can be captured and managed only with the use of information technology. If these data can be stored in reusable forms, they can be shared over distributed networks. The variety of practices associated with data management and range of understanding of what constitutes "data," which are well known issues in social studies of science (Bowker, 2005), present practical problems for wireless sensing data management. e-Science initiatives state the requirement for better tools, but say little about what the criteria should be for building them. More understanding is needed about practices, behaviors, and incentives associated with the collection, use, and management of scientific data. These findings are important input to the design of effective digital library systems, services, and policies.

Studies of scientific data practices, per se, are few and far between, especially for the "long-tail" of science research domains. The term "long-tail" is borrowed from Chris Anderson (2004) who applied it to internet commerce, where the potential value found in rarer items that are difficult to stock in brick and mortar stores due to shelf space constraints and low demand can be realized in the scale of the internet. Heidorn (2008) applies the term to scientific data sets, pointing out that there is little demand for the data sets generated from a research project, but there are so many of these data sets that they the aggregate impact they have on science is significant. Long-tail data sets come from long-tail science, which is a more natural complement to big science (Price, 1963) which generates big data (Lynch, 2008).

Investigations of scholarly activities in scientific laboratories (Latour, 1987; Latour & Woolgar, 1979; 1986) offer guidance, as do studies in biology and bioinformatics (Bowker, 2000a; Cragin, Palmer, Carlson & Witt, 2010; Zimmerman, 2003; 2007; 2008). Long-tail 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-sampling of biological variables, sensors take the burden of collecting contextual variables, such as ambient temperature, wind speed and direction, and chemical concentrations in water and/or soil. Similarly, computer scientists and engineers find these technologies of interest for research on robotics, sensors, vision, networks, and fault detection.

The development of repositories for scientific data and the policies of funding agencies to promote the deposit of data in those systems is predicated on the assumption that these data will be reused by others (Wellcome TrustSharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility, 2003; Van de Sompel, Hammond, Neylon & Weibel, 2006). However, data sharing is more common in big science than in long-tail science. Scholars in long-tail science often assume that their data are not of value beyond a specific study or research group. Heads of small labs often have difficulty reconstructing datasets or analyses done by prior lab members, as each person used his or her own methods of data capture and analysis. Local description methods are common in fields such as environmental sciences where data sources vary widely by study (Estrin, Michener & Bonito, 2003; Zimmerman, 2003; 2007; Zimmerman & Nardi, 2006).

The degree of instrumentation of data collection is a factor in data sharing due both to the cost of equipment and to the potential reduction in manual effort to generate data. Sharing expensive equipment is among the main drivers for collaboration. In these cases, collaboration, instrumentation, and data sharing are likely to be correlated (David & Spence, 2003). The relationship between instrumentation and data sharing may be more general, however. A small but detailed study conducted at one research university found that scholars whose data collection and analysis were most automated were the most likely to share raw data and analyses; these also tended to be the larger research groups. When data production was automated but other preparation was labor-intensive, scholars were less likely to share data. Those whose data collection and analysis were the least automated and most labor-intensive were most likely to guard their data. These behaviors held across disciplines; they were not specific to science (Pritchard, Carver & Anand, 2004).

According to Lynch (2008), to have big data does not have to mean sheer volume of data collected, as in high energy physics or astronomy, but data can “be big by being of lasting significance” and he goes on to encourage sound preservation practices to accomplish this end. The scaling-up of data management practices to support sharing as identified in a report by Arzberger, Schroeder et al (2004b) include: “openness; transparency of access and active dissemination; the assignment and assumption of formal responsibilities; interoperability; quality control; operational efficiency and flexibility; respect for private intellectual property and other ethical and legal matters; accountability; and professionalism.” If scientific data are to be leveraged for larger communities, cyberinfrastructure must reflect scientific practices in ways that make documenting and sharing data attractive. These may include mechanisms for personal digital libraries, attribution and provenance support, embargo periods for access, and security (Arzberger et al., 2004a; Borgman, 2004; 2007; Bowker, 2005; Hilgartner & Brandt-Rauf, 1994).

In ecology there have been a number of notes in the literature with data management best practices. Cook et al (2001) set out the basic data management practices in figure 1. These practices all emphasize long-term access to the data on the part of the data producer. Ten years later, Whitlock (2011) focuses more on the benefits of depositing data in a data repository, finding the right repository for your data, attaching appropriate metadata, and not trusting a personal website to act as a long-term data archive. Martín and Ballard (2010) provide a more comprehensive set of data management best practices for bird monitoring and diversity data in their white paper. The outline of data management practices in figure 2 is presented at the beginning of the report and then fleshed out in the body.

- Assign descriptive file names
- Use consistent and stable file formats
- Define the parameters
- Use consistent data organization
- Perform basic quality assurance
- Assign descriptive data set titles
- Provide documentation

Figure 1: Data management best practices (Cook et al., 2001).

- Policy and Administration
    - data policy
    - roles and responsibilities
        - data ownership
        - data custodianship
- Collection and Capture
    - data quality
    - data documentation and organization
        - dataset titles and file names
        - file contents
        - metadata
    - data standards
    - data life-cycle control
        - data specification and modeling (database design)
        - database maintenance
        - data audit
        - data storage and archiving
- Longevity and Use
    - data security
    - data access, data sharing, and dissemination
    - data publishing

Figure 2: The full spectrum of data management activities (Martín & Ballard, 2010).

Despite the assumed value of data sharing for e-Science, scientists have few incentives to sharing their data. Firstly, they are rewarded for publication, not for data management. Secondly, documenting data sufficiently for others to use them requires considerably more time and effort than documenting them only for the use of a small research team. Documenting data for later use also requires much more effort than what is required to publish data summaries in a journal article or conference paper, and the publication alone is not enough to interpret the data. Scientists may be willing to share their data, but only after publication and only with certain conditions (e.g., attribution, non-commercialization). Some researchers are sharing their data, but not through deposit in repositories, they conduct personal interactions with the data user. The interaction allows for the building of trust between both parties, assessment of the validity of the data on the part of the data user, and for the data producer to ensure that their conditions are met (Wallis et al., in progress).

Data Life Cycle Approach

Data management and sharing have some temporal component relating to the scientific process, such as the relationship between when they would be willing to share data and publication. The development of life cycle models has been adopted by archives, where they can be used to plan management and preservation activities to ensure continuity (Higgins, 2008). For instance, an early information resource life cycle model tracks resources through stages which enhance the resource for re-use: generation, institutionalization, maintenance, enhancement, and dissemination (Levitan, 1982). Archival institutions like the National Records Administration use life cycle models to make key decisions about the preservation of materials (Thibodeau, 2007). Life cycle models can also be used to coordinate the actors and activities at the various stages.

In recent years the life cycle approach has been employed in modeling digital curation. Data management overlaps with digital curation, digital data being one of many digital resource types and management being a part of curation. There are also analog data, such as physical water samples, which do not necessarily fit within digital curation practices although, are in the minority, and are not easily shared. Pennock (2007) gives the following justification for the adoption of this approach for understanding digital curation:

> “Approaching digital information management from a life cycle perspective facilitates continuity of service; this in turn supports verification of the provenance of digital data despite technological and organisational changes in their context, and helps to maximise the initial investment made in creating or gathering them” (Pennock, 2007).

Lee, Tibbo et al (2007) define an education program for digital curation around a series of five guiding principles, the second principle of which is, “digital curation activities span the entire life cycle of digital resources.” The authors provide the following brief list of the life cycle stages found in figure 4, without explanation as to what each stage entails. This data life cycle begins before the actual creation of the digital object. Lee, Tibbo et al define digital curation as “much more than preservation of bits,” and indicate that the context and data producers are important to the curation process.

- Pre-Creation Design and Planning
- Creation
- Primary Use Environment (Active Use)
- Transfer to Archives
- Archives (Preservation Environment)
- Transfer Copies or Surrogates to Secondary Use Environment
- Secondary Use Environment

Figure 4: Life cycle of digital resources (Lee, C. A. et al., 2007).

Despite the emphasis placed on the data producers and the context surrounding data creation, essentially the scientist and the scientific process, tend to be treated as a black box, and the data producers are only ever visible in the hand-off of data products to the data curators. This invisibility of the resource producer and the process of production are even more clear in the digital curation life cycle developed by Higgins, as seen in figure 5 below. The life cycle developed by Higgins is intended to be generalized enough to describe the digital curation life cycle of any digital resource, including digital research data. The model is remarkably complex, tying together parallel cycles and cycles that happen repeatedly throughout the life cycle. The full outside ring is comprised of “sequential actions,” beginning with conceptualize at the top, and moving into the ring moving clockwise. Those actions outside the sequential actions are “occasional actions,” and those inside are ongoing parallel processes.

The processes that lead to the creation of a digital resource are positioned, literally, in a black box outside of the digital curation. When applying this model to the curation of data, all the planning, collection, processing, analysis, etc. work performed by the data producer happens in the first phase, "Conceptualize," after which the data are handed off to curators who begin the actual curation process. This model leaves little room for the data producer to be a part of the process.

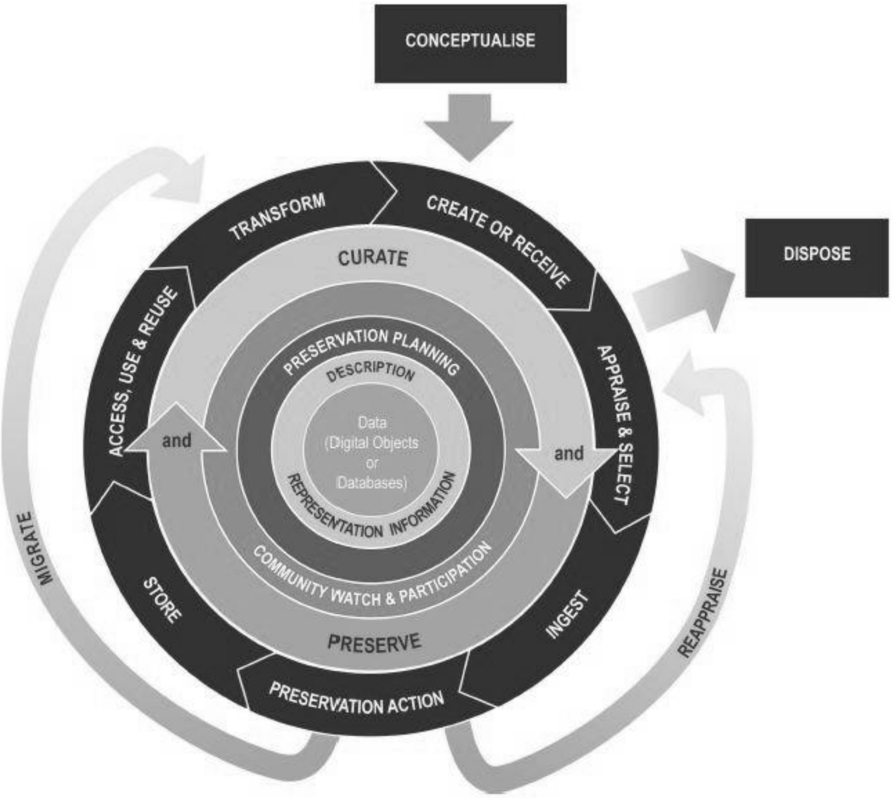
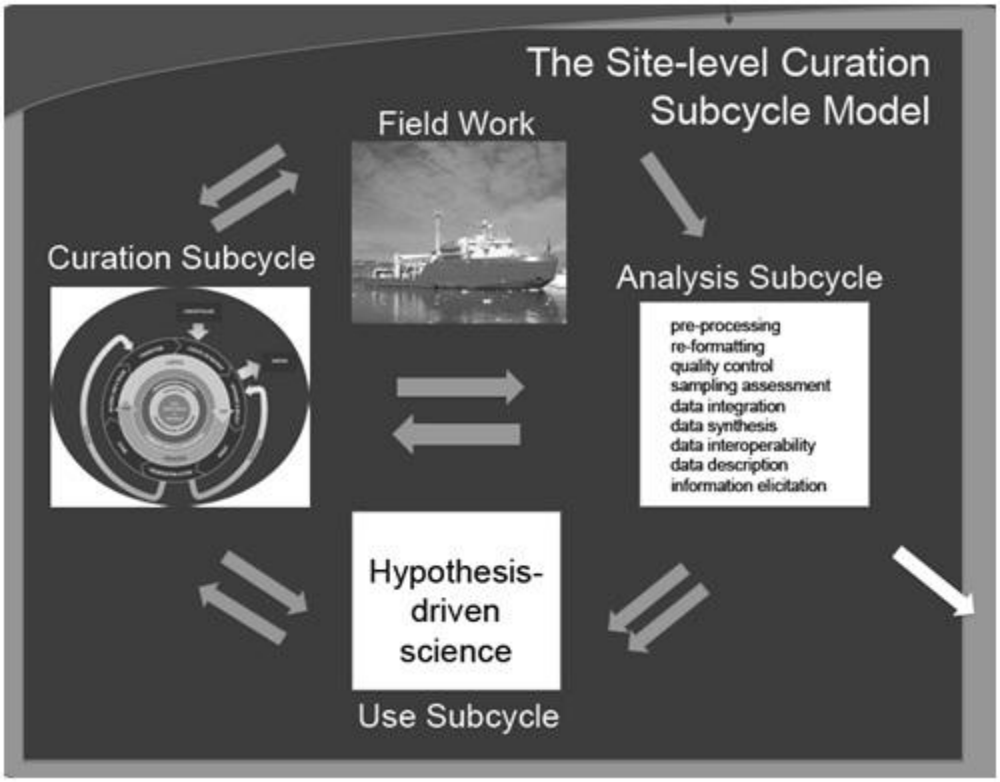


Figure 5: Digital Curation Life cycle (Higgins, 2008).

Baker, Millerand et al (2009) take the Higgins Digital Curation Life cycle and demonstrate how digital curation is applied to marine biology data. As seen in figure 6, the curation life cycle is one sub-cycle in a series of parallel cycles of work performed by the data managers at a Long Term Ecological Research site. Data managers at the LTER are also trained in the discipline they are supporting, and are able to perform the work of researchers as well as data managers. Within the sub-cycle model, the raw data are viewed as an object of value, which enters the curation life cycle just after being collected. The curation sub-cycle is separated from the rest of the research processes within the hypothesis-driven science sub-cycle. Analysis is separated from both management and hypothesis-driven science. According to personal communication with Baker (2011), the model was intended to describe how copies of field work data are distributed out to a variety of processes that are relatively independent. The separate cycles form a network of resources that all trace back to an original dataset. In each sub-cycle different individuals handle the data. From the model, as presented in this poster with minimal documentation, it is difficult to see how, where, and when the value added by each party integrate together. The LTER data management processes are quite specialized, and may not be generalizable beyond the LTER site described in the model.



The Site-level Curation Subcycle Model

Field Work

Curation Subcycle

Analysis Subcycle

pre-processing
re-formatting
quality control
sampling assessment
data integration
data synthesis
data interoperability
data description
information elicitation

Hypothesis-driven science

Use Subcycle

Figure 6: Data Management Sub-cycle Model. An excerpt from "Growing Information Infrastructure: data lifecycles and subcycles" (Baker, Karen et al., 2009).

Prior research into the life cycle of CENS scientific data has focused on the creation and use of data as part of research. Following the data from collection to analysis through processes such as cleaning, integration, and derivation allows the capture of provenance. We (Wallis, Borgman, Mayernik & Pepe, 2008) mapped the processes leading to the creation of usable research data in the ecological sciences at CENS. We identified nine different stages at which scientists make decisions that ultimately affect the data that are used in the final publication, which can be seen in figure 7 below.

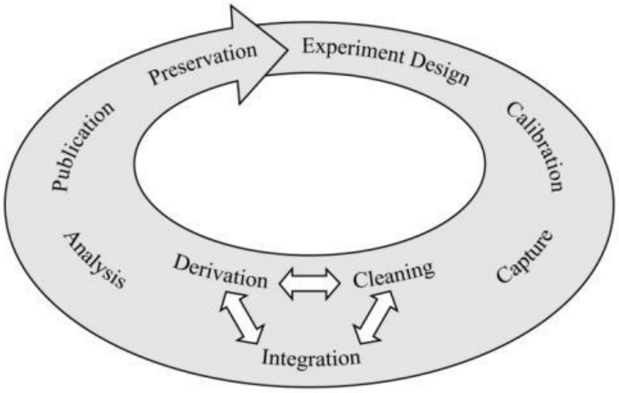


Figure 7: The life cycle of CENS research data (Wallis, Borgman et al., 2008).

In this data life cycle model, we initially mapped preservation activities at the end of the process. Researchers admitted they were not performing any specific preservation activities, beyond making the data usable for their own research. We argued that for data curators to acquire and process the data, they would need to be involved throughout the life cycle to capture the value added by the data producers.

This research was extended to address the intertwined data life cycles of the ecological data and the technology data from multi-disciplinary collaborations at CENS (Wallis, Pepe et al., 2008). We found that the collaborators we studied needed to mesh together their respective practices throughout the life cycle as can be seen in table 2. In the outside columns, the data life cycle stages are instantiated based on single-disciplinary work of the technology and application science researchers respectively. The inside column is the instantiation of an interdisciplinary data life cycle where both groups are working together. In some of the stages the work is combined, for instance everyone collected data for others to use. In the cleaning, integration and derivation stage, the data were handled in series by the collaborators rather than handled in parallel. In some stages, such as calibration/setup, one set of practices displaces the other rather than combined. And in some stages, new practices entirely were required, such as negotiating researchable questions during experiment design or visualizing the data during analysis. Misalignment of practices resulted in research delays. Placing collaborators and tasks on the data life cycle exposed the interdependencies between science and technology researchers (Borgman, Wallis & Mayernik, forthcoming).

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

Table 2: Three instantiations of the data life cycle: science, technology, and science/technology collaboration (Wallis, Pepe et al., 2008).

The life cycle of CENS research data was also extended in Pepe, Mayernik et al (Pepe, Alberto, Mayernik, Borgman & Van de Sompel, 2010), where the authors identified resources produced at different stages of the data life cycle. The paper presents an approach to automatic capture of data provenance, by collecting the resources together in an Open Archives Initiative Object Reuse and Exchange wrapper. When someone gains access to the data, they can then access all of the resources that describe the data at various stages: planning, calibration, and collection; cleaning, processing, and analysis; publication and preservation. The life cycle approach highlighted the timeline of events and products, so that an appropriate infrastructure for capturing the products could be developed.

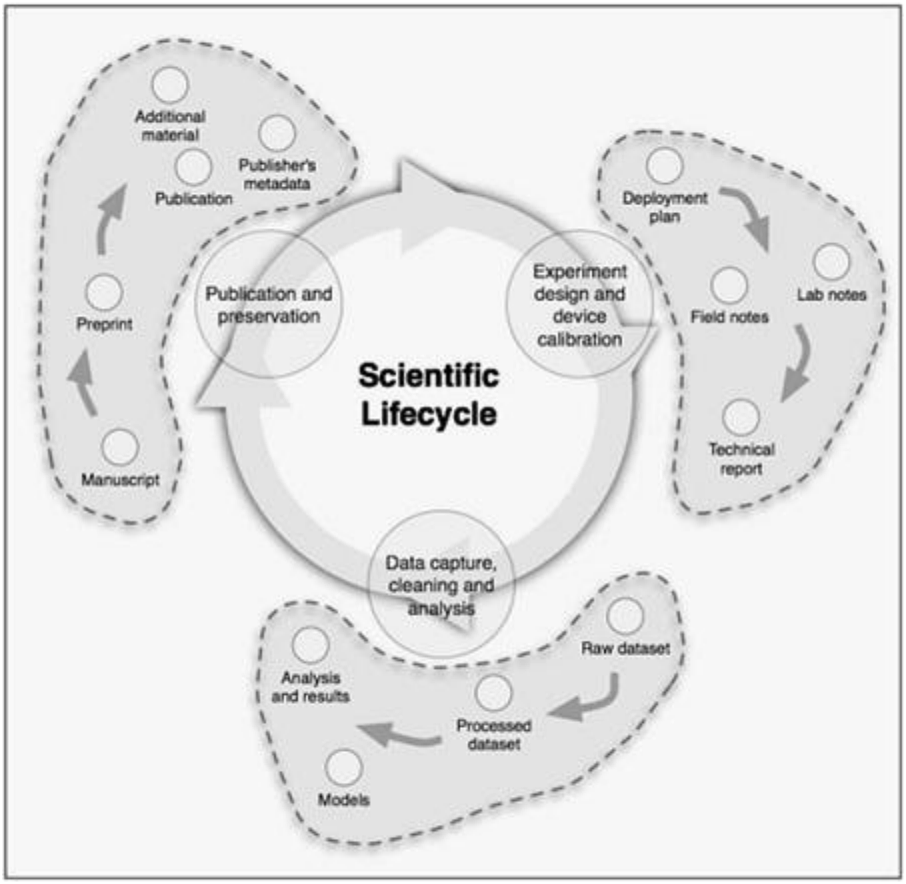


Figure 8: Map of the resources produced throughout the CENS data life cycle that could be aggregated into ORE objects describing the data (Pepe, Alberto et al., 2010).

Data in Science and Technology Collaborations

Study of large collaborations is an area of interest for infrastructure studies (Cummings & Kiesler, 2005; Olson, Zimmerman & Bos, 2008). Infrastructure can be used to negotiate the difficulties of distance and disciplinary differences. One of the keys to a successful collaboration as identified by Olson, et al (2008) is that researchers must share some common ground. Common ground is used to mean: a) previous successful collaboration, b) sharing a common vocabulary or a means of translation, c) and sharing a common working style. These factors allow for the creation of mutual knowledge that brings together collaborators. Borgman (2007) agrees that the process of negotiating shared meaning can be a positive one. Through exchange, one community's implicit assumptions are made explicit to the other. In collaborations between researchers where the collected data forms a shared resource, data both bridge and demarcate the lines between communities, acting as boundary objects (Star & Griesemer, 1989). A focus on data can reveal much about communities and relationships (Borgman et al., forthcoming).

Shrum et al (2007, p. 123-) distinguish between three types of collaborations involving instruments, based on the degree of interdependence and autonomy: (1) Use of standardized and familiar instrumentation, where the science contribution is based on how instruments are deployed; (2) adaptation of extant instrumentation to improve the science; and (3) design and construction of "an unprecedented instrument." While Shrum et al were studying physicists, the analogy to sensor networks applies. Static sensor network deployments fall in the first category, while field deployments in CENS fall into all three categories. The type of instrument-based collaboration also influences the choice, use, and interpretation of data (Collins, 1998; 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 data and phenomena.

Scientists often rely on professional software engineers to construct tools for data collection and analysis. Here the collaboration challenges lay 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.

Lawrence (2006: 393) identified this tension as central to a large cyberinfrastructure research effort: “Officially, the project was a computer science research grant, but at the same time it was expected to develop technologies for use in a disciplinary context,” a situation very similar to that under study in this dissertation. Scientists in the application domain needed technologies for use in their own research, thus they had to define features and architecture sufficiently that the systems would yield the research data they required. Conversely, the computer scientists were engaged in their own research, and desired as much flexibility as possible in pursuing their own questions (Lawrence, 2006).

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, and scientists must rely on technologists to ensure that field excursions are successful (Borgman et al., forthcoming).

Stakeholders in Data Management

Data management has many stakeholders, not the least of which are the researcher/s who produced the data intended for their own research. The last few sections have addressed complications and the possible uses of scientific data that have been managed and shared beyond the contemporary use. In addition to other researchers, funding agencies, institutions, publishers, and other parties have a stake in data management. A number of data management strategy documents have come out of the library and curation world during the last decade, which identify stakeholders and stakeholder responsibilities. These strategy documents are described below.

The Wellcome Trust Report (Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility, 2003) describes a "system of tripartite responsibility in genomic research data" that relies on: resource producers, resource users, and funding agencies. The resource producers are responsible for depositing their data, the resource users are responsible for providing attribution to the resource producers, and the funding agencies hold everyone accountable. A category of stakeholder that is not included in this report is that of the publisher. Publishers play a vital role in encouraging deposit of genomic data in Genbank, the NIH genetic sequence database (http://www.ncbi.nlm.nih.gov/genbank/). The data curators running the repository itself could be considered stakeholders as well, because the success of a repository relies mainly on the deposit of content. In this case, Genbank is run by a funding agency associated with the research community, thus Genbank already falls under the agency stakeholder category.

A report on scientific research data management from UKOLN (Lyon, 2007) lists stakeholders' roles, rights, and responsibilities as derived from interviews with researchers in the UK. In addition to the three stakeholders categories delineated by the Wellcome Trust report, which are called the scientist, user, and funder respectively, the UKOLN report also includes the institution, data centre, and publisher as stakeholders. Data management responsibilities described in the report were not a reflection of current practices per se, but an idealized distribution of responsibilities to support the goal of accessible data in repositories that are interpretable through metadata. In this report, the scientist's role, rights, responsibilities, and relationships are summarized in table 1, an excerpt of a much larger table providing similar information for each of the stakeholders identified. Unlike the Wellcome Trust report, the scientists' responsibilities do not explicitly include deposit of data in a repository.

Role	Rights	Responsibilities	Relationships
Scientist: creation and use of data	Of first use. To be acknowledged. To expect IPR to be honoured. To receive data training and advice.	Manage data for life of project. Meet standards for good practice. Comply with funder / institutional data policies and respect IPR of others. Work up data for use by others.	With institution as employee. With subject community With data centre. With funder of work.

Table 1: Excerpted row from Summary table of Roles, Rights, Responsibilities, and Relationships (Lyon, 2007).

The Long-Lived Digital Data Collections Report outlines four main actors who play important roles in the data collection and management process, as seen in figure 3 below. This report diverges from the Wellcome Trust and UKOLN reports, in that the actors are characterized by the data management activities they perform on data, rather than their actions being defined by their role. Because of this, some individuals can fall in more than one category, such as the scientist who may be both a data author and data manger. In this case the scientist would be responsible both for the production and maintenance of digital data. These actors are arguably also stakeholders, but this list does not include all stakeholders nor describe their stakes in data management.

- *Data authors:* the scientists, educators, students, and others involved in research that produces digital data.
- *Data managers:* the organizations and data scientists responsible for database operation and maintenance.
- *Data scientists:* the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection.
- *Data users:* the larger scientific and education communities, including their representative professional and scientific communities.

Figure 3: Four main actors in data collection and management (CODATA-CENDI Forum on the National Science Board Report on Long-Lived Digital Data Collections, 2005).

For the rest of this research, “data producer” will be used to denote the stakeholders that fall into the resource producer, scientists, and data author categories from the three respective reports. The responsibilities ascribed to data producers are not specific, and leave room for interpretation by the stakeholders. For instance, responsibilities listed for data producers in the UKOLN report are vague in order to accommodate domain variation. At the same time, they are so vague as to require nothing beyond what researchers are most likely already doing for themselves: managing data for their own consumption, meeting standards for good practice, complying with funder data policies where there are data policies, and “working up data for use by others.” Only the third responsibility is at all specific, but could be moot if there were no funder policies regarding data management. The fourth responsibility, as mentioned earlier does not specifically require the data be deposited in data repositories, which is the aim of the UKOLN report. How much “working up” is necessary so they be used by others is an ongoing question.

Author Contribution in Scholarly Communication

The author byline in a publication is a place where researchers identify themselves as responsible for the reported research. In publications with multiple authors, that responsibility for the research is diffused among the authors and may be unequally divided. The criteria used to determine author status and author order vary even within disciplines. When compounded with collaboration the determination of author criteria and author order becomes even more difficult for the reader to gauge research responsibility from the byline.

In 1955, the standard number of authors per publication was one (Rennie, Yank & Emanuel, 1997). When there was only one author, the individual responsible for the research was unambiguous. Zuckerman (1968) describes the typical author at that time, “(a) scientist may have had technical assistants working with him, but he was unambiguously the investigator, not merely the 'principal investigator.'" In a study of the prevalence for team authored publications (Wuchty, Jones & Uzzi, 2007), in the time from 1955 to 2000, the number of authors per publication has dramatically increased. The percentage of team authored publication in the sciences has risen from 50% to over 80%, with an increase of mean team size from 1.8 to 3.5 authors per publication. In the social sciences, the percentage of team-authored publications has risen from under 20% to above 50%, with an increase of mean team size from 1.22 to 1.84 authors per publication. The humanities have seen modest increases, where single-authored publications make up over 90% of the publications indexed by ISI Web of Science. The increase of authors per publication has led to a reduction in visibility of the role of individual investigators, or what Zuckerman (1968) described as a "profound change in the social organization of scientific work." At the far extreme of team authorship are hyperauthored publications, a term introduced by Cronin (2001) and applied mainly in the high energy physics community to publications that list a hundred or more authors in the byline.

As Solomon (2009) points out, contrary to the definition of author found in the Oxford English Dictionary, "(o)ne who sets forth written statements" no longer holds if it ever held for academic publishing in computer science. As described by Birnholtz (2006) the functions fulfilled by authorship include: "(a) attributing credit for discoveries to a person or group of people, (b) assigning ownership to this person or persons, and (c) enabling the accrual of reputation."

In attempting to suggest an authorship policy for computer science, Solomon compares policies from the American Psychological Association (APA), Committee on Publication Ethics (COPE), Elsevier, and the International Committee of Medical Journal Editors (ICMJE) on points of author definition, responsibilities, and order. The definitions of authorship across the policies are quite similar, and all point out that activities beyond writing contribute to the creation of the publication. The author responsibilities set out are also similar across policies, being an author on a publication implies that the author is publicly responsible for the contents of the publication. Where the policies really differ is in the author order. The APA policy suggests that authors are ordered by contribution, with the first author contributing the most, and the last author contributing the least. The APA policy also outlines a way to address equal contributions through the use of a contribution statement. The ICMJE policy does not suggest an order per se, but requests that regardless of the author order the authors include a statement explaining the order and contributions of each author. The COPE and Elsevier policies do not provide an author order (Solomon, 2009).

Author order is used to confer professional reputation and to make hiring and promotion decisions (Savitz, 1999). There appear to be three major strategies for presenting the author order in a journal byline: alphabetical by last name, strict order by contribution, and first/last, which is the same as order by contribution, but with the PI last.

Alphabetical ordering of authors is prevalent in a few fields, and in others has become less prevalent over time. Engers, Gans et al (1999) modeled the behavior surrounding alphabetical author ordering within economics journals. Within the five major economics journals the incidence of ordering alphabetically was just under 90%. In comparison to other fields, the closer disciplines, such as economic history, finance, and law had an 80% incidence rate. The disciplines farther from economics, such as sociology and psychology (40%), chemistry (50%), and medicine (5%), exhibited variable incidence of alphabetical author order (Engers et al., 1999). Alphabetical author order is also common in the hyperauthorship world of high energy physics (Birnholtz, 2006). Between 1996 and 2006, the ratio of publications in computer science with alphabetical author order to those with non-alphabetical went from 1:1 to 1:2. Prior to 1996 the 1:1 ratio held steady for decades (Solomon, 2009). In three biology journals the percentage of publications with alphabetized authors was roughly 30 percent (Laband & Tollison, 2000).

In a study of alphabetical author order and research quality in economics, it was found that first tier journals had over 10% more publications with alphabetical author order than the second tier journals (Joseph, Laband & Patil, 2005). The authors also found that publications with alphabetical author order had a higher rate of citation, but this likely has more to do with the quality of work in the first tier journals in comparison to the second tier journals than the author order. In a similar study within finance literature, the same trend between first and second tier journals held, but that could be attributed not just to the quality of the journal, but also the length of the publication, team size, and the presence of European authors (Brown, Chan & Chen, 2011).

Several articles mention alphabetical ordering as a way of assigning equal credit to the authors (Engers et al., 1999; Tscharntke, Hochberg, Rand, Resh & Krauss, 2007). When there is an order that is not alphabetical, the authors are assumed to be ordered by contribution. By ordering the authors alphabetically, the reader is presented with no contribution information, and must assume that the authors contributed equally. Engers, Gans et al (1999) are concerned that the use of alphabetical author order can be abused at the expense of an author who contributed more than her co-authors. The authors ultimately conclude that the prevalence of alphabetical author order within economic publications is more likely a result of dysfunctional collaborations, where the researchers cannot effectively negotiate the contribution order and rely on alphabetical order as a non-confrontational option rather than an intentional choice to signal equal authorship (Engers et al., 1999).

Unlike economics, finance, and high energy physics, many of the experimental sciences emphasize the contributions of the first and last author positions (Zuckerman, 1968). The first author is the research lead, and widely accepted to be the most significant contributor to the research. The last author is typically the lab head or other senior researcher who oversaw the conduct of the research. The other authors are ranked by their respective contribution. In this ordering the same position may have different meaning. In a paper with four authors the fourth author is a significant contributor, but in a paper with five authors, the fourth author is likely to be the least significant contributor.

In a study of perceived author contribution based on position (Wren et al., 2007), it was found that promotion committees perceptions of the first and middle authors varied depending on the total number of authors. The perception of the last author's contribution remained the same regardless of whether there were three or five authors. These findings were consistent with prior work (Davies, Langley & Speert, 1996; Shapiro, Wenger & Shapiro, 1994). Wren, Kozak et al (2007) found that the only time the perception of the last author's contribution dropped was when the last author was not the corresponding author.

In collaborative work, the first/last author positions can be used to identify two significant collaborators (Ambrosone & Kadlubar, 1997). Savitz (1999) points out that this may provide credit to multiple collaborators, but this signal is cryptic and may just as easily be misunderstood. It is important to note that Ambrosone and Kadlubar are referring to collaboration between researchers and clinicians in the same field where both collaborators would receive benefit from publishing.

In a study of how age and professional rank influence author order, Costas and Bordons (2011) analyzed author order in over a thousand publications in domains of biology and biomedicine, materials science, and natural resources. Within these fields the authors found that the first author position was typically occupied by young researchers or individuals with lower professional rank, while the last author position is occupied by veteran researchers or individuals of higher rank. There was some variation between the fields, largely due to the size of collaborations observed in each field. Where there is a higher degree of collaboration, such as in biology and biomedicine, professional rank is more significant than age in determining who occupies the first and last positions.

Although first/last is the dominant authorship order in the sciences, this approach is not without faults. Some fear that the standard use of the first/last author order will negatively impact the more senior authors (Buehring, Buehring & Gerard, 2007). The increase in multi-authored papers has induced publishers to limit the number of authors listed in a reference citation. These limits tend to be five or six authors before an “et al” is used in place of the other authors. For papers with more than five or six authors, the senior authors become invisible in the references. The authors recommend the use of order by contribution to ensure that the authors who actually contributed significantly are included before the five or six author cut-off. Riesenberg and Lundberg (1990) call for a strict ordering by contribution in biomedical literature, claiming that positioning the lab head as the last author encourages gift authorship. Gift authorship is the inclusion of an individual who does not fulfill the requirement for authorship for purposes of compensation or to increase the credibility of the research by adding a trusted name (Bennett & Taylor, 2003).

There is very little literature about the strict order by contribution, also known as the "sequence-determines-credit" approach (Tscharntke et al., 2007). This approach was recommended for both biomedical and environmental publications as a clearer method of indicating author contribution than other possible orders (Rennie et al., 1997; Tscharntke et al., 2007). As mentioned earlier, the APA guidelines for author order recommend using this order for psychological publications. A drawback of this approach is the entrenched first/last author emphasis.

The wide spread use of first/last would make it difficult for lead researchers to give up the emphasis placed on the last author position, as well as making it difficult for those trying to evaluate the work to dismiss the last author as a position of authority. First, second, and last positions have meaning to researchers and evaluators, which is why these are the author positions for which researchers bargain (Burman, 1982). According to Savitz (1999), the reader must use subtle clues, such as name recognition, in order to determine whether the last author had a significant contribution to the overall concept of the research or whether they just contributed the least.

In a study comparing economics literature which relies largely on alphabetical listing of authors and agricultural economics literature where authors are ordered by contribution, it was found that the emphasis on priority in agricultural economics was likely a result of applying for external funding. Within economics, funding comes from the institution itself, rather than from external funding sources (Laband, 2002). A later study of the agricultural economics promotion and review committee attitudes indicated that the variation in credit awarded by position in a multi-authored publication was much less significant than the difference in credit awarded to an author of a sole-authored publication when compared to a multiauthored publication (Hilmer & Hilmer, 2005).

In a study of why researchers collaborate and how author order is determined, Floyd, Schroeder et al (1994) found that the meaning of "contribution" differed among collaborators and encourages readers or promotion and tenure committees to be wary of taking author order as a face-value indication of contribution. Rennie, Yank et al (1997) suggest the byline include contributors rather than authors, as the term author itself may be misleading. Multiple proposals of how to clarify author contribution for the benefit of both the authors and the readers have been made, such as grading contribution (Pichini, Pulido & García-Algar, 2005) or including a statement of contribution (Buehring et al., 2007; Rennie et al., 1997; Savitz, 1999), as is now required for some medical publications.

After an egregious breach of ethical conduct in the biomedical literature, medical journal editors formed a committee to reduce the possibility of this type of breach happening in the future. Bennett and Taylor (2003) have since developed a typology of authorship misconduct identified in biomedical literature. The five types include: gift authorship, pressured authorship, ghost authorship, fragmentation, and duplication. The first three of these provided the impetus for the development of the ICMJE policy on authorship. Gift authorship was introduced earlier. Pressured authorship is when an individual uses their authority to be included as an author even if they did not fulfill the authorship criteria. Ghost authorship is when significant contributors who should be included as authors are left off the author list (Bennett & Taylor, 2003).

Initially, the ICMJE developed a clear set of criteria that determined authorship. Maintaining a strict authorship definition was intended to reduce frivolous publication, honorary authorship, and ensure that authors are publishing responsibly. Studies repeatedly demonstrated that both students and medical researchers were not aware of the criteria, and the policy had done little to reduce authorship misconduct (Bhopal et al., 1997; Hoen, Walvoort & Overbeke, 1998). The ICMJE later revised the policy to require the addition of a statement to the publication indicating what each author contributed and how each author has fulfilled the ICMJE authorship criteria (Bennett & Taylor, 2003).

There are many studies that utilize these disclosure statements as research data to capture the factors determining authorship and how author order is constructed (Baerlocher, Newton, Gautam, Tomlinson & Detsky, 2007; Rennie, Flanagin & Yank, 2000; Yank & Rennie, 1999). The utility of the contribution statement is still being determined, but there have been proposals suggesting how to clarify the contribution statement. In a study of author order in medical journals (Baerlocher et al., 2007), author contributions were classified into types for a more standard signifier in the contribution statement. The types of contribution identified are primary, contributing, and senior or supervisory. On any given paper multiple authors can be placed in each category.

Biomedicine is not the only domain looking to adopt disclosure of author contributions. There has been a flurry of notes published in the environmental sciences recommending adoption of author disclosure statements. Weltzin, Belote et al (2006) suggests that authors be called contributors as per Rennie, Yank et al (1997) and a statement of contribution be included in each publication. A correspondence in PLoS Biology explored the various ways to interpret author order for the assignation of credit within the environmental sciences (Tscharntke et al., 2007). The authors conclude that regardless of the author order chosen a statement of disclosure would clarify interpretation of contribution.

Studies of Data Practices at CENS

The research reported in this dissertation fits into a much larger data practices and cyberinfrastructure research agenda conducted in CENS. The research agenda includes studies of scholarly collaboration in CENS (Edwards, Mayernik, Batcheller, Bowker & Borgman, 2011; Mayernik, Wallis, Pepe & Borgman, 2008; Pepe, Borgman, Wallis & Mayernik, 2007; Wallis, Borgman et al., 2008; Wallis, Pepe et al., 2008) and the development of information architecture to support CENS' data practices (Borgman, Wallis & Enyedy, 2006; 2007; Borgman, Wallis, Mayernik & Pepe, 2007; Pepe, Alberto, Mayernik, Borgman & Van de Sompel, 2009, in review; Pepe, Alberto et al., 2010). Prior research at CENS includes two rounds of interviews, field observation, network analysis, document analysis, and user testing. During both rounds of interviews, the participants (faculty, research staff, and students) were asked who was responsible for data to capture a baseline understanding of researcher roles. We had naively expected them to identify who was author, owner, or responsible for data, but analysis of the answers to these questions demonstrated that this was not the case. Answers were often hesitant and/or ambiguous. Collaborating researchers provided contradictory answers. Answers did not align with current theoretical understanding of stakeholders and their responsibilities as described by data curator policies as seen in the prior section. The terms "data producer" (Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility, 2003) and "data author" (CODATA-CENDI Forum on the National Science Board Report on Long-Lived Digital Data Collections, 2005) were not how the researchers described themselves, and using these terms may alienate the very people they are intended to describe.

Terminology is one of the most compelling problems when asking researchers about responsibility for data curation (Wallis & Borgman, 2012). Authorship, ownership, and responsibility all evoke complex connotations that complicate the ability to get at the heart of the issue of who is responsible for data. The researchers at CENS appear to have a conception of what it means to be an author, and being a “data author” does not fit within this concept. For many of the participants the problem was one of creativity, that “author” denoted an act of creation, which goes against the idea that the data are measurements of nature, and inherently true. Because participants were distracted by the creative connotations they were unable to see the “authority” conferred on the author. As author, they would have authority to take action with the data, such as releasing the data to the public. Authority would be conferred on these researchers because they designed the data collection effort and performed the data collection tasks, and thus be the authority on why and how the data were collected. Rarely were we able to get the participants own terms for these concepts because the concepts did not exist, as with authorship, or had multiple interpretations, as with responsibility.

From the answers to the ownership and responsibility questions we compiled a list of tasks for which a researcher might be responsible and a list of rights as data creators (Wallis & Borgman, 2012). Participants identified that the data owner would have the following rights: i) the right to publish their data online, ii) the right to publish from their data without contest, iii) recognition as the authority on their data, iv) the authority to answer any questions about the data, and v) the authority to document their data. Participants identified that the data owner would have the following responsibilities: i) ensuring data quality, ii) ensuring documentation quality, iii) protection of sensitive data, iv) provision of data access, v) long-term maintenance of data, and v) support for data reuse. There was some overlap between the above rights and responsibilities and those rights and responsibilities recommended by the UKOLN report (Lyon, 2007), where rights such as “first use” are balanced with responsibilities such as “managing data for the life of the project”, but the overlap was not complete. To be readily adopted by researchers, any recommendations would need to better match the practices and methods of the researchers with regards to data management. The question posed by this research is: what do researchers consider their roles, rights, and responsibilities for data management, and who should be held accountable?

Summary of Key Points

The following are a summary of the key points to take away from the review of the relevant literature. Research data are extremely diverse, varying by purpose, method of collection, and level of processing. Data management is the active management of data to ensure persistence and accessibility of data for the reuse. Long-tail research data can be made valuable through reuse, which requires proper data management. Data practices of data producers vary by discipline, and are not well studied; data management practices even less so.

Data from collaborative research is handled by individuals who come from different cultures of data practice. Multiple groups have a stake in the success of data management: data producers, data users, research funders, data curators, among others. Data management responsibilities have been recommended for each stakeholder group from the top-down. Responsibility for research varies between contributors and is encoded in the author byline of publications.

From prior research at CENS, we have seen that it is difficult to identify the individual “responsible” for data, and that the rights and responsibilities as identified by the data producers differ from top-down recommendations. The life cycle approach has been used to map the data life cycle and the digital curation life cycle, and is a way to model the interaction of parties, processes, and the life cycle object throughout the life cycle stages. Data management and responsibility in the sciences is fertile ground for research. Recently enacted policies, such as the National Science Foundation’s recent data management plan requirement for funding proposals, make the research timely.

CHAPTER 3: PROBLEM STATEMENT

There are vast numbers of data resources being produced every day by long-tail researchers. To take advantage of these data they need to be properly managed to support future reuse. Data curators are willing to jump into the fray and assist researchers in their data management, but in order to facilitate better data management practices, we must first understand existing data management practices. As covered in the previous chapter, there are already visions of what data management looks like from the top-down view of research funders and data curators. The question that remains is: what does data management look like from the bottom-up? This question was formalized into the following 3 research questions:

1. What tasks are performed by researchers to manage their data?
2. How are data management tasks distributed among members of the research group?
3. For what data management tasks do researchers perceive they are responsible?

The following questions must be answered to understand the distribution of responsibility within scientific research collaborations and the factors that affect distribution.

Mapping Data Management Tasks to the Data Life Cycle

Prior research in mapping data life cycles has focused on creating generalized models of how data resources are handled at different stages of research. Digital curation life cycles have been developed to model how digital resources are handled during the stages of curation. While the life cycle models can both be applied to research data, the end of each is different. Researchers need to publish their research, and data are viewed as by-products, whereas curators value the data as a potential resource for future use. The intersection of these models has yet to be understood, and the researchers themselves perform some data management activities to support their own use of the data. The first research question therefore addressed not only what data management tasks were performed, but when were they performed.

R1 What tasks are performed by researchers to manage their data?

R1.1 What stages of the data life cycle are observed, and what purpose does each fulfill?

R1.2 What is the disposition of data at each stage of the data life cycle?

R1.3 What data management tasks are observed?

R1.4 What data management tasks are associated with each stage?

R1.5 For whom are the data managed?

H1 Researchers manage data for their own use. Data management tasks will be similar to core archival functions. Data management tasks will occur throughout the course of the data life cycle, attached to the activities that define each stage.

Data Management Task Distribution

Within a research group, data are handled by different people throughout the data life cycle. Members of the research group have different relationships with the data, and different relationships with one another. The next research question addresses who performed the data management tasks and how they came to be given the tasks.

R2 How are data management tasks distributed among members of the research group?

R2.1 Who performed which data management tasks?

R2.2 How were roles distributed?

H2 All the researchers will perform at least some data management tasks. People beyond the author group will be accountable for some tasks. Researcher status and other demographic factors will affect data management task distribution. Division of labor will not be communicated explicitly between researchers, but will instead be assumed.

Responsibility for Data

In prior research, CENS researchers often said that they were responsible for data, but had inconsistent answers as to what that responsibility entailed. R1 and R2 were meant to capture the actual data management practices of researchers. R3 was meant to capture the perceived data management responsibilities of researchers.

R3 For what data management tasks do researchers perceive they are responsible?

R3.1 Who is responsible for data and what does responsibility entail?

R3.2 Were the responsibilities fulfilled and to what standard?

R3.3 Who holds whom responsible?

H3 Researchers will have some notion of who should be responsible for data, but will vary in who is responsible and what responsibility entails. The researchers will not have a high opinion of how well their responsibilities were fulfilled. Researchers will hold themselves responsible for data, rather than being held responsible by some other person.

CHAPTER 4: METHODS

The exploratory research reported in this dissertation relied on qualitative methods to capture data management practices. In order to construct rich descriptions of the data management processes performed by researchers and how they were distributed within a group the investigation was framed using a single research product – a research publication. A publication describes: a specific research project; a specific set of data collected, processed, and analyzed to answer a specific research question; and was performed by a specific group of people who are clearly identified in the byline and acknowledgements. The study relied on mixed methods: semi-structured interviews to collect descriptions of practices from science and technology researchers, document analysis of publications that provided formal descriptions of data used by the researchers and the processes to which the data were subjected, and field observation of the researchers as they collected and processed their data. The resulting qualitative data was used to construct an ethnographic description of the practices surrounding the research reported in each of six research publications. The case descriptions were interrogated using grounded theory, beginning with the hypotheses stated in chapter 3 as well as new hypotheses that emerged during my data collection and analysis. The six cases were selected from research performed at the Center for Embedded Networked Sensing, a highly collaborative research community that is already facing the data deluge, because of the use of embedded networked sensing platforms being developed at the center.

Research Site

I have spent the past eight years documenting and facilitating the data practices of a distributed, interdisciplinary research center, the Center for Embedded Networked Sensing (CENS). CENS is a National Science Foundation (NSF) Science and Technology Center based at UCLA, with four other partner institutions in California (University of Southern California, California Institute of Technology, University of California Merced, and University of California Riverside) that began in 2002 and will end in 2012 (Center for Embedded Networked Sensing, 2009).

The goal of the research center was to develop and implement innovative embedded networked sensing systems by bringing together technology researchers and the intended future users of the systems, the so-called application disciplines. Research at the center has been organized by the technology research areas, such as multi-scaled actuated sensing, sensors, and programming, and by application science area. About 80% of CENS collaborators (which until 2012 included roughly 300 faculty, students, staff, and post-doctoral researchers) are in computer science or engineering. This group cuts 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 or remote sensing via satellite. The remaining 20% of CENS collaborators are from application science disciplines, such as environmental sciences, biology, seismology, and geology. There are also a few collaborators from the arts, architecture, or medicine, who are principally concerned with cell-phone-based mobile applications. The research performed by the application discipline members has largely included performing investigations using the embedded networked sensors being developed by the technology researchers.

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 utility. 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. Although the initial framework for CENS was based on autonomous networks, early scientific results revealed the difficulty of specifying field requirements far enough in advance to operate systems remotely. Most CENS' research is now based on dynamic "human in the loop" deployments where investigators can adjust monitoring conditions in real-time (Mayernik, Wallis & Borgman, in review). CENS employs an array of sensor technologies, many of which feed data to one another to perform computation in the network, allowing for real-time feedback.

CENS has been an extremely generative research center, with nearly one thousand publications, thousands of research posters, and other proofs of research, such as code and technology made available. The center fosters a culture of sharing, not only of equipment and code, but also data and expertise. The sharing reaches beyond those at CENS. In a recent study of data sharing and data reuse at CENS, we found that nearly half of the researchers interviewed had shared their research data with researchers beyond the CENS community (Wallis et al., in progress). Many outside researchers do not have the luxury of being able to share their research data. Data hoarding may be reasonable if a researcher has very little data. However, when the researchers have more data than they can possibly handle, which has been remarked by a number of participants, then sharing is less of a career risk. Data sharing is predicated on researchers' willingness and ability to share data, and within the CENS community these predicates are fulfilled.

CENS the center is currently winding down, as NSF Science and Technology Center (STC) grants have a built-in expiration. After ten years, or five if not renewed, the NSF would no longer fund the center. STC's have the option of continuing under outside funding, but they cannot apply to the NSF for funding at the center-level. For the majority of CENS researchers, the research that happens within CENS is only one of many projects in their portfolios. The research center infrastructure, including the support staff and weekly activities, will end with the funding, but the current research trajectory will continue. During the last four years work at the center has increasingly focused on the use of cellular phones as a sensing platform. What began as a one-off project has expanded to roughly half of the research performed at the center. This research area has been so generative, that when CENS ends, a significant portion of the students and faculty will have transitioned into a new center where cellular phones are the driving technology.

The research reported in this dissertation falls within the last two years of CENS's funding. The benefits of researching data management during the center wind-down is three-fold: a) the research performed by CENS research groups is well established at this point, allowing for reflection on past research reported in publications, b) the members of CENS themselves are thinking about persistence of the research products as the CENS infrastructure is deteriorating , and c) the data practices research group has been part of the CENS community for the duration, and has captured a significant understanding of the community dynamics, research, and data on which to base theories. To compound these benefits, after years of attending CENS functions, going into the field, and carrying heavy equipment for researchers, I have established a rapport with members of the CENS community allowing for greater access.

Selecting researchers only from within the CENS community has limitations, and will ultimately reduce the broad applicability of the findings. Technology researchers at CENS spend more time making usable equipment and systems than their peers outside the center, who focus on developing proofs-of-concept. Application scientists at CENS spend more time testing out new equipment than their peers who rely on established methods. Thus, the research being performed by this community is atypical when compared to those from their home disciplines. That said, many application science researchers also claimed that the use of embedded networked sensing systems will be the future of their field. The technology researchers alike saw the future of their fields as involving more close collaboration with application scientists. If their predictions hold, the currently atypical researchers from CENS will become the norm over the next decade.

Data Collection Procedures

Ethnography is a way to "report on social life that focuses on detailed and accurate description rather than explanation" (Babbie, 2007). The accuracy of description relies on reporting reality in the terms of the people being studied, rather than the way the ethnographer understands them. In order to be able to attain this level of understanding the ethnographer needs to become an insider.

Participant observation is a data collection strategy (Glaser & Strauss, 1967) where multiple forms of data collection are brought together to allow the researcher to triangulate. For this dissertation, I have chosen to use field observation, interviews, and document analysis to form the points of my triangle. Field observation is a way to access the actual practices. Field observation is experientially rich and as a result the most difficult to capture in ways that allow the researcher to code. The interview is a means of capturing an individual's point of view, their terminology, and what they consider important. But this method is also limited, because many practices are implicit, and cannot be made explicit unless confronted by the reality of performance. The field observation and interview are a powerful combination, making it possible to experience firsthand the phenomena being described secondhand during interviews. The final point of triangulation, is the analysis of documents created during practices, known as document analysis. Documents provide a record of what was deemed necessary to record, as well as terminology and organization structures, all of which inform the understanding of practices.

Grounded theory is a way to "generate a theory from the constant comparing of unfolding observations" (Babbie, 2007). Hypotheses are generated from the data and then tested by interrogating the data at multiple levels of abstraction (Glaser & Strauss, 1967). To maintain integrity of the grounded theory approach the researcher must make comparisons, ask questions, and sample from the data. Another important part of grounded theory is the emphasis on taking notes, or memoing. Memoing is a practice where the researcher is constantly taking notes, about code, theories, and operations or procedures. Memos allow for the constant comparison as well as the ability to step back and check data against interpretations.

To answer the research questions presented in the prior chapter, I utilized these three methods to construct ethnographic descriptions of: the research performed by a research group, the reported data, the data management tasks performed by members of the research group, and their perceptions of data responsibility. Demographic variables were collected for each participant: researcher status, discipline, and gender. Researcher status was collected, as the PI, graduate students, research and technical staff were likely to have different responsibilities. The researcher's discipline was collected, as disciplines of practice tend to have their own cultural norms. The gender of each researcher was collected, because the distribution of responsibility is a form of delegation, a power relationship where gender may affect the outcome.

None of these methods capture what researchers actually should be doing in the way of data management activities. Participants may be doing X or perceive they ought to be doing X, but this does not necessarily mean that they ought to do X. As a result, developing a proscriptive set of “best practices” for data management responsibility distribution for researchers was not a feasible outcome. In order to evaluate the efficacy of certain data management practices, the researchers would need to be studied on a much longer time-scale. The strength of this work rests in the description of the actual and perceived distribution of data management responsibilities. Thick descriptions of these phenomena provide a rich understanding of data management practices within multi-disciplinary collaborations. The resulting models will be useful for informing policy and infrastructure development, and to make the data management practices of the data producers visible to data curators.

Because I performed ethnographic research at UCLA, my research design and materials were submitted for approval by the Office for the Protection of Human Subjects Institutional Review Board. The research I performed posed very little risk to the participants. Participation was entirely voluntary, and participants were able to opt out of research at any time if they felt it necessary.

Sampling Strategy

For the purposes of this dissertation, the individual from a research group was the unit of analysis. Academic research groups are typically comprised of a PI, their students, and any associated research or technical staff that make up their “lab.” But the collaborative nature of the work at CENS crosses multiple PI’s labs, and research groups come together around a research project, such as the development of new software tools, or testing equipment. By studying the individuals that comprise the group, a rich account of individual and group practices can be constructed, using multiple points of view of the same set of events. The alignment and dissent between participant accounts informed the interpretation of results.

A significant distinction is made between the technology researchers and application scientists at CENS. The purpose of their involvement is quite different, developing technology and using new technology, respectively. In prior studies of this community, practices vary significantly between these groups as well. To leverage this variation, both the groups that published technology findings and the groups that published science findings were desired. Not all CENS research falls cleanly in one category or the other, for instance some science researchers publish technology findings, and the inverse. Groups that published some mixture of science and technology findings were also desired, to fully cover the sub-groups present at CENS.

Every year, descriptions of research projects being performed by groups at CENS were compiled into an annual report to the NSF and other funders. The annual report is a different genre of document from the other academic publications, because it is written to highlight the productivity of the interdisciplinary collaboration for the research funders, as this was one of the reasons the center has received funding. CENS research publications, on the other hand, are limited to single-disciplinary reporting because that is what most academic journals support. Very few journals, such as Nature and Science, will accept truly collaborator reporting of results. The few collaborative publications that were accepted in these journals were high-level overviews of the potential applications of the technologies being developed at CENS.

Using the CENS 2011 Annual Report, I was able to identify all of the current CENS research groups, in order to make my sample selection. Potential groups needed to fulfill these criteria: the projects need to be ongoing or recently concluded so that the research experience would be fresh in the participants' minds, excluding research projects in which I participated (those projects within the DATA category), and excluding "participatory sensing" projects. The participatory sensing research projects were not of interest because they tended to stray beyond the borders of technology and science research. Encouraging the reuse of scientific data is the ultimate goal of this research, and as such the participatory sensing groups did not fit within the scope of the dissertation and were discarded from the pool.

The CENS 2011 Annual Report differs from prior annual reports in a way that impacted sampling. In each report a significant portion was dedicated to reporting on CENS research, broken out by research area. In 2011, the report does not have a specific section for technology research, as can seen in the excerpt from the document contents in table 3 below. For contrast, an expert from the 2010 Annual Report table of contents is also included. All of the 2011 subsections are represented in the 2010 table of contents, as well as technology research areas, such as programming and platforms (2.9), multiscale actuated sensing (2.10), and embeddable sensors (2.11). The only difference between the 2010 report and that of 2009 is a shift in terminology, urban sensing became participatory sensing. Reports have contained minor changes from year to year, such as the naming of a research area, or either the addition or subtraction of one area. But the shift from 2010 to 2011 was significant and indicative of the overall shift in the research performed in the center. CENS is still performing technology research, but the descriptions of that research have been subsumed into the applications for which they are being developed.

2011	2 Research	21
	2.1 Research Goals and Objectives	23
	2.2 Participatory Sensing (PART)	25
	2.3 Terrestrial Ecology Observing Systems (TEOS)	73
	2.4 Contaminant Transport and Management (CONTAM)	91
	2.5 Aquatic Microbial Observing Systems (AQU)	103
	2.6 Seismic Applications (SEI)	117
	2.7 Research Testbeds and Data Practices (DATA)…	127
2010	2 Research	25
	2.1 Research Goals and Objectives	27
	2.2 Research Testbeds	28
	2.3 Terrestrial Ecology Observing Systems (TEOS)	30
	2.4 Contaminant Transport and Management (CONTAM)	55
	2.5 Aquatic Microbial Observing Systems (AQU)	75
	2.6 Seismic Applications (SEI)	97
	2.7 Participatory Sensing (PART)	118
	2.8 Statistics and data practices (SDP)	163
	2.9 Programming and Platforms (PRO)	168
	2.10 Multiscale Actuated Sensing (MAS)	190
	2.11 Embeddable Sensors (SEN)	233

Table 3: Excerpts from the table of contents in the 2011 and 2010 CENS Annual Reports.

From sections 2.3 through 2.6 of the CENS 2011 Annual Report, I generated a list of 17 recent research projects. I then classified the projects by the type of research they were likely to report, based on the disciplines of the people listed as part of the research group performing the project. The prior two annual reports were used to verify the classification, in addition to my prior knowledge of the CENS community. Seven of the projects were application science projects which would be of interest to their domain colleagues. Five of the projects were technology research projects which would be of interest to their domain colleagues. And the remaining five projects were a mixture of science and technology that would likely be of interest to colleagues from both science and technology domains. From the 17 total projects, seven projects were selected: three application science, three technology research, and one group that was both science and technology.

A research project can be an unbounded project, ranging from a 2-day deployment and quick write-up to a decade-long deployment that has involved many generations of graduate students participate and provided many publications. A publication is a more bounded frame to view data, the research group, and their practices. Within the CENS community there was no set relationship between the data and the publication. A single dataset could be used for multiple publications and a single publication could pull from multiple datasets. The publication effectively binds the data reported therein, because there is some research question being answered using some set of data as evidence. Just as the data is effectively bound by the publication, the members of the research groups are explicitly indicated in the publication byline and in the acknowledgements.

For each research project a publication was selected. The 2011 Annual Report provided a list of publications, but the list was not current as it had been prepared nearly a year before the selection process occurred. Instead, recent publications from each of the groups were identified based on the research project members. Google Scholar, Microsoft Academic Search, and faculty webpages were used to identify the most recent publications from individual members of the selected project groups. From this batch, publications were selected when the author list contained similar members to the research group in the Annual Report description. By matching similar author I was able to identify research that was considered within the purview of CENS, as it had been included in the Annual Report. Of the one to two publications left from each project, one was chosen from each to act as a seed for contacting authors.

From prior knowledge of the community and understanding the relationships between individual group members I was able to identify the lead author given the research being presented in the publication. The lead author from each of the seven publications was solicited for an interview. The solicitation provided the participant the ability to suggest a different publication if they were so inclined, provided it was still recent CENS research. The solicitation script can be found in the appendices. Of the seven lead authors solicited, only six responded. Of the six remaining, four chose to use the publication provided in the solicitation. For the other two cases, publications in revision were supplied by the authors.

The publications chosen by the authors shifted the distribution of research projects per science, technology, and mixed categories. After the publication adjustments, there were two technology research projects, two application science research projects, and two research projects that were a mix of science and technology. The Fungi and Hypoxia Cases are application science groups reporting on application science findings that were made possible by CENS technology. The Power and Glider Cases are technology groups reporting technology developments, where the former is application independent and the latter has incorporated application science to increase the accuracy of their algorithm. The mixed cases are: the Webcam Case, a purely methods paper written by an application science group testing the viability of technology for measuring application science phenomena; and the Stream Case, a group whose discipline straddles technology and application science, and their methods paper presents primarily a demonstration of the technology as well as some new scientific findings.

Authors from each publication were selected for solicitation based on their role in the research and from recommendations by authors who were already interviewed. As mentioned, the lead author from each publication was solicited. The led author tended to be the one who came up with the idea for the research, had designed the study, lead the data collection, performed all the data processing and analysis, and wrote up the results. This person would have handled the data at different points in different forms, and would have intimate knowledge of where it was stored and the sort of documentation it had been given. Their perspective of the data management practices would be invaluable. From my knowledge of the community I was also able to identify the PI with which the lead author was associated. The PI of the group may not have handled the data, but they witnessed the project from data collection to publication as a supervisor. The PI tends to have a higher-level perspective of the work in comparison to others from the discipline at large. In prior interviews and field observation, the PIs have acted as gatekeepers, introducing me to the other members of their labs and assuring them that social scientists do not bite. For these reasons the PI from each publication was then solicited. These two authors were then asked who of the other authors would contribute to an understanding of the data practices that occurred during the research reported in the publication.

At the beginning of the interviews, the goal was to interview three authors per publication. My rationale was as follows: one author would not provide enough texture, especially in a one to two hour interview, two authors would provide multiple points of view and fill in blanks missed with just one author, three authors would allow me to triangulate between the accounts, and four authors would just become redundant. For each publication at least three of the co-authors were solicited, but not every person solicited responded or was available for an interview. Of the twenty authors solicited, sixteen participated.

After interviewing two, three, and four co-authors for different publications, the experiences were evaluated. Interviewing four co-authors did not contribute more findings than did three. Capturing three co-authors was ideal, providing a very rich picture of the data management practices, although difficult to achieve. Two co-authors were found to provide sufficient detail, so long as one of them was the lead author. In two of the cases, the Glider and Hypoxia Cases, the author lists overlapped. These groups collaborate regularly, with one group being technology researchers and the other group their application science partners. The PIs from both groups are listed as authors on both papers in honor of the collaboration. Both PIs were interviewed, but only about the paper from their discipline. In the Stream Case all four of the authors were interviewed, and in the Hypoxia and Webcam Cases, three of the co-authors were interviewed. In the other three cases, two of the co-authors were interviewed, as can be seen in table 4 below.

Case	Authors	Solicited	Participants
Fungi	3	3	2
Hypoxia	6*	3	3
Power	6	3	2
Glider	6*	4	2
Stream	4	4	4
Webcam	5	3	3
Total	28	20	16

Table 4: The number of authors, authors solicited, and participated in the research from each case publication. The Hypoxia and Glider Cases shared two common authors, the PIs from each of the collaborating labs.

Data Collection

Data were collected using three approaches: document analysis, interviews, and field observation. These three approaches yielded four sources of data used for analysis: coded documents, interview transcripts, field notes, and memos about emerging themes.

Document analysis

The publications selected from each case served as the focus of document analysis. A given publication was read prior to interviewing the participants to familiarize myself with the research reported and the terminology used by the authors. A list of data reported in the research was also compiled, and used during the interview to confirm the types of data and specific variables used in the research. Publications were coded for descriptions of the research, documentation of data processing, and contributions from individuals not included in the author list, specifically in the acknowledgements section. Additional documents were collected from the participants during the course of the interviews, such as data management plans and protocol sheets. These additional documents informed interpretation of results. Reflections on the case publications and other documents were captured in memos.

Interviews

Participants were interviewed with two separate interview protocols, which can be found in appendices I and II respectively. The first protocol was the "follow-the-data" protocol that was developed to solicit accounts of the creation and handling of research data during the preparation of a publication. The second protocol was the "CENS wind-down" protocol, which was a series of questions to understand the current state of research during the last year of CENS, reflections on the center, and how people were preparing for the center's end. The questions in both protocols were a blend of open-ended questions to encourage narrative responses and focused questions to capture specific responses.

The follow-the-data protocol, was adapted from prior interview protocols developed by the CENS Data Practices research group. Questions in the interview protocol captured the paper topic and context, various contributions of the co-authors and how they came to be distributed, the tasks applied to the data, who was perceived to be responsible for data management and what being responsible entails, and whether participants felt they had fulfilled their data management responsibilities. Other questions captured baseline understanding of the participant's research and the role of data in their research. The following questions from this protocol were used in particular to construct descriptions for each research case.

To understand the context of the data and data practices descriptions of the research reported in each case publication were solicited using the following questions:

1.1. What type of research do you do?
1.2. What is the main research project are you working on now?
2.1 Is this article coming from the main research project you working on now?

2.1.1 If not, which project is it drawing from? [from those discussed earlier]

2.2 Where did the particular topic for this paper come from? (concept or idea)

2.2.1 Under what domain would you classify this research?

2.2.2 Do you think this made an incremental or radical contribution to the field?

2.2.3 How does this paper compare to others you have written?

2.2.4 How does this paper compare to others from the field? (typical, tech heavy, application science heavy, etc)

To understand the context of the data practices, descriptions of the data reported in each case publication were solicited using the following questions:

3.1 We have identified this list of variables as used in the research reported in this paper. Can you confirm that this list is correct?

4.1 For this paper did you use data from external data source (ie anyone not included in the author list)?

5.1 Would the particular data used for this article [table/figure] be relevant to others to re-use? (Data)

To understand the variety and distribution of data practices, descriptions of data practices and who performed which practice were solicited using the following questions:

2.3 Can you characterize your contribution to this paper? [involved in planning, interpretation, data collection or analysis, etc]

2.3.1 Who managed the data used in this paper while it was being prepared?

2.3.2 What were some of the different contributions of the collaborators for this paper with regards to collecting, managing, analyzing the data, and producing tables/ graphics?

3.2 Could you walk through some of the steps from designing the data collection plan to analysis? What happens as data are created, cleaned, analyzed, managed, etc.?

3.2.1 How did you select and verify your data for this paper?

3.2.2 What tools do you use to interpret and manage your data (Excel, R, MatLab, formulae, etc)?

3.2.3 Did your group keep records or a data log of how you acquired and processed the data? (workflow tools, etc)

4.2 Where does the data used for this paper reside? Where is it stored and accessed? (Could you show us?)

4.2.1 Could you locate the data for this [table/graphic/image]?

4.3 What system do you use for naming files, data collections, or versions of your publication or of the tables and graphics used in it?

4.3.1 Can you tell us or show us where you keep your versions?

4.3.2 Do you have additional source files containing data or other resources you used for the publication?

4.4 Do you keep the subsets of data used for the paper? Can you show us where?

To understand how researchers the process, descriptions of how data practices were distributed were solicited using the following questions:

2.3.3 How did everyone know what they needed to do for this paper?

2.3.3.1 Was there a formal process or was it understood?

2.3.3.2 Who holds whom accountable?

To understand how researchers perceived they were responsible for data, descriptions of data management responsibility were solicited with the following question:

5.4.3 Who is responsible for the dataset? What does responsibility entail? [Try to elicit types of responsibility]

To understand how researchers perceived they were fulfilling their data responsibilities, descriptions of data responsibility fulfillment were solicited with the following questions:

3.4 Are you having any trouble managing your data collection? In the short term or long term? Why?

3.4.2 Have you written a data management plan yet? If so, can you describe the experience?

3.4.3 Do you feel you have fulfilled your data management responsibility?

3.4.3.1 How would you compare your data management practices to those of your CENS colleagues?

3.4.3.2 How would you compare your data management practices to those of your domain peers?

5.5 For others to use this dataset, what sorts of additional information might you need to add to the data that is not in the article?

5.5.3 Would you need to include additional information about the instrument, other equipment or software tools used? For example, did you write your own code that might also need to be shared to use the data?

Prior interview protocols on which the follow-the-data protocol was based, had never quite matched the practices observed in field observation. In the prior protocols we had asked participants to ground their answer to questions using the last dataset they had handled. The vagueness of participant answers and the divergence of their answers from what was observed, lead us to believe that this approach was not grounded enough. By focusing the questions around a specific publication, the answers provided were more specific, and a better match to what had been observed in the field.

For the CENS wind-down data collection, being run by Lizzy Rolando and myself, we wanted to capture as many participants as we could. Rather than contact participants separately about both studies, we solicited them once, and asked the questions from both protocols during each interview. When Lizzy was available to be a part of a participant’s interview, she would ask the CENS wind-down questions. When I was the only investigator, I asked the questions from both protocols. Initially this was done to reduce strain on the participants who would make time for multiple interviews, but as the interviews progressed I found that some of the responses collected during the wind-down protocols augmented the follow-the-data protocol responses. The way the questions were written approached the same data handling practices from different angles, and in asking about certain practices more than once, we were able to paint a more complete picture of their practices. The questions that were particularly useful from this protocol are listed below. Question 3 was another way to ask about the data used in each case, and the other questions filled out the picture of the data practices performed by the researchers. Lizzy has solicited twenty participants for the CENS wind-down in addition to the sixteen interviewed for this dissertation. By combining the protocols, I was able to maximize the data collected for both studies and minimize the disruption to the participants.

3. Within your work, what do you consider to be data?
4. How do you manage your data collection so that you can use it again in the short and long-term future?
   4.1 How are data currently shared within your team?
   4.2 What are your criteria for selecting and preserving data? For sharing?
   5.1 Do you make any of your data public available online? Do you use public repositories?

The interview duration using both protocols ranged from an hour to just over two hours, averaging one hour and 15 minutes. Interviews were recorded, transcribed using a transcription service, and then the transcriptions were checked by hand to ensure accuracy. Three of the participants opted to be interviewed as a group, which lasted over two hours. The group interview had some benefit, as there was less repetition of baseline information and research description, allowing for more depth with regards to their activities. The group interview may have led to more consensus between participant accounts, as authors may have been unwilling to contradict one another in person. At the same time, the participants in this group all worked together closely and had a real regard for one another.

Field observation

For the scope of this dissertation I relied on my prior field observation experiences and field notes to cross-verified the interview and document analysis data. In the past eight years I have spent roughly twenty days in the field with researchers, as they deployed equipment and collected data. I have also participated in meetings and working groups, as well as attended community functions, such as the Annual Research Review, the Research Retreat, poster sessions, weekly lectures, and the more social coffee hours. I have maintained a desk in the UCLA CENS buildout, where almost a third of the center was housed.

Data Analyses

The interview transcript, notes, and memos were imported as resources into NVIVO where they were coded using the codebook found in the appendices. The codebook was developed to support both the dissertation research and the CENS wind-down research, so not all codes are applicable here. The codebook was developed based on the research questions and the interview questions that mapped to them, in addition to themes that emerged from my notes and memos. While coding the first few interviews, the codebook was revised to capture nuances. The rest of the interviews were coded with the updated codebook, and the initial interviews were re-coded to align with the other interviews.

Coded sources for given case were brought together and used to construct a picture of the case, by answering the following questions:

- **What is the publication?** What research is reported? What discipline does the research fall under? Is the paper typical of this discipline? What is the contribution this paper makes to the discipline? What is the context for this research paper? What research lead up to this point? Where does this paper fit in the larger research trajectory?

- **What data are reported in the publication?** What are the data sources and types? What resources were used for the collection of data?
- **What data management tasks are applied to the data?** How were the tasks performed, on what, and when?
- **Who are the participants in the data management for this paper?** What were their contributions to the data management in the paper? How were these roles distributed?
- **What is responsibility for data management within the context of the paper?** Who do participants think are responsible? What does responsibility entail? Who held whom responsible? Did participants fulfill their responsibilities?

Beginning with the hypotheses developed from prior research, the data were interrogated. Using grounded theory, the task of hypothesis development and refinement continued throughout research. Hypotheses were tested at successively deeper levels of abstraction using the data (Glaser & Strauss, 1967). Results were broken out by case to provide context, and to describe the division of labor between group members; and by research question to compare across the cases.

CHAPTER 5: RESULTS BY RESEARCH CASE

Results are presented in two chapters: in the first chapter the six cases are introduced, and in the second chapter the results are compared across all six cases to address the research questions. The names and designators for the six cases can be seen in table 5 below.

Case Designator	Case Name
Fungi Case	Modeling the role of fungi in soil respiration
Hypoxia Case	The role of an algal bloom in an hypoxic event
Power Case	Demonstrating and mitigating hardware variation
Glider Case	Path planning for ocean gliders
Stream Case	Demonstrating variability in whole-stream metabolism
Webcam Case	Capturing Spring green-up with public webcams

Table 5: Case names and designator used in the text, and the orientation of the publication.

The six cases varied by the number of disciplines collaborating and the category of the research findings. The variation between cases are summarized in table 6 below. The cases ranged from single to multi-disciplinary, with one group comprised of representatives from three distinct disciplines. Two of the cases were single-discipline groups and the other four cases were comprised of individuals representing two or more disciplines. Although researchers collaborated from multiple disciplines the publication that formed the sampling frame came from a single discipline. The discipline for each case was different, except for the Fungi Case and the Webcam Case which were both ecology publications. The publications were classified broadly as presenting technology findings, application science findings, or some blend of application science and technology findings. For example, the research in the Webcam Case described a new method for harnessing technology to measure a known phenomenon more efficiently. The researchers describe the research as “methods for environmental sensing” which is atypical of ecology research, and falls closer to technology research than scientific research. The Stream Case is also classified as in between science and technology because the discipline itself is somewhere between science and technology. The researchers indicated that they were studying the technology, methods, and the stream system.

Case Designator	# Disciplines	Publication Discipline	Publication Classification
Fungi Case	1	Ecology	Application Science
Hypoxia Case	2	Marine Biology	Application Science
Power Case	2	Computer Science	Technology
Glider Case	3	Robotics	Technology
Stream Case	1	Environmental Engineering	Science and Technology
Webcam Case	2	Ecology	Science and Technology

Table 6: Publication discipline, number of disciplines represented, and broad genre of research for each case.

The CENS collaboration spans five universities; all five universities were represented by the author pool. Participants were solicited from all five institutions, but only participants from four institutions agreed to participate. The following table depicts the institutional affiliations for the authors of each case, as well as the departmental affiliations if authors represented more than one department at an institution. The discipline of each publication is indicated with an asterisks in each case. The Glider and Hypoxia Cases have authors from the same departments, because the labs frequently collaborate with one another. The authors from the Power Case are from the same department, but represent two disciplines, electrical engineering and computer science, which have a significant overlap. More specific maps of how the authors from each case relate to one another are presented in the case descriptions that follow.

Case	Inst 1	Inst 2		Inst 3	Inst 4	Inst 5			
		Dept 1	Dept2			Dept 1	Dept 2	Dept 3	Dept 3
Discipline	Ocean	Robotics	Marine	Ecology	Env Eng	CENS	Ecology	CS	Elec Eng
Fungi				3*					
Hypoxia		2	4*						
Power								4*	2
Glider	2	2*	2						
Stream					4*				
Webcam						3	2*		

Table 7: Distribution of the publication authors across institutions and departments by case.

Descriptions of each case capture the following four aspects: (i) the research reported in the publication, (ii) the contributors to the reported research, (iii) the data reported in the publication, and (iv) the characterization of individual contributions to the reported research. More specific descriptions of data management tasks performed by each individual are captured in the next chapter. These descriptions were constructed from the interviews, document analysis, and field observation.

Fungi Case: Modeling the Role of Fungi in Soil Respiration

Research Reported in the Fungi Case Publication

The publication examined in the Fungi Case reports research on the role of ectomycorrhizal fungi in soil sequestration of carbon dioxide. Ectomycorrhizal fungi are fungi that extend in root-like structures from the end of plant roots to symbiotically exchange water for nutrients. The authors of the publication were part of a larger group collecting data about these structures. The specific research reported focuses on issues with carbon sequestration models that do not take these fungi into account. The authors were led to believe that there was a "decoupling" of some of the variables that were linked within the existing models. With the research reported in the publication the authors demonstrate how taking fungal involvement into account can re-couple the model.

The research reported in this publication falls within the broader discipline of ecology, and the narrower discipline of soil ecology because according to the first author the carbon cycle research is focused underground. The research reported is atypical of ecology due to the volume of available data. The research reported in this publication came from analysis of data collected by sensors deployed from 2004 to 2011. The volume of data generated by using sensors allows researchers to ask questions of dependence between variables and determine “what’s driving what.” Structural equation modeling, the method employed in the research is a new method for the ecology discipline. The contribution of the publication is to adjust how carbon flux is derived, because “carbon may be diverted to the mycorrhizal fungi.” Prior publications from this group address the technology developed to automatically capture images of root growth in clear tubes planted in the ground and measuring carbon dioxide at multiple depths to build predictive models of carbon production and absorption. The research relied heavily on sensor networks and technologies developed under the larger CENS umbrella, but only reports results that are of interest to the application scientists.

Contributors to the Research Reported in the Fungi Case Publication

Of the three Fungi Case publication authors, the first and second authors were interviewed. The third author is the PI of a lab, and the first and second authors were his students at the time. Both students have since graduated and are in tenure-track faculty positions. A much larger group of people was mentioned in the acknowledgements. Four people from the third author’s lab are acknowledged for helping to collect minirhizotron images over the years. Another two people from the same lab are acknowledged for assisting with image processing on the collected images. Two people from the research site are acknowledged for their help installing and maintaining the sensor network. And another faculty member is thanked for comments on the publication.

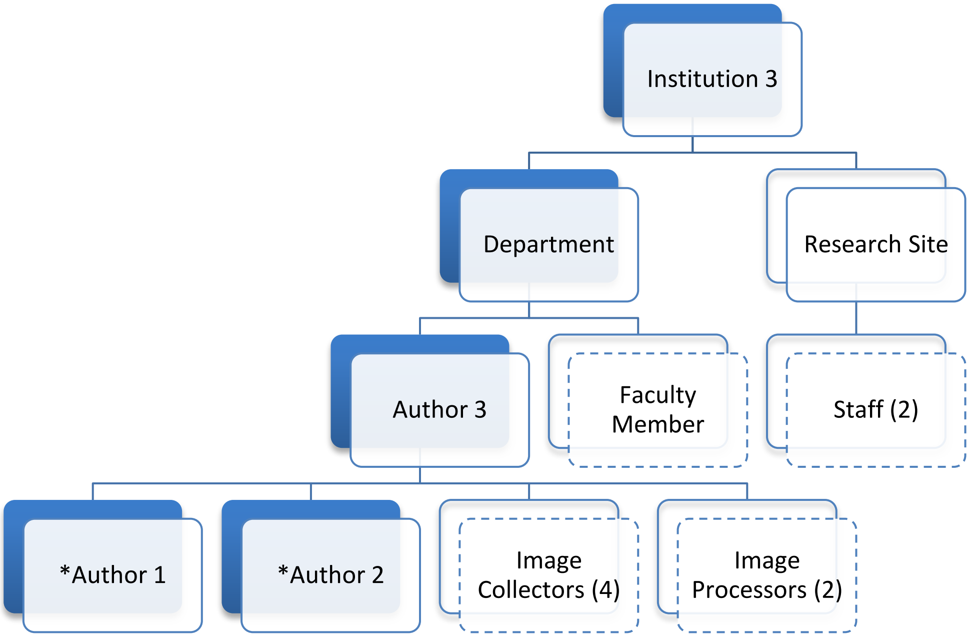


Figure 9: Institutional, departmental, and lab affiliations of the Fungi Case contributors.

The relationships between authors and other contributors from the Fungi Case can be seen in figure 9 above. The top two layers depict the institutional and departmental affiliations; in this case, the research site is actually a department at the same institution as the researchers. The middle layer shows the group leads, and the bottom layer contains the authors who work under other authors. Those authors with asterisks were interview participants. Dashed boxes denote those acknowledged in the publication or indicated to have contributed. Future cases will use the same visual cues to demonstrate the institutional, departmental, and lab affiliations of authors.

The first author describes the inspiration for the publication as being somewhat opportunistic, driven more by the availability of data than by a specific research question. The three authors came together during a brainstorming session and discussed what could be asked of the data. The first author mentioned that all three authors came together because they all had a “special interest” in the role mycorrizhal fungi play in carbon flux. For the first two authors this research was a side project to their respective dissertation work.

Data Reported in the Fungi Case Publication

The data used for the research reported included photographic images, sensor data measuring environmental factors, and a series of variables derived from the collected data. The images coming off the minirhizotron tubes buried in the ground were analyzed to determine root and fungi growth rates. The sensors measured environmental variables, such as soil temperature, moisture content, photosynthetically active radiation and carbon dioxide concentration. The sensors directly measure voltages and the first author points out that they, “actually have to change that voltage difference into a reading for us.” With all of these variables the authors were able to construct measures of root respiration and overall carbon flux. In order to create the structural equation models a series of correlation coefficients for the variables were derived. In order to provide some ecological context for the research, the authors provided descriptive measures such as feet above sea level, dominant tree species, and mean annual precipitation in the publication.

Types	Data	Subset or Derivative Data
Minirhizotron images	EM rhizomorph dynamics	EM rhizomorph production (rate)
	Fine root dynamics	Fine root production (rate)
Sensor array data (Environmental factors)	Soil temperature	
	Soil moisture	
	Photosynthetically Active Radiation (PAR)	
	CO2 concentration	
Biotic factors	fine root production	
	EM rhizomorphs production	
	Autotrophic (root) respiration	Root respiration
		EM fungal respiration
	heterotrophic (decomposer) respiration	
CO2 dynamics	Soil CO2 production	
	CO2 production (Pi)	
	Efflux of CO2 from soil surface (Rs)	
	Soil CO2	
	Soil CO2 dynamics	
	Net primary production	
	Diffusivity of CO2 in the soil profile	
	Flux of CO2	
	CO2 efflux	
	Flux divergence	
	Difference in efflux	
	CO2 assimilation rates	
	Carbon allocation	
Derived	Significant differences	
	Running averages	Mean daily averages (1day, 2day, 3day, 4 day)
	X^2 (goodness of fit)	Variance inflation factor
		Condition index/indices
	Cross-correlation	Correlation coefficients
		Expected correlation matrices
		Observed correlation matrices
Independent variables	Feet above sea level	
	Dominant tree species	
	Types of rhizomorphs	
	Time	
	Depth	
	Height	
	Mean annual precipitation	
	Leafing out	

Table 8: Data reported in the Fungi Case, organized into types based on publication and author input.

The first author believes that the data collected during the research project at this site would be of use to other researchers, though not necessarily the subset used in the research reported in the publication. The overall collection includes over six years of sensor-collected data and roughly 30,000 minirhizotron images. In particular, the first author offers that other ecologists would be interested in using this data.

Contributions to the Research Reported in the Fungi Case Publication

The first author characterizes his own contributions as identifying the appropriate data for use in the analysis. In order to use the structural equation model method the first author needed to identify coherent spans of data, where all of the variables of interest were collected simultaneously. Once the coherent data were identified, he organized them in such a way as would allow him to perform analyses.

> "So I'm not going to say that I collected all the data because that would not be true. But I will say that I organized the data such that we could start addressing these questions, performed the majority of the analyses, particularly the structural equation analysis for this particular paper" (Fungi Case, First Author).

The second author characterizes the first author's contribution as "playing with the data" and coming up with interesting questions.

According to the first author the second author did some of the statistical analyses, as well as providing "comments and contributions" to the writing of the publication. The second author on the other hand, when characterizing his own contribution to the paper, includes his contribution to the overall project. The second author was part of the team during the initial deployment of the sensing equipment. "So I designed the experiment. I was at the [research site] when we designed the plot and I installed the sensors. I programmed all the data loggers" (Fungi Case, Second Author). His contribution to the research reported included collecting and processing the data, all of the verification processing of the sensor data, generating the carbon flux data, processing some of the minirhizotron images, performing some of the analysis, providing a figure, and being "heavily involved" in writing the publication.

The third author, besides being part of the initial brainstorm session that lead to the research reported in the publication, is given credit for his comments and contributions to the writing of the publication by the first author. Like his own assessment of his contribution, the second author gives the third author more credit for contributions to the overall research project and getting this specific publication published.

> "[The 3rd author], well, he was the PI of the project, so he provided the overall... Of course, he wrote the grant and he got the money, but we actually had problems publishing that paper. … And then [the 3rd author] suggested to give another spin to the paper, the way it was written" (Fungi Case, Second Author).

The authors had trouble getting the paper published and only after the third author encouraged the first author to focus more on the technology side of the research were they able to get it accepted. So in the end this paper ended up being not just a presentation of findings about an ecological phenomenon, they were also a demonstration of the viability of the technologies and method used by this research group.

Author	Contribution according to First Author	Contribution according to the Second Author
First	Selected and organized data; performed analysis; wrote publication	Played with the data; developed research question
Second	Statistics and analysis; commented draft	Data collection; cleaning and processing data; analysis; wrote draft with first author
Third	Commented draft	Provided research direction; got funding; spun draft to make it publishable

Table 9: Summary of author contributions in the Fungi Case by participant account.

The second author also mentions the contributions of at least one of his colleagues from the lab who helped him with the minirhizotron data. There have been multiple people over the years who have stewarded that data, and they are all named in the acknowledgements.

The distribution of these tasks among members of the research group appears to have emerged naturally. As the first author explains here this research was different from other research because they already had the data. The second author had a similar assessment of how people knew who was contributing what. He offers, “when you have that large dataset, there's a lot more questions than what you can answer.” Because of this, he points out, members of the teams would pick the questions of interest and involve the other group members when necessary. The second author also makes the point that this research was outside of both the first and second authors’ dissertations, because there were no specific roles and they had the freedom to work with other people.

Hypoxia Case: The Role of Algal Blooms in Hypoxic Events

Research Reported in the Hypoxia Case Publication

The Hypoxia Case publication falls within the specific discipline of plankton ecology, which falls under both marine biology and biological oceanography disciplines. The publication reports on findings collected during a hypoxic event at a harbor in Southern California. A lack of dissolved oxygen in the water led to a massive fish kill, with millions of sardines dying over roughly a 24-hour period. The lack of oxygen persisted for nearly a week and dramatically affected the algal populations present in the harbor. The authors on this publication are chiefly interested in the studying the algae and the environmental conditions affecting them. This work found that the hypoxic conditions led to a significant change in the algal communities present, rather than significant changes in the algae changing the harbor conditions. While the research is unable to definitively say what caused the hypoxic event, the researchers were able to disprove many theories about what led to the fish kill event through elimination. The findings rely on data collected using sensing systems built by two of the authors, but the findings are only really significant to the application scientists, in this case the marine biologists. The research reported is atypical of the marine biology and biological oceanography disciplines, because of the volume of data that were collected before, during, and after the hypoxic event. The researchers do rely on methods, variables, processing, and analyses typical to the disciplines. The research reported in the publication could be characterized as opportunistic, because the data collection was driven by the event occurring, rather than some specific research question.

The overall contribution of the research reported in this publication is both a better understanding of the algal community's role in hypoxic events. The research is radical in that it overthrows the long-standing notion that algal communities are responsible for hypoxic events. Hypoxic events are quite rare and are geographically focused, limiting the generalizability of the research. The other contribution the research reported in the publication makes is demonstrating the efficacy of sensor networks in capturing rare events. Due to the unpredictable nature of the phenomenon, study of HABs can only begin after a bloom has accumulated and equipment can be deployed to the location. In this case, the equipment had already been in place for five years, following a prior hypoxic event in the same location.

Contributors to the Research Reported in the Hypoxia Case Publication

Of the six authors on the Hypoxia Case publication, the first, second, and sixth authors were interviewed. The authors on the publication come from two different labs at the same university, headed by the fifth and sixth authors. The first three authors are associated with the sixth author's lab. The first author was a graduate student researcher, the second author is a research faculty member, and the third author is a graduate student. The fifth author heads a lab that collaborates with the sixth author's lab, and the fourth author is a research staff member who works with him. The fifth and sixth authors are also the sixth and fourth authors respectively from the publication examined in the Glider Case. In addition to the authors, the acknowledgements list the following contributors to the research reported: two students and a technician from the sixth author's lab who helped collect and process samples; a faculty member who collaborates with the fifth and sixth authors for his guidance with the community structure calculations; a researcher at another institution for contributing a map of the harbor; the county life guards, the city fire department, and the city in general for access to the site and assistance.

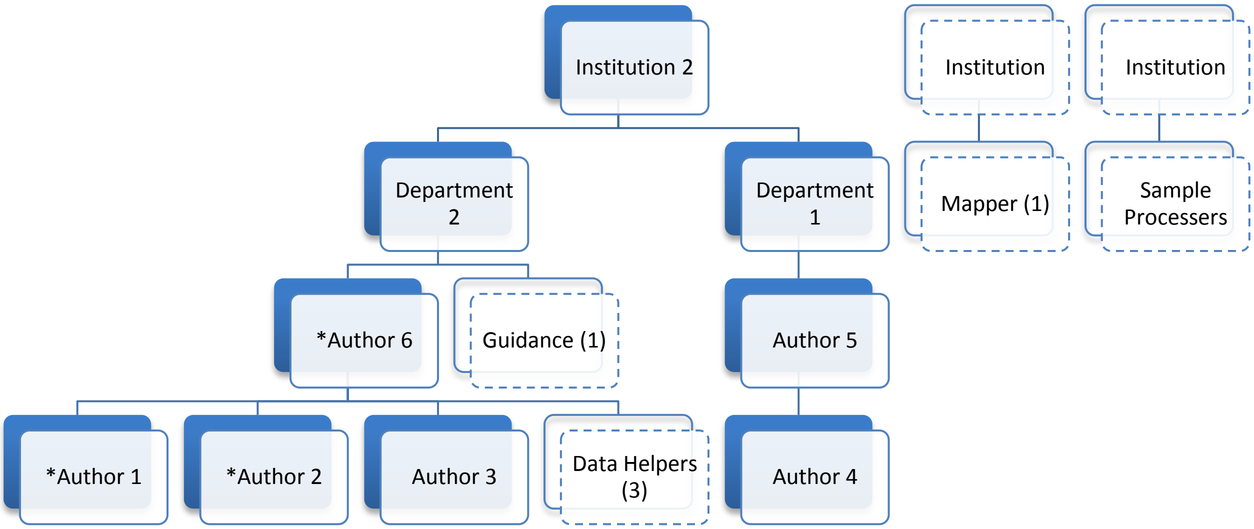


Figure 10: Institutional, departmental, and lab affiliations of the Hypoxia Case contributors.

After the hypoxic event began, the first author took the lead to make sure more data were collected to fill out the regularly collected data for higher spatial and temporal resolution and a broader range of parameters. For the first author the research reported fits within a larger trajectory studying the environmental and biological conditions that lead to the occurrence of harmful algal blooms. The second author saw this research fitting into another aspect of the first author's research, that of the movement of organisms, nutrients, and other 'stuff' through the water column. The sixth author saw the research reported here as being part of not only a larger trajectory to understand harmful algal blooms, but also part of a much larger trajectory to understand algae more generally. He is concerned with what he calls, "the base of the food chain," and how shifts in algal communities affect the rest of the ecosystem. The second author sees this research as being part of a larger trajectory to incorporate sensing systems in the study of marine biology.

Data Reported in the Hypoxia Case Publication

The publication lists the various data collected as part of the research, including equipment, calibration information, sampling rates, how they overcame “bio-fouling” on the sensors, and how the samples were processed. The data and data types listed below in table 10 were confirmed by the first author.

Types	Data	Subset
in situ sensor data	Salinity	
	Temperature	
	Dissolved oxygen	
	Chlorophyll a	
Water samples	Inorganic nutrient analyses (analates)	Nitrates
		Phosphate
		Silicate
	In vitro chlorophyll a	
	Abundances	Heterotrophs
		Dinoflagellates
		Diatoms
		Euglenids
		Other phytoplankton
Derived	Observed taxonomic richness	
	Like species richness (# of taxa)	
	Diversity (change in richness)	
	Evenness (distribution of the richness)	

Table 10: Data reported in the Hypoxia Case, organized into types based on publication and author input.

The research group collects in situ sensor data, including salinity, temperature, dissolved oxygen, and chlorophyll α. They also collect water samples which are then processed to measure inorganic nutrients, in vitro chlorophyll α, and plankton community abundances. The inorganic nutrients measured from the samples include nitrates, phosphate, and silicate. The in vitro chlorophyll α measure is more accurate than the in situ measure, but requires more effort, so instead the in vitro is used to verify and adjust the in situ measures. Abundances of both zooplankton and phytoplankton are counted. From the abundance data other data are derived, including taxonomic richness, how the richness is distributed, and how the richness changes over time. According the first author, these derived variables are, “statistics or indices used to describe various aspects of the community.” The researchers from this group tend to use meteorological data from external sources, such as NOAA, but according to the first author they did not use any external data for the research reported in the publication.

According to the second author, the sensors have been collecting data since they were installed and data are downloaded from them regularly. The process of creating measures from water samples is quite involved and over the years has been performed by multiple people from this group. Water samples are usually collected every week. A single water sample will be split many ways to create “filters,” which are quite literally the disposable filters of various pore sizes used to remove particulates from the sample. Filters of different pore-sizes are used to collect different algal species and even smaller things such as nutrients and toxins.

The data collected for this case could be useful to other researchers. According to the first author, these data would be of interest to those working with similar species for comparison research. Although she points out the community structure data would be of interest to others, she does not see much value in the raw data for use by others. Instead the processed data would be more usable. The second author agrees that the data might be useful to other researchers, but how these data could be used and what sort of documentation would be necessary, “would depend on the question.” Beyond academia these data would be of interest to other communities, such as government, water quality researchers, fisheries, and other groups.

Contributors to the Research Reported in the Hypoxia Case Publication

The first author characterized her contribution as being the person in charge of the data collection and analysis for the event. She asked the sixth author if she could have this role, and he assented to it. As with the rest of the lab, she was out collecting samples and pulling data off sensors throughout the event. She "spent a lot of time with the microscope", counting and characterizing the plankton communities. She ran the analyses, created the figures, and wrote up the research with input from the rest of the authors. The second author also characterized the contribution of the first author as making the research her own. The first author had been researching the algal communities in the harbor prior to the event, so when the event occurred she was well positioned to say where additional equipment was needed and adjust the physical sampling frequency.

The second author characterized her contribution as the one who initially collected physical samples at the site, although that job has been handed off to a series of lab members over the years. For the research reported in the publication she assisted with the statistics and community structure calculations and how best to tie together the various data sources from the event. Like other members of the lab, the second author helped with the data collection during the event, deploying new sensors, sampling, and setting up equipment. The first author gives the second author credit for the community diversity statistics.

The first author characterized the contribution of the third author as helping in the field collecting samples and pulling data off the sensors, as well as performing some of the basic data processing in MatLab. The second author characterized the third author's contribution as fundamental, because she was the lab technician at the time of the event and the more sophisticated sensing equipment was her responsibility. The first author characterized the contribution of the fourth author as helping in the field. The second author characterized the contribution of the fourth author as deploying and maintaining the sensors, dubbing him "MacGyver." This author fabricated some of the temperature sensors deployed in the harbor.

The sixth author characterized his own contribution as being one of groundwork. He trained the first author, put together the project, maintains cordial relations with the city and site, and other supervisory contributions. The relationships between his group and the research site are not only important for access, but for alerts as well. When dead fish started appearing the sixth author was called by someone at the harbor, and the group was able to mobilize in response. The sixth author also had an editorial contribution to the writing and figures, and overall quality assurance of the research. The second author characterized the contributions of the fifth and sixth authors as being the "heads and brains" of the project. The first author indicates that the fifth author was included author as an author because the research fell under the collaboration between the fifth and sixth authors, and the equipment in the harbor was maintained by his group.

Author	Contribution according to First Author	Contribution according to the Second Author	Contribution according to the Sixth Author
First	In charge of data collection and analysis; wrote up research	Took ownership of the research	Hands-on data collection, processing, and analysis
Second	Some statistics	Initial data collection; data collection during event; some statistics	
Third	Collecting samples; basic data processing	Maintaining and pulling data off sensing equipment	
Fourth		Deploying and maintaining sensing equipment	
Fifth	Equipment maintained by his lab	Head and brains of the project	
Sixth	Instrumental in keeping the sensors in the harbor, through funding and relationships		Groundwork with site; training lab members; editorial contribution

Table 11: Summary of author contributions in the Hypoxia Case by participant account.

In addition to the authors and the individuals who are mentioned in the acknowledgements, another group of people contributed to the research. The nutrient samples were processed by a lab at another institution for “a nominal fee.” The members of the sixth author’s lab could run the samples on their own, but it is less cost effective.

The division of labor among the researchers was based on expertise and availability for the research reported. The second author thought the work came together organically. Subsets of the authors would get together, work on some problem, decide what the next steps were, and who would do what. The first author gives a similar account of meetings where the division of labor was discussed. At least during the data collection she mentions that whoever was available would go out to the site, because they needed as many people as possible. The first author pointed out that this event was really quite exciting for the lab, and “everyone was very keen” to get some research out of the data being collected.

The second author described the group as everyone holding themselves accountable. The first author described the situation as being variable over time. When the event first began, the excitement drove the work, but once that was over, the first author found she sometimes needed to “nudge people” to give her things she needed, such as some analysis or comments on a draft. The sixth author mentioned that he holds his students accountable for documenting their processes, as well as proper storage and backing-up of their work.

Power Case: Demonstrating and Mitigating Hardware Variation

Research Reported in the Power Case Publication

The research reported in the Power Case publication falls within the discipline of computer science. As electronics become smaller, new variability is arising between individual devices that should nominally function the same. At this scale the manufacturing tolerances are not tight enough; even a single atom difference in the placement of components on a processor can lead to significantly different performance. Manufacturers provide an estimate of the performance of the processor, but the estimates are always on the safe side and can underestimate the performance. When researchers put sensing equipment in the field, they try to balance the frequency of data collection with the battery life of the equipment. Pieces of software called power managers handle this balance, but they rely on the manufacturer's performance estimates. When the estimates are conservative, the researchers are not able to take full advantage of the equipment for collecting data, and may sacrifice sampling density unnecessarily.

The publication demonstrates the variability of power across microprocessors at different temperatures, which can reduce the accuracy of estimates for power management. The authors propose a software solution that dynamically adjusts the task manager based on the equipment's performance rather than rely on the manufacturer estimates, overall boosting the possible sensing frequency by a factor of seven. Because the research examines hardware and provides a computer science solution, the research falls along the boundary between electrical engineering and computer science. The Power Case publication primarily presents the computer science findings, where a companion article goes into more depth about the variability testing and modeling hardware performance.

The research reported is both original and incremental. This is not the first time smart algorithms have been used for power management, but this is a novel application. The hardware research communities have long-predicted that variations across individual devices would emerge. Rather than a far off prediction, these variations can be observed today, and proved a complication for sensor network researchers at CENS. Demonstrating the variation and using a simple software stack solution to mitigate the problem was low hanging fruit for these researchers.

Contributors to the Research Reported in the Power Case Publication

Of the six authors on the Power Case publication, the first and sixth authors were interviewed. The variability project is run by the fifth and sixth authors, who are both faculty in the electrical engineering department. The sixth author has a joint appointment in the computer science department. The first, second, and third authors were all students of the sixth author. The first author is still a student, but the other two have recently graduated. The fourth author is a student of the fifth author. There is no acknowledgement section in the publication, but other students from the sixth author's lab were mentioned during the interviews as administrators of the shared data repository. According to the sixth author there are two to three students serving as informal administrators to the repository at any given time.

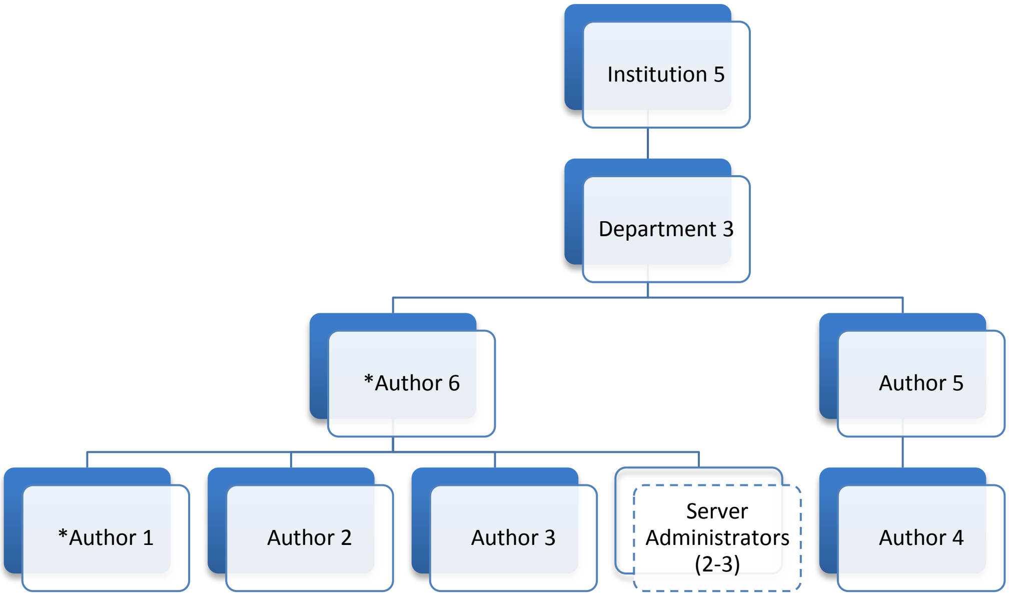


Figure 11: Institutional, departmental, and lab affiliations of the Power Case contributors.

The sixth author was originally made aware of the problems working with hardware as a result of his involvement with CENS's sensor network research. Embedded sensor networks rely heavily on efficient power management. Not only would identical devices behave differently at the outset, but they would also age differently. The sixth author wrote and received a center grant for the study of the device variability, with this research being the first reported findings. The first author was generally interested in the development of operating systems for embedded networked sensors. When the hardware variability research was developed he saw an opportunity for developing an operating system to take advantage of the hardware variation. He is working towards an operating system that will completely adapt all running parameters to the individual device.

Data Reported in the Power Case Publication

The first author described three phases of data collection and analysis: hardware measurement, characterization, and benchmarking the software. During the hardware measurement phase, the researchers tested the variance across nominally identical processors by measuring duty cycling, lifetime, battery capacity, sleep power, active power, and energy left untapped while the hardware was in different modes and at different temperatures. These are called “performance measures” by the sixth author. For every instance of device A in mode B at temperature C, a raw file of performance measures was generated. They had ten devices in total, two modes (asleep and awake), and ran between twelve and fifteen different temperature points. From these data the authors derived the average power for each device over time at a given temperature and mode. In the characterization phase, the researchers took the temperature profiles aggregated for each device in each mode, and fit the profiles to a model of device performance. The model was used to develop the software stack algorithm, so that the stack could make more accurate performance estimates for a device given the observed variations. While benchmarking the software stack performance, the researchers measured how long the system took to run and other “microbenchmarking" data. They had previously found that the device performance was more dependent on ambient temperature than mode, so they ran a temperature simulation. During the evaluation phase they simulated a long-running system using temperature profiles they downloaded from the National Climactic Data Center. The temperature profiles provided a more realistic scenario of the temperature variation a device would encounter when embedded in the environment. They ran the scenario varying the available battery power, capacity, and other variables as well.

Type	Data
Processor Performance	Duty Cycle
	Lifetime
	Battery capacity
	Energy left untapped
Power	Sleep power
	Active power
Temperature	Ambient temperature
	Thermal dynamics of packaged chip

Table 12: Data reported in the Power Case, organized into types based on publication and author input.

According to the first author, the other partners on the variability project would be interested in using the data from this research. He also suspects that peers would be interested in using the data because they are so novel.

Contributors in the Research Reported in the Power Case Publication

The author contributions, like the data collected, follows the phases of data collection. The first author mentions receiving help from the fourth author in collecting the hardware measurements. The first author began collecting power profiles, and the fourth author expanded the range of instances. The fourth author also fit the measurement to the model during the characterization phase of the research.

> "So I had done the kind of initial set up and initial experiments and then [the fourth author] kind of took over and did larger range of measurements across

> temperature and across more instances. And I guess his most important contribution is then after having these measurements kind of fitting them to a model" (Power Case, First Author).

The first author attributed the work on developing the software stack to himself and the second author. The final evaluation of the stack through simulation was attributed to the second and third authors. The first author claims he was not involved in that phase.

According to the first author, the first three authors wrote most of the paper, with each author writing up the phase they contributed. The fifth and sixth authors also contributed to the writing of the publication, but in more of an advisory capacity. "So, [the fifth author] is [the fourth author]'s advisor. And [the sixth author] is my advisor. So they had kind of a say in all of this pretty much" (Power Case, First Author). The sixth author characterizes his role as advisor, and goes on to say that advisors are "the useless ones." The sixth author also describes himself as "the conduit to the prior work," and says he brought the idea for the paper to the group as part of the center grant proposal he and the fifth author wrote.

Author	Contribution according to First Author	Contribution according to the Sixth Author
First	Collected hardware measurements; wrote software stack; wrote up section	
Second	Wrote software stack; simulation and evaluation of software stack; wrote up section	
Third	Simulation and evaluation of software stack; wrote up section	
Fourth	Collected hardware measurements; fit model; wrote up section	
Fifth	Advised research	Wrote grant proposal and came up with idea
Sixth	Advised research	Wrote grant proposal and came up with idea; conduit to prior work

Table 13: Summary of author contributions in the Power Case by participant account.

According to the sixth author there was no formal distribution of roles. The first author drew his understanding of what needed to be done in terms of the phases of research from the overall variability project. From there, contributions were made by area of expertise. Through weekly or bi-weekly meetings the authors were able to discuss progress and next steps. The first author also mentions that a subset of the authors would meet ad hoc in order to discuss issues that arose during research. The sixth author gives a similar account of the meeting schedule, adding that the meetings were held regularly at the beginning of the research, but became more impromptu as the research progressed. The sixth author mentions specifically that he does not have a formal method for assigning work, and prefers to let it evolve naturally. The first author described the collaboration as running “pretty smooth,” and everyone held themselves accountable.

Glider Case: Path Planning for Ocean Gliders

Research Reported in the Glider Case Publication

The publication discussed in the Glider Case is a robotics paper that reports on the development of path-planning algorithms used in driving an ocean-bound autonomous underwater vehicle (AUV) during sampling missions. The underwater vehicle is tasked with following an ephemeral phenomenon, a harmful algal bloom (HAB). These blooms are virtually impossible to predict, and where they travel is determined not only by ocean currents, but by available nutrients and other competing species of algae. This publication reports on the incorporation of prediction data generated by ocean models to increase the path planning algorithm accuracy of an AUV following a harmful algal bloom. The publication includes results from two field experiments and one simulation. In this case field experiments are two-days to one-month long deployments of the AUV following some prior HAB observed in the ocean model, the test being whether or not the AUV managed to follow the path of the HAB.

To present the path planning experiment and results, the authors explained the biology of harmful algal blooms, the scientific application for which the path-planning algorithms were developed. The publication spends more time on the application for the technology than would typically be presented in a robotics publication. The number of experiments reported in the publication also contributes to its abnormality within the field of robotics. Robotics researchers tend to use simple robots within laboratory conditions where an experiment could last only a minute at a time, allowing the researchers to perform many experiments before reporting to the community. According to the sixth author, the research reported is part of a long-term problem to be solved over the next twenty years; incorporating ocean, biological, and chemical models into robotic path-planning decisions to improve accuracy.

Contributors to the Research Reported in the Glider Case Publication

Of the six authors on the Glider Case publication, the first and sixth authors were interviewed. The sixth author is a faculty member who collaborates with the fourth and fifth authors. The fourth and fifth authors are faculty members at the same institution as the sixth author, but in different departments. The fourth and fifth authors are application scientists for whom the technology developed in the sixth author's lab is meant to support. The collaboration between the fourth, fifth, and sixth authors, is described by the sixth author as a "federation of three labs." According to the sixth author this federation is cemented by sharing grants between labs. The first author was a postdoctoral researcher in the lab of the sixth author. The third author is at a different institution and heads a group that maintains the ocean model. The second author is a researcher within the third author's group.

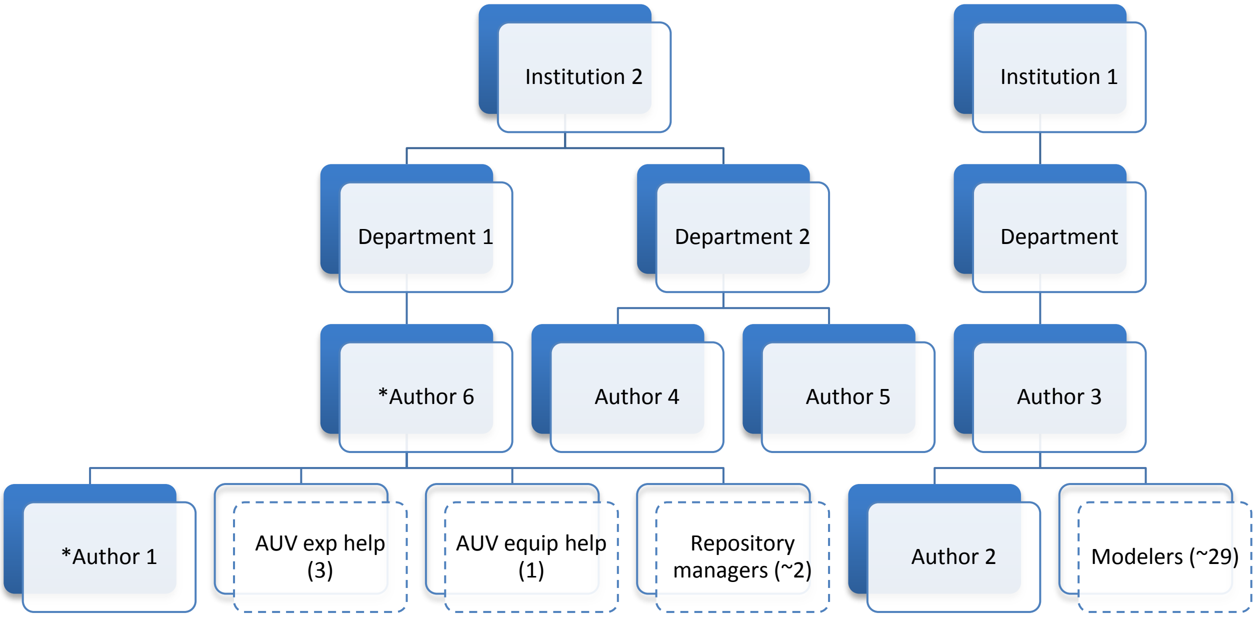


Figure 10: Institutional, departmental, and lab affiliations of the Glider Case contributors.

For the second and third authors the data collected by the AUVs was fed back into the ocean model to "improve the fitness," but otherwise the research reported did not advance their research. The fourth and fifth authors were concerned with the development of better technologies to support their scientific application. The sixth author describes the way AUVs have been used for marine applications in the past, and the AUV would be put in an area of interest and follow a simple "lawnmower" path for collecting data. A lawnmower path lacks the nuance to track HABs which change course unpredictably. The fourth author has two options, either purchase multiple $150,000 AUVs or use the one he already owns more efficiently.

Data Reported in the Glider Case Publication

As mentioned above, the research reported came from two field experiments and a simulation, which generated location data about the AUV. During the two field experiments, the AUV would surface every four hours and email the researchers its current location. During the simulation experiment, location data were generated by the simulation. All of this position data were compared to actual data from an HAB to determine how well the AUV or AUVs tracked the center and boundaries of the HAB. The researchers fed prediction data from the ocean model into their path-planning algorithm. The ocean model prediction data were downloaded from the model website and fed into the AUV every time it surfaced. According to the sixth author the prediction data are themselves generated from satellite forecast data.

The AUV also captured data about the ocean and other data of interest to the collaborators, such as chlorophyll and fluorescence line height. These data were logged on the AUV and sent to the ocean model researchers and marine biologists respectively. If the AUV had been tracking an actual HAB instead of a virtual one, the marine biology data collected would have fed into the bloom-tracking algorithms, but for the purposes of this research the data collected by the AUV was not being fed into the algorithm and are thus not reported in the publication.

According to the first author, the data collected by the AUV about the ocean and the marine biological phenomena would be of interest to a variety of other communities, such as physical oceanographers, microbiologists, sanitation districts, fisheries, and remote sensing. Other robotics researchers may be interested in the position data, but it is less likely because in the ocean, “there is no way to ground truth your data” (Glider Case, first author). The first author points out that researchers who run experiments using simple robots in laboratory conditions are able to ground truth their data, verify their data at the time of collection, making it more usable by others.

Type	Data	Subsets
Mission	Time between surfacings of glider (T)	
	Time glider is deployed at sea	
	# of trajectories	
	Average distance of trajectory	
	Depth-averaged current estimation	
Location	Present location of glider (L)	
	Speed of glider (v)	
	Distance traveled by glider (dh)	
	Number of hours spent traveling (h)	
ROMS Predictions	High Frequency radar surface current measurements	
	Moorings data	
	Sensor data	
	Satellite data	
	4D velocity predictions	Ocean current predictions
		Vertical current profile predictions
Ocean Plume tracking	Geographical locations encompassing plume's extent (D)	Comprised of drifters
		# of drifters
		hourly predictions of location of each drifter in D
	Convex Hull – Minimum Bounding Ellipsoid (E)	
	Centroid of Convex Hull (C) (assumed to centroid of plume)	
	Geographical distance between C at 4 hours and the starting location (dg)	Lower bound for distance (dl)
		Upper bound distance (du)
Communications	Length of communication	
Plume Discovery Data	Satellite imagery	
	Direct observations	
	Fluorescence line height	
	Chlorophyll	
Trajectory of glider	Waypoint Selection Algorithm	
	Waypoints	
	# of waypoints generated	
	Path for glider (LCn) or (Lp) or (pCn)	
	Velocity of ocean current	
	Aiming point (A)	
Derived	Median error between actual surfacing location and prescribed surfacing location	
	Reduction rate in navigational error	
	Optimization parameter	
	Speed of the plume	
	Boundary depth for glider	
	Boundary depth for plume	
	Depth-averaged current vector	

Table 14: Data reported in the Glider Case, organized into types based on publication and author input.

Contributions to the Research Reported in the Glider Case Publication

According to the first author, he contributed the vast majority to the work on the research reported as well as the overall writing of the publication, “to be bluntly honest, probably over 90% is mine.” The second and third authors were brought on as authors to acknowledge the assistance they provided with the ocean model.

> “[The 3rd author] had very little contribution other than providing the model and the initial collaboration that we started with the first two papers, and then just email correspondence as to asking questions about what variables are this and how does this work and things” (Glider Case, First Author).

The second author specifically had a significant contribution in making prediction data from the ocean model available in a way that would allow the first author to update the AUV. The first and second author worked together to develop an interface to allow the data to be grabbed quickly when the field experiments were being run.

> “[The 2nd author] probably had the largest contribution in that she was instrumental in making sure data were available and setting up the FTP of transferring the data from the vehicles back into the model and helping close the loop and we developed a Google API to model some drifters and she helped with that. … So a lot of the grunt work, if I had any questions, ran through [the 2nd author] and so quite a bit of that probably a good 10% I would give to her” (Glider Case, First Author).

The sixth author gives a similar account to the contributions of the second and third authors, but is more specific about how the ocean model was tailored to the needs of the field experiments reported.

> “The [3rd author’s group] actually gave us the ocean forecasts. And so we directly interfaced with their software system … then it produces the ocean forecasts. And so they made that available and I believe they read a draft, then commented on the

> draft paper, but their main contribution was the model. It's a specialized model so we need it tuned just for the area where our robot is working and so that was what they provided" (Glider Case, Sixth Author).

The fourth and fifth authors contributed their application science knowledge, and through discussions of how the harmful algal bloom phenomenon moved through the ocean. As mentioned earlier a clear understanding of the application allows for a clear understanding of the path planning problem being researched. The fifth author was especially helpful in this regard when providing feedback on the manuscript.

> "I had significant contributions from [the 5th author] basically in conversation. In that, to help me understand what was happening in the physical ocean and understand the oceanography and the processes involved. He was instrumental in being able to right some of the sections that I wrote. I can say that a lot of it is paraphrasing of a conversation I had with him" (Glider Case, First Author).

The first author also credits the last three authors, who make up the lab federation, with the overall motivation for the research being reported. "The science is self-motivated by [the 4th and 5th authors], [the 6th author] had the big picture idea, and I pulled it all together to formulate it" (Glider Case, First Author). The sixth author describes his own role as being part of the planning and writing for the research reported, as well as through his supervision of the first author. "I was involved in the planning and I was not involved in the implementation at all. It was all led by [the first author]" (Glider Case, Sixth Author).

Author	Contribution according to First Author	Contribution according to the Sixth Author
First	90% of the work	Led the research
Second	Worked with the 1st author to make the ocean model data usable in the glider algorithm	Made the model available and tuned it for the area the glider covered; commented on draft
Third	Provided access to the ocean data model	
Fourth	Provided comments on draft	Wrote sections; consulted on science
Fifth	Provided guidance on understanding the ocean system; commented on draft	
Sixth	Had the big picture idea; supervision	Planning; supervision

Table 15: Summary of author contributions in the Glider Case by participant account.

Beyond the authors, many other people contributed to running the AUV, storing the data in the lab repository, and managing the ocean model used to supply prediction data to the AUV. The first author described how the mission lead interacted with the AUV during experiment runs, or “missions.” The lead needs to be available every three to four hours to interact with the AUV for the entire time the AUV is on a mission. Members of the lab rely on one another to keep an eye on the AUVs, which is a great relief to the person running a mission who can get very little sleep during the mission.

> “When we run these, the lead on the mission is in-charge of the vehicles. So, you're basically awake for three to four weeks at a time standing in front of the screen watching to make sure they come back. … And so every four hours they were calling in. … Well, it happens with all of us. … And we get text messages, we get emails, and so people are watching” (Glider Case, First Author).

Since a number of people in the lab run missions on the same AUVs, there is reciprocity among the lab members. The help provided by three of the lab members with tracking the AUV during the mission is acknowledged in the publication. The authors also acknowledge contributions from a staff member from the sixth author’s lab for, “his work with [AUV] hardware making field implementations possible and simple.” According to the first author, the group managing the ocean model is quite large and while two of the group are included as authors there are roughly thirty people in total working to support the model. Within the sixth author’s lab there are a number of students responsible for the lab server and the system developed to pull in data from emails and the AUV Secure Digital (SD) cards.

> “So, and all the tools to massage the data once it comes off the robot are all sort of home brewed, right? And they tend to live because two or three students or postdocs sort of maintain them” (Glider Case, Sixth Author).

These students may be the same ones acknowledged in the publication for their help with the AUV experiments, but that is unclear.

The various contributions by the authors and the wider communities they represent in the research were well understood by the time they worked on this publication. The sixth author characterizes the collaboration on the research reported as being the same as prior work, because the authors have collaborated on five or six papers together. He indicates that the work, including writing the publication, was distributed by area of expertise, with application science authors contributing application science and the robotics authors contributing robotics. The first author corroborates that this is an understood process at this point, although his account of what was done by each author is a bit different. According to the first author, he wrote all the sections and then asked for feedback from the appropriate co-author.

Stream Case: Demonstrating Variability in Whole-stream Metabolism

Research Reported in the Stream Case Publication

The publication selected for Stream Case is from the disciplines of environmental engineering, stream ecology, hydrology, and ecohydrology. Whole-stream metabolism is “a calculated term representing the net dissolved oxygen (DO) change per day resulting from biological activity” (Stream Case Publication). Whole-stream metabolism estimates are based on a single sampling point in the stream, but this approach does not capture an accurate estimate of the entire stream metabolism. The research reported demonstrated the variability of whole-stream metabolism calculations at multiple sampling points across a stream, comparing calculations from three stationary sites: one in the middle of the river, and the others close to each bank. These three stations were in full sun, partial shade, and full shade. The authors found significant differences across the three sites based on sun and depth of the stream bed. The specific topic for this research came from sensor research with CENS and from a more recent collaboration with an international group who perform on metabolism calculations in lakes. Rivers are more dynamic than lakes, and there was some question as to whether metabolism calculations could be used to say anything about rivers and streams. From this question, the group began testing their sensing system in the streams to measure whole-stream metabolism, just to “see what would happen.”

This publication first presents scientific findings and second demonstrates the method and technology used to accomplish the method. As a proof-of-concept the stream chosen was small and easy to work with in terms of deploying equipment. In the future the researchers intend to use the same equipment and calculations in a much larger river system. For the purposes of this research the publication has been classified somewhere between application science and technology, because environmental engineers and the research reported are somewhere between application science and technology research.

Contributors to the Research Reported in the Stream Case Publication

All four authors on the Stream Case publication were interviewed. The authors were from the same lab; where the fourth author is the head of the lab, and the other authors are his students and staff. The first and second authors are current Ph.D. candidates about to finish their dissertations in the next year, and the third author is a technical support staff member. The third author was hired after participating in the fourth author’s lab as an undergraduate researcher and splits his time between the fourth author’s projects. No other individuals are acknowledged in the publication. During the interviews, a prior student from the fourth author’s lab is mentioned for his development of the data repository they all use.

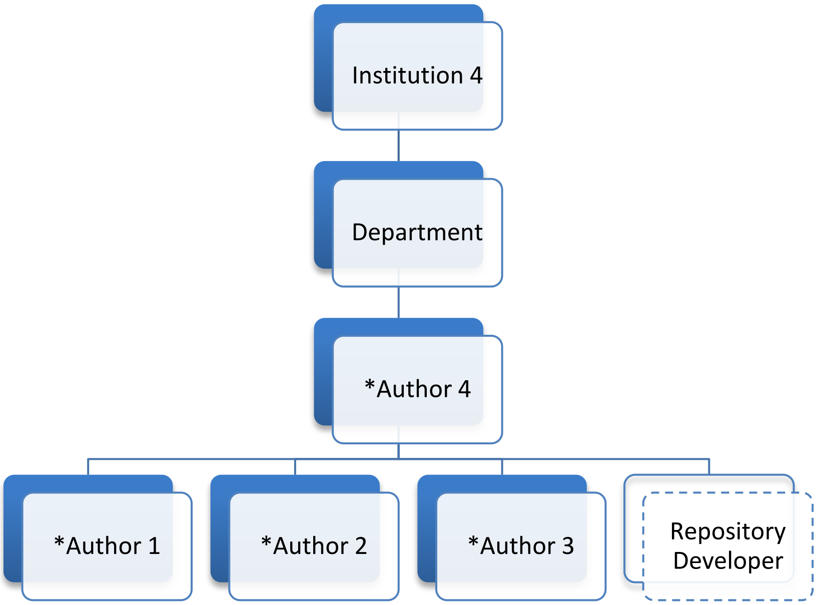


Figure 13: Institutional, departmental and lab affiliations of the Stream Case contributors.

The fourth author describes how this research is part of a larger goal to manage reservoirs in an agricultural area so that they feed into healthy streams, which is what he would classify as an “environmental engineering question.” To achieve this goal the research group is studying “river water conditions in time and space.” According to the fourth author the Stream Case publication is one building block in a predictive model. From the hydrological side, the third author characterizes the research reported as contributing to the ground water portion of the predictive model. The first author describes the overall research trajectory as assessing how water management and agricultural practices affect river systems. She goes on to describe two of the current ways the research group are trying to address this problem: Eulerian and Lagrangian. Her own research and the research reported in the publication itself falls within the Eulerian category; planted sensors measure water flowing past. The second author’s primary research is within the Lagrangian category; sensors float along with the water to measure how the water changes as it encounters new conditions. From the research reported, the research group will continue to test new variables for their predictive model.

Data Reported in the Stream Case Publication

According to the fourth author what counts as data are the observations from sensors and any products derived from those data. He also mentions that the research group tends to work with simulated data and observational data. For the research reported, only observational data were collected. The other authors from this group have similar understandings of what constitutes data.

The data used for this paper included independent variables that captured geomorphic properties of the stream and meteorological conditions, such as river topography, air temperature, solar radiation, photosynthetically active radiation, wind speed, and relative humidity. The data included dependent variables used to measure the physical, chemical, and biological properties of the water, such as dissolved oxygen, water velocity and temperature, specific conductivity, and chlorophyll activity. According to the second author the sensors capture indirect measures for many of these variables, so the actual phenomenological values must be derived from the sensor readings. From these independent and dependent variables some variables are derived, including net daily metabolism, which is a measure of whole-stream metabolism.

Types	Data
Geomorphic properties	River topography (bathymetry)
Physical/chemical/biological properties of water	Dissolved oxygen (DO)
	Water velocity
	Water temperature
	Specific conductivity
	Chlorophyll-a
Meteorological conditions	Air temperature
	Photosynthetically Active Radiation (PAR)
	Solar radiation
	Wind speed
	Relative humidity
Derived variables	Gross primary production (GPP)
	Community respiration (CR24)
	Net daily metabolism (NDM)
	Photosynthesis to respiration ratio (P/R)

Table 16: Data reported in the Stream Case, organized into types based on publication and author input.

In addition to the data that were collected and derived by researchers within the group, they also used data from an external source. The data from the USGS gage stations nearby the research site are used to check the experimental conditions, rather than to contribute to the pool of variables. While performing the research they did not want to be surprised by a "big variation of flow" through the equipment, and the gage stations would have provided them with forewarning.

The data collected as part of this research may be of use to other researchers. The fourth author identifies some possible consumers of the data, such as state agencies, water resource agencies, the water quality control board, and possibly other environmental engineers. He mentions that stream ecologists would specifically not be interested in the data collected because the site is a stream impacted by agricultural waste, and is no longer a "native" place. Regardless of who would be interested in the data, the fourth author does not think there is really enough data to be of any use to others. The data were collected over three days, which is not long enough to observe most phenomena of interest.

Contributions to the Research Reported in the Stream Case Publication

All of the authors contributed to the research reported. At some points in time all of the authors contributed and at other points in time the contributions were broken up by the authors' are of expertise.

The fourth author characterized his contribution as starting with the idea of trying to use the metabolism calculation method. When the first author took the lead, the fourth author retired to "lead editor" and was available for his students to consult. He also participated in the collection of data, and even claims to have collected the best data because he "got up and changed the batteries at like three in the morning."

The third author characterized his contributions as the deployment design and assisting in data collection, which were limited to the beginning of the research effort. The first author characterized the third author's expertise as knowing about the equipment and how best to deploy it. "So usually [the 3rd author], he does a lot of the hardware, like where to anchor in and where to... Like if we need to put something in a certain place, is this the best place for it" (Stream Case, First Author).

Both the first and fourth authors characterized the second author's contribution as collecting, cleaning, and processing the data. The fourth author even dubbed the second author "Mr. Scripts" due to his skills with the statistical package R. The first author also attributed the figures to the second author. The second author's assessment of his own contribution is in line with what was said by the others, including producing figures and processing data. The figures were not created alone by the second author, at least the first three authors got together to discuss them and, according to the first author, they sometimes brought in the fourth author to determine, "the best way to present what we want to present."

The first author was given credit for writing up the findings by the second author, “She's been writing mostly.” (Stream Case, Second Author) Just as data collection is a group activity, so is reviewing and revising the publication. To sum up the interaction on the project the third author offered the following, “We get a lot of feedback from each other. … Pretty much all phases of the project from conception to implementing it and analysis and all that” (Stream Case, Third Author).

Author	Contribution according to the First Author	Contribution according to the Second Author	Contribution according to the Third Author	Contribution according to the Fourth Author
First		Writing the draft		Took the idea and ran with it
Second	Cleaning and processing the data; developed figures; feedback	Cleaning and processing the data; developed figures; feedback	Feedback	Cleaning and processing the data; supporting role
Third	Handled the equipment and placement; feedback		Designed equipment deployment; collected data; feedback	Supporting role
Fourth	Data collection; Feedback		Feedback	Came up with idea; assisted in data collection; editor

Table 17: Summary of author contributions in the Stream Case by participant account.

The group had no need for a formal distribution of who should do what at this point, because they have been working together for a number of years successfully. The third author characterized their interaction as “understood.” The second author mentioned that they try to play to individual strengths, for instance the third author is better with Photoshop and the second author is better at R, which dictated who did what. The fourth author also cited the time this group has worked together as playing a role in how each person understood what was expected of them.

> “I think they're mature enough now. I mean they've been around, and this is the… looking at time series is sort of [the 1st author]'s dissertation so she was kind of

> like the lead and everybody else is... They play a little bit of a supporting role, but pretty much once the data is moved into the right folder, they're off back on their other things.” (Stream Case, Fourth Author)

The authors all worked in the same cubicle area, except the fourth author who had an office across the hall, which afforded frequent interaction. According to the first author, they typically work as a group and are all very supportive of one another’s work.

All of the authors are agreed that they held themselves accountable. The third author remarks, “With our dynamics we're able to just to ask, ‘Hey, when can I expect that?’” The first author sees this as a matter of mutual respect for one another’s space. The fourth author describes them as “a good group,” and points out that he has not needed to mediate any issues between members. Even though the fourth author is the PI of the group, the others do not see him holding them accountable. The third author even described the PI’s approach as being “hands off.” Although according to the second author occasionally the fourth author will need to prod one of them for a draft or some other deliverable.

Webcam Case: Capturing Spring Green-Up with Public Webcams

Research Reported in the Webcam Case Publication

The research reported in the Webcam Case publication was an evaluation of how well public internet cameras captured spring “green-up.” Green-up is when plants sprout leaves, literally going from bare branches to green leaves. The timing of this phenomenon varies by latitude, but is also “sensitive to changes in environmental conditions,” and is used as an indicator in Global Climate Change. Images captured from webcams were compared to remote sensing products in order to determine efficacy. Although webcam images were complicated by “varying image exposure and color correction noise,” the webcams demonstrated equal or better performance when compared to the remote sensing images. Thus, the publication demonstrated the viability of using publicly available cameras for studying large-scale ecological phenomena. The publication falls within the discipline of ecology, and can be characterized as methods paper. In this case, the method reported is cheaper and complementary to existing plant ecology methods, including remote sensing.

Contributors to the Research Reported in the Webcam Case Publication

Of the five authors on the publication, the first three authors were interviewed. The fifth author is an ecology faculty member, who hired the first author to work at CENS. The first author is a research staff member at CENS, and serves as the resident ecologist. The second author is a graduate student of the fifth author. The third author was a technical staff member at CENS, who worked primarily with the first author. The fourth author is a computer science faculty member and in the upper ranks of CENS. Beyond the authors, two other individuals, a CENS research staff member and an engineer from the infrared astronomy lab, were acknowledged in the publication for their contribution to the data analysis. This is the only case where the first author and the PI are not affiliated with the same department, although they were in the same department before the first author was hired at CENS. This connection is manifested in figure 14 below by a line connecting the first and fifth authors. The fifth author is listed as the corresponding author.

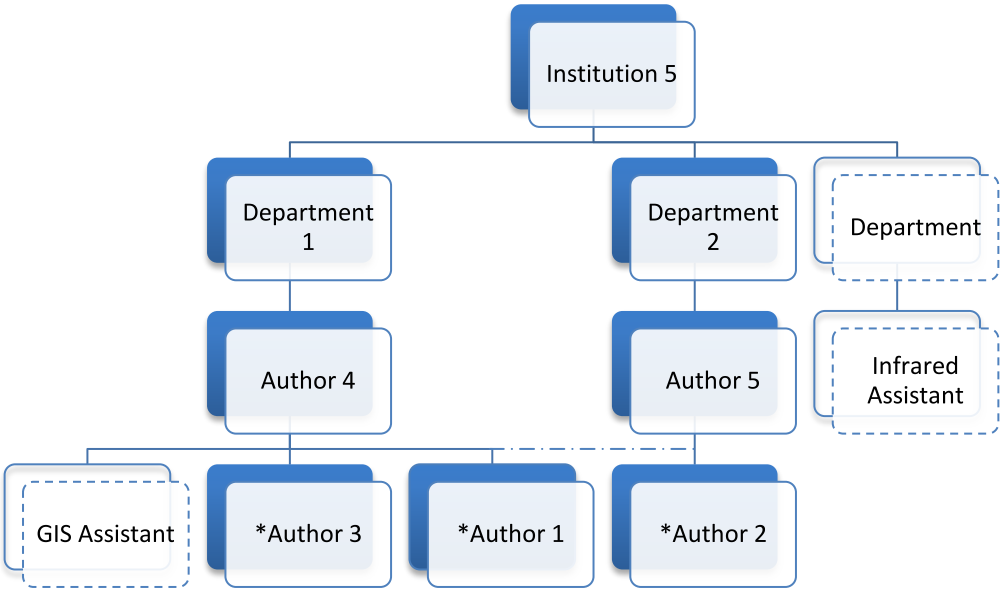


Figure 14: Institutional, departmental, and lab affiliations of the Webcam Case contributors.

The first author described the research reported in the publication as an extension of his work with individual web cameras at a local reserve. After collecting images from these cameras for a year, he realized if they had more cameras they could track trends over a larger area. According to the first author, the fourth author was very excited about the idea. For the second author the research was a side-project from her dissertation research. The third author saw the research reported in the publication as an extension of another project he and the first author had worked on that collected citizen science data about when flowers were blooming to capture a more precise measure of environmental dynamics.

Data Reported in the Webcam Case Publication

The webcam data were collected from public webcams identified across the contiguous US. Webcams were found through Google searches for "nature cam" and "tree cam." When a list of webcam IP addresses had been compiled, a script collected images from the cameras twice daily for a year. Not all the camera feeds remained available throughout the entire year. The publication reports that 1400 publicly available webcams were identified that captured vegetation images. Descriptive metadata were discovered about each camera from the IP address and website, including physical location, elevation, resolution, etc. Webcams were classified based on the type of plant cover they captured. From the individual images a number of variables were derived, such as grey-scale pixel average and average excess green per pixel. The third author has continued to collect webcam images twice daily since the research was published.

The remote sensing data used to evaluate the webcam performance were acquired from freely available NASA MODIS satellite datasets. According to the first author, they acquired satellite images from a few different sources, all of which provided processed images. They also used eight and sixteen-day derived products. In order to match the satellite data to the webcam images, a significant amount of processing was required on the part of the second author, with some help from the third author. As reported in the paper, satellite-based land cover classification was also captured to classify the webcams. From both the webcam images and the remote sensing data the mean date of Spring and visual Spring were derived and compared. A series of error estimates and significance measures were also derived.

Type	**Data**
Webcam Images	Image from public camera
	Time period for data collection
	Date/timestamp
	Sampling time for photo from website
Webcam classification	Segmentation Category (Deciduous, Evergreen, Understory, Nonvegetated)
	Visual Camera Classification (Mixed forest, evergreen, understory annuals, open shrublands, urban, deciduous, pasture, etc.)
	Satellite Classification (Mixed forest, Evergreen Needle-leaf forest, Open Shrublands, Urban and built-up, etc.)
Camera	IP address
	Public camera identifier (location #)
	US State where camera is located
	Elevation of camera
	Number of times photos were taken from website
	Camera resolution
	Camera performance (Based on consistency of images, lack of malfunctions, and stationary camera location)
	Lat. & Long. of camera

Camera Aggregate	# of public cameras discovered
	# of public cameras that were geo-referenced
	# of cameras used for test group
	Rate of loss of cameras
Image Processing	Grey-scale pixel average
	Average excess green (includes value of green pixel, value of red pixel, and value of blue pixel)
Phenomenological	Showing green in spring
	Fully green in summer
	First signs of leaf senescence in fall
Satellite Images	MODIS images of surface reflectance
	Resolution of MODIS images
	Satellite-based land cover classification
	Dates of collection for MODIS images
	Title
	Scan
	Line
	NDVI (includes red and near infrared bands of surface reflectance product)
	Per pixel quality assurance scientific data
	Daily 3x3 pixel matrix
	Median of high quality pixels in 3x3 matrix at a point in time
	NDVI time series
	Hyperbolic tangent (tanh)
	Time in Julian Day (t)
Derived	Greenness minimum (w1+w2)
	Total amplitude of greenness signal (w1-w2)
	Highest rates of change of the fitted greenness function (u and v)
	Standard error for estimates of spring
	Standard error with 20% filter
	# of poor quality days
	# of consecutive poor quality days
	Median total number of poor quality days
	Median gap of consecutive poor quality days
	# of cameras modeled using whole-image excess green
	Mean date of Spring
	Maximum date of estimated Spring
	Minimum date of estimated Spring
	Mean date of visual spring
	Maximum date of visual spring
	Minimum date of visual spring
	Root-Mean-Square-Error
	Root-Mean-Square-Error with 20% filter
	# of hand-segmented areas
	# of hand-segmented areas that had modeled excess green values
	And # that did not
	Estimated date for Spring in specific hand-segmented area
	Maximum difference between two hand-segmented locations in one area

	Minimum difference between two hand-segmented locations in one area
	Pearson’s correlation coefficient (r)
	Statistical significance (p)
	Mean difference between 50% green visual ground truth and modeled estimates for deciduous
	Mean difference between 50% green visual ground truth and modeled estimates for understory
	Mean difference between 10% nongreen leaf color and modeled estimates for deciduous
	Mean difference between 10% nongreen leaf color truth and modeled estimates for deciduous

Table 18: Data reported in the Webcam Case, organized into types based on publication and author input.

According to the first author, the images collected from the webcams may be of interest to other people, although he thinks they would be more likely to identify their own cameras and collect their own data. He points out that the MODIS data are freely available, so no one would want the MODIS data they captured. The second author mentioned that the authors had discussed other possible ways the data could be analyzed, and that others may be interested in performing those analyses. She goes on to say that whomever wanted to use the data would need the R and Python scripts from the authors. The third author mentioned that the data would be useful for image processing as a reference dataset.

Contributions to the Research Reported in the Webcam Case Publication

According to the second author, the first three authors worked closely together, literally and figuratively, as they shared a row of desks at the CENS building. The first and second authors developed the overall direction of the research. The second and third authors worked together to collect both webcam and satellite data, then worked together to process the satellite data. The first and third authors worked together to process the webcam data. Then the first and second authors compared the webcam and satellite data and split the writing responsibilities.

The first author characterized his contribution as coming up with the idea for the research, performed the image analysis on the webcam images. The second author characterizes the contribution of the first author as having the inspiration for the research. The third author also attributes the initial idea to the first author.

The second author characterized her contribution as identifying the webcams, and discussing the best times during the day for regularly capturing the webcam images. She also describes herself as being in charge of locating and collecting the satellite imagery, extracting the data and modeling it. The first author characterized the contribution of the second author as identifying the webcam data sources and performing the image analysis on the satellite images. She also verified the instances where the webcam data analysis failed. Both the first and second authors mentioned that she developed the statistical methods used.

The third author characterized his own role as one of implementation, as the first author came up with the idea, and the third author came up with a way to implement it. He mentioned that he stored and extracted the satellite data, as well as organized all of the data for analysis. He wrote a simple browser that would allow the first and second authors to move more intuitively through the massive amounts of image data that had been collected. The first author characterized the contribution of the third author as writing the automatic data collection scripts and handling the storage and retrieval of the images. The second author indicated that the third author was in charge of the "scraping and storing" of the webcam images. She also attributed the webcam image analysis to the third author; although the first author claimed that he had performed this work.

> "And at the same time [the third author] was going through the camera image data, and coming up with a comparable measure of greenness from the RGB channels, and we're comparing that to red and infrared reflectance from the satellite imagery, so an index" (Webcam Case, Second Author).

The third author mentioned that the webcam image analysis began as his task and soon changed hands to the first author, which might account for the second author's confusion in her account of who performed the analysis.

> "When it started out, I think I did most of it. But then as it went on I think, mostly due to him growing impatient, waiting for me to get things done. I mean he started picking up like, you know, Python and scripting, learning how to do image processing tasks" (Webcam Case, Third Author).

According to the second author, the third author helped her extract the satellite data from the format in which they came, which the second author describes as, "a huge help."

The fourth and fifth authors were not mentioned as having active roles in the data handling, but they did motivate the research. The first author mentioned that the fourth author's enthusiasm for the initial topic is what pushed them to perform the research. And the first author characterized the contribution of the fifth author as being "very encouraging."

Author	Contribution according to First Author	Contribution according to the Second Author	Contribution according to the Third Author
First	Came up with idea; performed webcam image analysis; split writing	Inspiration for research	Came up with idea; image analysis
Second	Identified webcams; performed satellite image analysis; developed statistical methods; split writing	Identifying webcams; collecting satellite images; data extracting and modeling	Satellite image analysis
Third	Writing data collection scripts; image data storage and retrieval	Scraping and storing webcam images; extracting satellite data; webcam image analysis	Stored and extracted satellite data; organized data for analysis
Fourth	Motivated research		
Fifth	Very encouraging		

Table 19: Summary of author contributions in the Webcam Case by participant account.

The Webcam Case publication is the only one to contain an authorship statement, which is as follows.

> "Authorship: [the first author] and [the second author] were equally responsible for designing and performing the research, analyzing data, and writing the paper;

> [the third author] was responsible for performing the research, analyzing data, and contributing analytic tools; [the fourth author] and [the fifth author] were responsible for providing scientific leadership and collaboratively writing the paper” (Webcam Case Publication).

The authorship statement provided roughly the same division of labor as the interview responses.

As with the other cases, the division of labor seems to be more informal than formal. The physical proximity of the first three authors made interacting with one another easy, according to the first author. The second author added that when a problem arose, they could just deal with it as a group because they were right next to one another. According to the second author, these three authors had worked together previously, so the division of labor was informal. They would meet together as needed, to assess their progress and decide who was going to do what to move the research forward. The third author characterized the distribution of research tasks as being understood.

The authors interviewed gave varying accounts of who held whom accountable. According to the first author they were under duress to get this research out before someone with similar research, so the three first authors were all motivated to get the work done as quickly as possible. Rather than one person holding another responsible, these researchers were more likely to help whoever was running late on their piece. According to the second author, all of the authors held themselves accountable. The third author mentioned that he reported to the first author, and that the first author was “the one in charge of the project.”

Summary of Results by Case

This chapter described the context of a publication from each case of six cases. The context included a brief description, the contributors, the data, and the contributions to the research reported in each publication. Before moving on to discuss the data management performed by the researchers on these cases I would like to highlight three themes in the data: the nature of research performed at CENS, the relationship between authorship on the publication and contribution to the research, and the diversity of research data.

CENS Research

The cases presented included two publications that reported purely scientific findings, two publications that reported purely technological findings, and two publications that reported a mixture of science and technology findings. The science cases were both described as being atypical for their respective fields because they had orders of magnitude more data than the typical publication. Both presented some science findings, as well as some demonstration of method. In the Fungi Case the structural equation model approach was demonstrated, and in the Hypoxia Case the in situ sensing infrastructure was demonstrated. Of the two technology cases, one was described as typical and the other was atypical. The Power Case presented research that demonstrated the presence of a predicted problem, and a software solution to mitigate. The Glider Case on the other hand was atypical because models from the scientific application were used to inform the navigational algorithm being tested. The Stream and Webcam Cases both presented findings that were a mixture of science and technology, and were described as methods papers. Methods papers are a genre of publication that is seen in both ecology and environmental engineering, but they are part of the minority. Publications reporting scientific findings make up the bulk of publications. The Stream Case reported on the utility of metabolism calculations and demonstrated the need for a series of measuring stations that transect rather than a single point measure when applying the method to a stream. The Webcam Case reported on the utility of publicly available webcams for capturing ecological phenomena. In both of these cases embedded sensing networks were being evaluated for their ability to support scientific applications.

Publications from five out of the six cases were atypical or in the minority for their discipline. This can be attributed to their being part of the CENS research community. None of these publications reported on development of embedded networked sensing, but they did report research that could not have been performed without the affordances provided by CENS. The science cases benefitted from the significant increase in data collected by embedded networked sensing equipment. The mixed science and technology cases tried out new applications for embedded networked sensing or took advantage of embedded networked sensors. And in the one atypical technology case, the Glider Case, the robotics authors leveraged their relationship with the marine biology researchers to increase the efficacy of their algorithm. Although the research reported in these publications was atypical, the authors believed that their respective fields would move in this direction, i.e., larger volumes of data, embedded networked sensing, and using actual phenomena to improve systems. The one publication typical of its discipline, that of the Power Case, was not necessarily CENS research, but was research inspired by problems that were observed in prior CENS research.

Authorship and Contribution

The publications ranged from single-discipline author groups to multi-disciplinary author groups. Regardless of the number of disciplines represented in the author list, each publication reported results to only one discipline. This may have been sampling bias, because in three of the four multi-disciplinary cases, only authors from the publication discipline were interviewed. Authors from the collaborating disciplines were solicited for interviews, but none of those authors replied to the solicitation. Because I received no response, it is unclear as to whether they did not feel comfortable talking about the publication or did not have the time. The participants who were interviewed were asked who else should be interviewed from the publication. The participants only indicated other individuals from the publication discipline. Contacting those authors from a collaborating discipline was even discouraged.

For two of the cases, only the authors from the publication discipline handled the data. The marine biologists from the Glider Case discussed how best to take advantage of the ocean model data, but did not handle the data in any way. The ocean modelers did contribute data to the algorithm, but they did not handle any of the data used in the reported analysis. Similarly, the robotics researchers from the Hypoxia Case helped deploy equipment, but did not handle any of the data processing or analysis. On the other hand, in the Power and Webcam Cases, the authors from both disciplines handled the data. In the Power Case, I was discouraged from contacting other authors because I would not have heard anything other than the description already provided by the two authors who were interviewed. In the Webcam Case, authors from both disciplines were interviewed.

In all six cases the lead author was in the first author position in the byline, and the PI from the publication discipline was in the last author position. Authors were otherwise ordered by contribution. The contribution implied by author order was somewhat confirmed in the contribution descriptions provided by the participants. However, two cases were notable exceptions. In the Power Case all of the authors made significant contributions, and this would not have been apparent from the author order. The first four authors all collected and analyzed data, and wrote up their sections of the publication. The last two authors came up with the idea and the funding, and then supervised the implementation, providing comments on the manuscript. In the Webcam Case, the first two authors had nearly equal contributions, which would not have been apparent from the author order. The first and second author took data from different sources, ran the same set of analyses and compared the outcomes, and shared the writing up of the publication. The only difference in contribution was that the first author came up with the idea, and oversaw the research, which put him just ahead of the second author in contribution. In this case the publication contained an author contribution statement that would remedy any misconceptions introduced by the author order.

The line between author and acknowledgement shifted across the cases. There were contributions made by individuals who were not authors, and were merely acknowledged. In the Fungi Case, there were acknowledged individuals who had helped collect the data for years, but had not participated in that particular analysis. In the Hypoxia Case, the entire lab and anyone else they could muster participated in the collection of data during the fish kill event, but only those who participated in the subsequent analyses were included as authors. In the Glider Case, an entire lab of individuals was acknowledged through two of the included authors. The reverse also occurred, where individuals were included as authors who had not contributed significantly to the current research. The Hypoxia Case is the easiest to identify of these cases. The fourth and fifth authors are included for their prior role in the establishment of the embedded networked sensors in the harbor, but were not otherwise involved in the research reported in the publication.

Data Diversity

Even in this small sample, the data reported in each of the case publications are very diverse. At a very basic level of comparison, the number of data variables reported, the cases ranged from eight variables in the Power Case to dozens of variables in the Webcam Case. The only variables reported in common across most cases were time and date-stamps. Despite both being ecology, there is almost no overlap in the data reported in the Fungi and Webcam Case publications.

There is overlap in the variables reported in the Hypoxia and Glider Case publications, perhaps due to the shared authors and the lng-standing collaborations between the research labs.

The data in five of the six cases split along a boundary between collected or generated data and derived data. The collected and generated data were measures of something in the world or generated by a simulation. The derived data were the products of data processing and analysis. For instance, in the Stream Case data was collected describing the river along the transect, the nutrients in the water as it flowed past the sensors, and the larger environmental conditions. These data were then used to calculate the daily changes in the dissolved nutrients. In the one exception, the Power Case, the data were used to generate a model, which was not considered to be data.

The participants from each case organized their data into groups; some were organized by method of collection, as in the Hypoxia Case, while others were organized by properties, as in the Stream Case. For the majority of the cases some blend between method and properties were used to organize their data.

CHAPTER 6: RESULTS BY RESEARCH QUESTION

The prior chapter introduced each of the six cases being studied. This chapter digs deeply into the data management tasks described by the researchers interviewed, encoded in the publications, and observed in practice. Results in this chapter cut across the six cases to address each of the research questions.

R1 What tasks are performed by researchers to manage their data?

R1.1 What stages of the data life cycle are observed, and what purpose do they fulfill?

R1.2 What is the disposition of data at each stage of the data life cycle?

R1.3 What data management tasks are observed?

R1.4 What data management tasks are associated with each stage?

R1.5 For whom are the data managed?

R2 How are data management tasks distributed among members of the research group?

R2.1 Who performed which data management tasks?

R2.2 How were roles distributed?

R3 For what data management tasks do researchers perceive they are responsible?

R3.1 Who is responsible for data and what does responsibility entail?

R3.2 Were the responsibilities fulfilled and to what standard?

R3.3 Who holds whom responsible?

R1 What tasks are performed by researchers to manage their data?

In the following section, a generalized model of the data life cycle is constructed, making note of the disposition of the data during each life cycle stage. Observed data management tasks are classified into a typology, and then mapped onto the stages of the data life cycle. Finally, the question of who are the intended recipients of the managed data are explored.

R1.1 What stages of the data life cycle are observed, and what purpose does each fulfill?

The following generalized data life cycle stages were observed across the cases. In each stage researchers perform tasks that affect their data, which data are collected, how they are collected, processing in preparation for analysis, and so on. The life cycle begins prior to the creation of the data, during Planning, in which the researchers decide what data to collect, and extend to preservation after the results are published. The data are connected to the performance and reporting of research. Data is typically collected for a specific project, with the exception of long-term data collection efforts. The findings of the project are then reported in one to a few publications. The final data from the end of one data life cycle informs the next research project, thus starting the next data life cycle.

Planning

Beginning with a research question, researchers identify the necessary variables to collect, the necessary equipment to collect them, a location where the variables can be collected, and other method details. CENS researchers tend to design the research method using experiences gained from prior research, such as what sampling frequencies are necessary to capture the phenomena while satisfying practical constraints, such as battery life. The planning at the beginning of the research ensures that researchers capture the data they need to support their hypothesis.

In all of the cases, researchers went through a planning stage. In the Fungi and Hypoxia Cases, the data had already been collected or was being collected prior to planning for the research reported in the publication, because the research stemmed from existing data collection efforts. For these cases there were Planning Stages for both the long-term data collection effort and the research reported in the publication. In the Fungi Case, planning the research involved planning a new method to mine the data collected during the long-term deployment. In the Hypoxia Case, planning the research occurred just after the researchers learned about the fish kill and needed to respond to the situation in order to maximize their data yield from the existing deployment supplemented with their reserve equipment.

Equipment Calibration

Before sensors and other equipment are deployed they are calibrated. Chemical sensors are calibrated using known solutions and physical sensors may have a protocol the researcher runs through in order to ensure that compasses are pointing north, etc. The researchers will typically record the offset between the actual measurement and the expected measurement for each piece of equipment. Some equipment are calibrated again in-field, a task referred to as "ground-truthing." The offsets recorded during the calibration of each piece of equipment are later removed from the data, during a process referred to by researchers as “data calibration.” Very complex equipment may require a series of tests and motions to allow the equipment to calibrate itself. Calibration ensures that the researchers are able to overcome the variation exhibited by equipment and construct true measures of the studied phenomenon.

In the Fungi Case, the equipment was calibrated prior to deployment, several years prior to the research being proposed. In the Hypoxia Case, the long-standing equipment deployed were swapped out and recalibrated every week or two weeks, to account for calibration drift, which happens quite rapidly to sensors in water. The equipment added to increase data collection resolution were calibrated in the lab with known solutions before deployment. In the Power Case, the measurement equipment were factory calibrated. In the Glider Case, the glider is a very complex piece of equipment that was calibrated on land, to ensure that the application science data collected was calibrated. For the technology researchers there was no way to calibrate the glider in a way that would capture offsets for their data. In the Stream Case, sensors were calibrated in the lab and then ground-truthed in the field. In the Webcam Case, there was no way for the researchers to calibrate webcams owned by other people, but they were able to collect the information for each webcam necessary to offset the image data based on sun-height and angle at each webcam location.

Data Acquisition

Data are collected or generated through the use of equipment or simulations. Sensors deployed in the field collect observations of physical phenomena. Some sensor measurements are direct (e.g., temperature, wind speed) and others are indirect (e.g., measure of fluorescence as an indicator of chlorophyll activity). Some data are generated through running simulations or experiments. Simulations use models and data to generate new data. Some data are acquired from external sources, such as the National Oceanographic and Atmospheric Administration and the United States Geological Society, to provide context for collected data or to warn researchers about poor data collecting conditions before they occur. During the data acquisition stage, researchers gather evidence to support or refute claims or to indicate where more research is needed. Multiple rounds of data collection may be employed.

In all six cases, data were acquired. In the Fungi Case, the data were collected prior to research and the researchers chose some subset of the data to work with for the research reported in the publication. In the Hypoxia Case, the data were collected for five years prior to, during, and after the hypoxic event. Not all of the data were included in the research reported in the publication, but knowledge derived from prior years about the algae communities informed the analysis. In the Power Case, data were collected in three stages to verify that the research problem existed, then to test and evaluate the software solution. In the Glider Case data were collected during two experiment runs and a simulation in the lab. They collected data for robotics, physical oceanography, and marine biology, even though only the robotics data were part of the research reported in the publication. The other data was logged and shared with their collaborators from other disciplines. In the Stream Case, data were collected during a three-day deployment. In the Webcam Case, the researchers automatically captured webcam images for an entire year. These researchers also acquired satellite images and image data from a variety of external sources.

Case	← Prior	Planning Stage	Post →		
Fungi Case	6 years	Post acquisition			
Hypoxia Case	5 years	During event	Collection continued		
Power Case		Prior to acquisition	Measurement	Testing	Evaluation
Glider Case		Prior to acquisition	Experiment 1	Experiment 2	Simulation
Stream Case		Prior to acquisition	Three-day Deployment		
Webcam Case		Prior to acquisition	Year of capture		

Table 20: Summary of the data acquisition stage and relationship to the planning stage for each case.

Data Processing

Data processing encompasses a variety of tasks, such as cleaning, integration, and derivation of the data, which are performed in preparation for data analyses. Data need to be calibrated to remove the equipment offsets recorded in the equipment calibration stage. Sensor data, in particular, are collected at a much higher density than is necessary for analysis to minimize the effects caused by outliers, so these data must be averaged into composite points before they can be analyzed. Data across multiple variables or multiple sites are integrated to allow for correlation or comparison respectively.

In the Fungi Case, data were run through scripts to identify outliers and otherwise clean the data. The first author identified which data to aggregate for the analysis, and any gaps in the data were filled. In the Hypoxia Case, the water samples were processed into units that could be used for species counts and for measuring nutrients. Sensor data were cleaned using statistical methods and data were smoothed through averaging, reducing the number of points in each timeseries to a more manageable size. In the Power Case, data were brought together to develop a model. In the Glider Case, data processing included mapping all of the location data collected during experiments. The first author indicated that the location data did not require any calibration or cleaning. In the Stream Case, data were run through filters for outlier removal, and the gaps were filled using algorithms. In the Webcam Case, the total number of webcams was reduced to thirty by processing and fitting the webcam data to models. The satellite images needed to be cut apart and reduced to match the webcam data. All the image data were run through image processing scripts to construct comparable image profiles.

Data Analysis

Researchers use statistical, modeling, and visualization tools that vary by research specialty and individual preference to perform analyses of their data. Analysis allows the researchers to generate and test hypotheses or otherwise draw conclusions from data. Data analysis generates the final data, which is then presented in the publication.

In the Fungi Case, the researchers generated a structural equation model. In the Hypoxia Case, the researchers combined timeseries of each variable from the data collection stations and plotted them against the timeline of observed events to ultimately disproved the dominant theories about what caused hypoxic events. In the Glider Case, the researchers evaluated the performance of the path-planning algorithm at following a simulated harmful algal bloom, by comparing where the glider went and where it should have gone in order to capture the simulated phenomenon. In the Power Case, the researchers demonstrated the variability of equipment by comparing the power curves created for each device, and evaluated the performance of the software solution to overcome the observed variation by calculating the overall efficiency of devices using the software solution. In the Stream Case, the researchers compared the timeseries captured in whole-stream metabolism calculations with metabolism calculations from a single point in the stream. In the Webcam Case, the researchers evaluated the accuracy of webcams in observing green-up by comparing with corresponding remote sensing data for the same locations.

Publication

Although data acquired during research may not be published on their own, the findings are included in scholarly publications such as journal articles, conference papers, posters, and technical reports. Data underlie the findings made during the research, and can be seen in publications as data presentations, which include maps, graphs, and tables of the values. The publication of results allows data to enter a much larger scholarly conversation where they are validated through peer review.

In the Fungi Case, the data was reduced down to a model of variable dependence, and the degree of dependence between each of the variables in the model. In the Hypoxia Case, the data presented in the publication include images of the dominant algal species and timeseries at different stations within the research site. In the Power Case, the sample temperature profiles are provided, as well as aggregate profiles used to create the models. The software stack performance over temperature, time, and mode were also provided. In the Glider Case, the data are visualized in a series of maps showing the path planned by the algorithm, and the desired path based on later knowledge. In the Stream Case, the researchers developed new ways to visualize their data, because they added new dimensions to those exploited by their peers, who tend to publish timeseries. In the Webcam Case, the data presented included sample webcam images and profiles, as well as a comparison between the green-up observed via satellite and webcams.

Preservation

If researchers perform any actions that specifically make the data available to a wider audience, these actions happen only after the research has been published. Preservation actions include culling the data and physical samples no longer of use to other researchers, the long-term storage and maintenance of data the researchers deem valuable enough to keep, and depositing data in discipline repositories. The preservation of data supports reuse of the data by other researchers both within and beyond the research group.

In only two cases were specific preservation activities reported by participants. In the Hypoxia Case, the first author will deposit her final data with the lab once the article has gone to publication, which at the time of writing this dissertation has not yet occurred. She will also delete intermediary data versions, and discard physical samples no longer of use to anyone in the lab. If she had been reporting genetic data in the publication, she would then deposit that data in Genbank. She did collect genetic data, but they were reported in a complementary publication. In the Webcam Case, the authors ran out of server space and discarded all of the satellite data that they could acquire again. Prioritizing data for preservation is normally considered a preservation activity in and of itself. In the other four cases, the data would remain where they were on personal and lab-shared servers, no specific actions were taken on the part of the researchers to preserve their data.

Life Cycle Model

The data life cycle stages described above fit all six cases studied, with some stages repeated or skipped depending on the nature of the research. For the Fungi Case, the data were collected prior to the development of the research question. As a result the planning, calibration, and collection stages looked different from the other field science and experimental cases. The researchers planned and acquired data, but in these stages the researchers identified the data they could feed use to build the structural equation model and therefore no calibration of equipment was necessary. The Hypoxia Case is also slightly different because the researchers were already collecting data prior to the hypoxic event, and on the fly they were forced to reassess their collection strategy, add equipment, and increase the sampling frequency on the existing equipment. The Fungi and Hypoxia Cases are both a sub-cycle in a much larger data life cycle. The data acquired in the Fungi Case came from a nearly decade-long data collection effort. The data acquired in the Hypoxia Case contributed to the larger collection of data collected at the marina over a similarly long period of time. In the Power and Glider Cases, planning through processing stages repeated themselves three times, like a sub-cycle, because they performed multiple rounds of data collection experiments and simulations. The Stream Case and the Webcam Case map directly to the data life cycle stages.

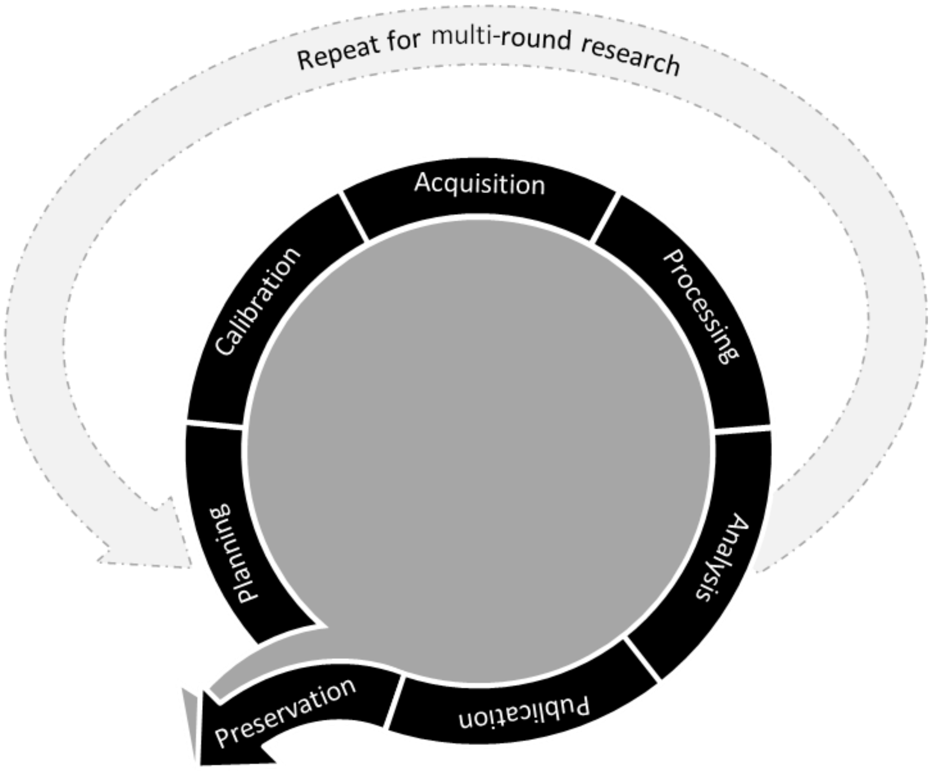


Figure 15: The life cycle of data, from the steps leading to acquisition to preservation.

R1.2 What is the disposition of data at each stage of the data life cycle?

The participants referred to the different versions of data they handled during research. Some common designators used for the types of data are: raw data, processed data, and final data. The raw data come straight from the instruments and are relatively untouched. During processing and analysis many versions of the processed data will be generated by the researchers. The final data are those reported in the publication. A few of the researchers mentioned the data not being "final" until the publication was through peer review and had been published. Until then, there was always the possibility that they would be asked to perform some other processing or analysis or bring in some other data to further demonstrate their findings to the community. When researchers mentioned any preservation activities, they tended to focus on raw and final data as the more valuable datasets to preserve. When these types of data come to the forefront is plotted against the life cycle stages in table 21 below, along with where the data types are located at those stages.

In the Fungi Case, the data collected at the research site were stored on a lab-shared server, and the first and second author would download the data they needed to work with to their local machines. Only after the work was published did they upload their personal copies of the data to the lab-shared server. In the Hypoxia Case, the data were collected to personal machines in the field, then uploaded to a lab-shared ftp server to share between the authors. Whenever an author needed to work with the data they would download the data to their local machine and processing or analysis was performed on the data. Occasionally the data would be uploaded to the server after these changes were made by the authors.

In the Power Case, the data were collected to personal machines and then shared in a DropBox folder, with new versions being added to the folder. In the Glider Case, the first author collected the location data to his email and then deposited them in the lab-shared repository, which parsed the data for him. He then downloaded the data to his personal machine to work with them. According to this author the raw data that had been parsed into the repository would be of interest to other researchers and not the processed and final data on his personal machine, although his personal machine was also backed up to the lab-shared server space and versioned using SVN, just in case.

In the Stream Case, the data were collected onto personal machines in the field, and then uploaded to the lab-shared repository when they got back from the field. The authors would download the data to their personal machines and then upload the changed versions once they were complete. In the Webcam Case, the authors downloaded the webcam and satellite data directly to servers. Whenever any of the authors wanted to perform analysis on a subset of the data, the data of interest were moved to a portable drive and transferred between authors in that way. The data were too cumbersome both in size and accessibility to transfer more easily.

Life Cycle Stages	Data Version	Disposition of Data
Planning	-	Prior to data collection
Equipment Calibration		
Data Acquisition	Raw Data	Raw data are collected to personal machines; uploaded to lab repository, lab server, or shared cloud storage
Data Processing	Processed versions	Raw data downloaded to personal machines, and new versions proliferate; researchers use file names and folder structures to support access; versions may be periodically uploaded to shared lab space
Data Analysis		
Publication	Final	Final data versions are maintained on personal machines, and uploaded to shared lab space.
Preservation	Raw & Final	Subset of researchers discarded intermediary versions from personal machines and shared lab spaces, ensuring the raw and final data were accessible

Table 21: Summary of the disposition of the data at each stage of the data life cycle.

Throughout the data life cycle the data are moved between personal machines and lab-shared spaces, in all six cases. Researchers tended to work with the data on their local machines. The copies uploaded to lab-shared spaces after acquisition, during processing and analysis, and after publication become part of the project record.

R1.3 What data management tasks are observed?

The participants mentioned performing many different tasks performed during the course of the research to make the data fit for the contemporary purpose and that would eventually support data reuse, if someone wanted to reuse the data. As this is the function of data management, I would argue that the following tasks should be considered data management. These tasks “add value” to the data collected. They begin prior to data acquisition and continue through preservation. Of the tasks that fulfilled these criteria, four classes of management tasks were observed in the sample: selection, verification, storage, and documentation. The types and tasks are explained more fully below.

Selection tasks

Selection tasks are similar to archival appraisal, in which a given item is assessed for its inherent value and whether it fits some criteria, and then either selected or discarded depending on the value and fit. Participants described making decisions that evaluated and selected some site, equipment, variable, etc. over another to answer their research questions. Prior to data collection, selection decisions affect what data will be collected, which greatly restricts how the data can be analyzed and what findings are valid. Selection tasks are performed to assure the quality of the final data that would be available for reuse. The first author from Glider Case acknowledged that he could have chosen any number of ways to test his hypothesis, and it was up to him as the first author to evaluate how best to perform the research.

Selecting research site

For all the research cases that had a research site, there was some decision making task that went into the selection thereof. The data collected at the site chosen would need to demonstrate the phenomena being studied, the site needed to be accessible, and other restrictions. In the Fungi Case, the research site was chosen because it was a protected reserve, expensive equipment could be safely left out for years and there were power outlets in otherwise natural settings to support sensing equipment. Other researchers were deploying equipment at the site, so the researchers could take advantage of network effects, such as the collection of micrometeorological data. In the Hypoxia Case, the equipment was deployed at the site because at least one hypoxia event had previously occurred there, thus increasing the likelihood of another event at that location. In the Glider Case, the site was selected in part because the area of the ocean was one where the group had previously worked, and they knew that the area was both accessible from shore and out of the way of shipping lanes that could destroy the glider. In the Stream Case, the researchers had begun the research at a much larger river and decided to scale back to a smaller site. The site was one where they had previously performed research, they had access to the site, and the site had infrastructure that would support their sampling. In the two cases where sites were not selected, either the research site was out of the researchers' control as in the Webcam Case or there was no site for the research as in the Power Case.

Case	Research Site Selection Criteria
Fungi Case	Based on accessibility, safety, infrastructure, and network effects
Hypoxia Case	Based on location of prior hypoxia events
Power Case	No site for research
Glider Case	Based on prior research, accessibility, and out of shipping lanes
Stream Case	Based on prior research, smaller site, accessibility, and infrastructure
Webcam Case	Webcam locations out of their control

Table 22: Summary of site selection and the criteria for site selection by research case.

Selecting appropriate variables, equipment, and sampling frequencies

The methods for studying the phenomena of interest limits the data that can and will be acquired to support the research. Which variables are chosen, which equipment are used, and how frequently the data are collected all affect the power of the research findings. For instance, some pieces of equipment are more expensive, but also more reliable, whereas other equipment is cheaper, but less reliable. The researchers would need to make a decision as to what were their priorities, more coverage with the less reliable equipment or less coverage with more reliable equipment. There is a similar trade-off between battery life of sensing equipment and the frequency of sampling. One of the goals of research performed at CENS was to capture phenomenon at a higher resolution, both spatial and temporal, and in so doing these researchers were able to observe what had previously been unobservable.

All of the cases studied exhibited the selection of variables, equipment, and sampling frequencies. There are disciplinary standards for which variables should be collected to properly capture the studied phenomenon, and which equipment are trusted to capture those variables. Sampling frequencies vary with the temporal scale of the phenomenon of study, and are thus less specific to the discipline and more specific to the phenomenon. A few of the cases do not follow the respective standards of either their discipline or the phenomenon. For instance, in the Glider Case, the glider was chosen by the application scientists as equipment they would like to use provided more efficient sampling algorithms, otherwise the robotics researchers would have chosen a simpler robot on which to test their algorithms.

In the Fungi Case, the variables, equipment, and sampling frequency were selected for the data collection effort that began years prior to the research reported in the Fungi Case publication. For this case there were two variable selection tasks, first when the equipment was installed the variables collected would capture the ecological phenomenon, and second, when developing the structural equation model the authors needed to select only coherent variables. Coherent variables were those sections of data that were collected during the same time period, so that some relationship can be built between them. Like Fungi Case, in the Hypoxia Case, there were two selection tasks, after the first hypoxia event when they installed equipment at the harbor, and then at the beginning of the second hypoxia event when the researchers needed to make decisions about what additional variables could be collected with the reserved equipment, and how to increase the sampling frequency to capture the rapidly evolving situation. In the Webcam Case, the researchers were not working with their own equipment, but they needed to select webcam feeds that would show the green-up phenomenon in order to evaluate the webcam functionality.

Case	Variable Selection Criteria	Equipment Selection Criteria	Sampling Frequency Selection Criteria
Fungi Case	Capture phenomenon; data coherency	Discipline standard and context	Capture phenomenon
Hypoxia Case	Capture phenomenon at greater detail	Standard equipment with additional	Based on prior sampling, increased for temporal density
Power Case	Standard measures	Standard equipment	Standard frequencies, adjusted breadth of sampling
Glider Case	Capture phenomenon	Chosen by research collaborators	Based on prior research, balance safety and time between surfacing
Stream Case	Capture phenomenon	Standard equipment and network	Based on prior research, balancing density with battery life
Webcam Case	Deriving variables that indicate phenomenon	Webcam feeds with visible plants	Based on prior research, and limited by satellite standards

Table 23: Summary of variable, equipment, and sampling frequency selection, as well as criteria for selection of each by research case.

Discarding outliers

Not all data acquired during collection or simulation can be used, some data are errors rather than true measures. Outliers are identified using statistics, sensor fault detection, and other methods. The removal of erroneous outliers results in a higher quality dataset on which analyses can be run. Five of the six cases mentioned removal of outliers in some capacity. In the Fungi Case and Stream Case, sensor faults were discarded. In the Hypoxia Case, statistical outliers were discarded. In the Webcam Case, the data processing method adopted by the authors would occasionally fail on some webcam stream, those streams were then discarded from the final analysis. In the Webcam Case, there was a similar problem where the measurement equipment would occasionally return results that were “wonky” according to the researchers, and the data from that run would be discarded and re-run.

Case	Discarding Outlier Criteria
Fungi Case	Sensor faults
Hypoxia Case	Statistical outliers
Power Case	When equipment gave “wonky” measures
Glider Case	No outliers to remove
Stream Case	Sensor faults
Webcam Case	Where the statistical processing failed

Table 24: Summary of outlier removal and the criteria for discarding outliers by research case.

Selecting data to tell a story

The researchers collected and tasked much more data than was necessary for supporting the findings reported in the publications. Some subset of the available data were selected to "tell a story" in the publication. In the four cases where this selection task was mentioned, each group had collected many more variables, locations, and at a higher spatial density than would be feasible either for analysis or presentation in the publication. In the Stream Case and Power Case, this selection task was not mentioned.

Case	Data Selected to Tell a Story
Fungi Case	Reduced spatial scales to construct model
Hypoxia Case	Only those locations where the hypoxia event occurred
Power Case	Not mentioned
Glider Case	Only the location data collected during experiments and in simulation
Stream Case	Not mentioned
Webcam Case	Subset of webcam streams to demonstrate proof-of-concept

Table 25: Summary of data selection to tell a story and the data selected by research case.

Discarding data no longer of use

Once research was published, the data will likely not be used by the authors for another question. Prior to publication there was a chance that the authors may need to run some analyses again or bring in other collected variables. Data that have been assessed by the researchers as potentially valuable for future use are saved and all others discarded. The criteria of what data are considered valuable varies over the two cases where researchers discarded data. In the Hypoxia Case, there is a lab policy that researchers should evaluate the likelihood of data and samples being used again by other members of the lab or researchers beyond the lab, such as data that is a subset of a long-term data collection effort. In the Webcam Case, the researchers were limited in the amount of available space for maintaining the image data collected. The authors discarded any data that they believed they could acquire again. In the Glider Case, Stream Case, and Power

Case all data are kept regardless of future value, and in the Fungi Case no discard of data after publication was mentioned.

Case	Criteria for Discarding Data No Longer of Use
Fungi Case	Not mentioned
Hypoxia Case	All raw data kept, along with physical samples that members of the lab were interested in using for future research, all other data and samples were discarded
Power Case	Keep everything
Glider Case	Keep everything
Stream Case	Keep everything
Webcam Case	All raw and processed data were discarded if they could be acquired again (ie satellite images) and the task was thoroughly described

Table 26: Summary of data retention after use and the criteria for discarding data by research case.

Discarding intermediary versions

In order to make the remaining data easier to access, intermediary versions produced during the processing stage of the data life cycle are discarded. Those participants who discarded data no longer of value to the research group also discarded intermediary versions of the data. The first author in the Stream Case also mentioned removing intermediary versions from her personal machine to support access to the raw and final data. In this case the intermediary versions are not removed from the shared data repository. In the Glider Case only the raw data are saved to the shared lab repository so there are no intermediary versions to discard. In the Fungi Case and Power Case, the discard of intermediary versions was not mentioned.

Case	When Intermediary Data Versions are Discarded
Fungi Case	Not mentioned
Hypoxia Case	Intermediary versions are discarded before deposit onlab-shared server
Power Case	Not mentioned
Glider Case	Only raw data are initially saved to the shared repository so there are no intermediary versions to discard
Stream Case	Intermediary versions are discarded from personal machines
Webcam Case	Intermediary versions of image data that can be acquired again were discarded

Table 27: Summary of intermediary data version retention and when intermediary data versions were discarded by research case.

Verification Tasks

Researchers do not just assume that the data they collect is "true," they actively verify the data to ensure the veracity of not only the data but how the data are presented in the publication and the conclusions made from the data. Verification may include statistical tasks and group discussions that contribute to the validity of the data. Participants described several points at which verification tasks were employed by the researchers.

Verification of methods

Acquiring data can be a costly task, and researchers make sure that the site, variables, equipment, and sampling frequencies are going to get them the data they need to prove or disprove their hypotheses. In five of the six cases, verification of the research plan was discussed by all the authors interviewed. In the Fungi Case, the first author used statistical tools to determine if the structural equation model was valid. In this case the data had already been collected and the authors were determining if the modeling method was valid rather than the data collection plan.

Case	Verification of Methods
Fungi Case	Statistical validation of the model method
Hypoxia Case	Research plan discussed with all authors
Power Case	Research plan discussed with all authors
Glider Case	Research plan discussed with all authors
Stream Case	Research plan discussed with all authors
Webcam Case	Research plan discussed with all authors

Table 28: Summary of method verification and how the methods were verified by research case.

Equipment calibration

Every piece of equipment has a slightly different offset that needs to be accounted for during the processing stage of the data life cycle in order to arrive at the actual measure of the variable. In the Glider Case, Stream Case, and Hypoxia Case, the researchers calibrated the equipment used while in the lab as per the manufacturer's instructions. In the Power Case the measuring equipment was calibrated prior to purchase and did not need to be calibrated by the researchers.

In the Webcam Case, the public webcams were not under the group's control, so they could not calibrate them. In the Fungi Case, the equipment had been calibrated many years prior, and was not necessary for research reported in the Fungi Case publication.

Case	Equipment Calibration Task
Fungi Case	Equipment calibrated when first installed, not for the research reported
Hypoxia Case	Sensors already in place had been calibrated prior to deployment, additional sensors were calibrated in the lab before deployed
Power Case	Measuring equipment are factory calibrated
Glider Case	Glider is calibrated per manufacturer instructions
Stream Case	Sensors calibrated in the lab prior to deployment
Webcam Case	The equipment was not theirs to calibrate, but offset information was gathered

Table 29: Summary of equipment calibration and method the equipment calibration by research case.

In-field verification

Researchers assess the research as it is proceeding, to ensure that they are collecting data and that the data being collected are what they expected. During the interviews participants mentioned tasks they use to reflexively assess the quality of the data. While observing comparable research in the field, regular quality assessment was demonstrated. In the Glider and Power Cases, the researchers kept tabs on the progress of their experiments as they ran. In the Stream Case, Hypoxia Case, and Webcam Case, the researchers visually inspected numerical data, water samples, and webcam images respectively.

Case	In-field Verification Task and Object
Fungi Case	Not mentioned
Hypoxia Case	Visual inspection of water samples
Power Case	Review of equipment measures captured
Glider Case	Tracking glider during mission to make sure it is going where expected
Stream Case	Visual inspection of data streaming onto personal computers
Webcam Case	Visual inspection of webcam images

Table 30: Summary of data verification during acquisition, what was verified, and how they were verified by research case.

Verification of Collected Data

Whether asked by others or in order to assure themselves of the validity of the results, researchers will occasionally reproduce their results. In two cases this task was mentioned. In the Stream Case, one of the publication reviewers was unsure as to whether an observed effect reported in the publication was actually an observed effect or a sensor variation. In order to appease the reviewer, they performed a bench test comparing the performance of the sensors in question. In the Power Case, the researchers would re-run their measurements when they collected results that seemed invalid.

Case	Verification of Collected Data
Fungi Case	Not mentioned
Hypoxia Case	Not mentioned
Power Case	When measures were “wonky”, measures were re-run
Glider Case	Not mentioned
Stream Case	Sensors were bench-tested at the request of reviewers
Webcam Case	Not mentioned

Table 31: Summary of verification of collected data, when they were verified, and how the data were verified by research case.

Data Calibration

Each datum collected by a piece of equipment contains the offset exhibited by that piece of equipment. During the equipment calibration stage, the equipment offsets were measured so that they could later be removed from the data. The task of calibrating the data transforms the measures into valid phenomenological measures by removing the offsets recorded during equipment calibration or ground-truthing of the equipment in the field.

In three of the cases the calibration of data were mentioned. In the Stream and Hypoxia Cases, the data were calibrated using the recorded offset for each piece of equipment. In the Webcam Case the researchers used position information about each camera, such as latitude, longitude, and altitude to calibrate images for the position of the sun during each of the two daily sampling times throughout the year. In the Power Case, the calibration of the equipment yields data that has already been calibrated, and does not need to be adjusted by the researchers. In the Glider Case, the initial equipment calibration only affects the application science data collected by the glider. The location data reported by the glider cannot be calibrated because there is no way to capture any kind of offset in the ocean. In the Fungi Case, data calibration was not mentioned.

Case	Data Calibration Task
Fungi Case	Not mentioned
Hypoxia Case	Data are calibrated given the recorded equipment offsets
Power Case	The calibration of the equipment returns valid measures, and there is no need for calibrating the data
Glider Case	There was no way to capture an offset, so the data cannot be calibrated
Stream Case	Data are calibrated given the recorded equipment offsets
Webcam Case	Although the equipment is not calibrated, the researchers use the position of the sun and camera position to calibrate images

Table 32: Summary of data calibration, and how the data were calibrated by research case.

Automated tasks for the removal of outliers

When criteria for what constitutes an outlier is verified via prior research or discussion between researchers, then the criteria can be stabilized. When outlier criteria are stable, researchers are able to encode them algorithmically. In three of the cases, participants mentioned the use of automated tasks to identify or remove outliers. In the Fungi Case, the second author had a script that would flag outliers. In the Stream Case, the second author used filters to remove "out-of-bounds" sensor faults. In the Webcam Case, the analysis of each webcam feed would identify feeds that did not task correctly. In the Hypoxia Case and Power Case there was no mention of the automatic removal of outliers, and in the Glider Case, there were no outliers identified.

Case	Automatic Removal of Outliers
Fungi Case	Script that flags statistical outliers
Hypoxia Case	Not mentioned
Power Case	Not mentioned
Glider Case	No outliers to remove
Stream Case	Low-pass and high-pass filters to catch sensor faults
Webcam Case	Statistical tasks identified webcam feeds that did not task correctly

Table 33: Summary of automated outlier removal, what outliers were identified, and how they were identified by research case.

Verification of outliers

What data are considered outliers can be very contentious. In our prior research with the CENS community, we reported cases of collaborating application science researchers and technology researchers debating whether n outlier was a sensor artifact or evidence of a phenomenon that heretofore had been invisible due to the resolution of data capture. In order to mitigate false-positives, the outliers must be verified. In five of the six cases, researchers mentioned discussing the validity of outliers with their co-authors. In the Fungi Case, the first author mentioned that he would discard obvious outliers, and only bring the borderline outliers to the group to be verified. In the Glider Case, the location data could have no outliers, so there was nothing to verify.

Case	Outlier Verification Task
Fungi Case	Borderline outliers were discussed with all authors
Hypoxia Case	Outliers were discussed with all authors
Power Case	The researchers capturing equipment measures discussed wonky-runs between one another
Glider Case	There were no outliers to verify
Stream Case	Outliers were discussed with all authors
Webcam Case	Outliers were discussed with all authors

Table 34: Summary of outlier verification and how outliers were verified by research case.

Verification of data presentation

Visual presentations of the data support their interpretation by others. The development of data presentations serves as a visual check of the data by the researchers. The visualizations are very carefully crafted and will go through frequent revision to clearly and accurately present the results. In four of the six cases the verification of data presentations were mentioned. In the Stream, Hypoxia, Webcam, and Power Cases, the data presentations were verified through discussions involving some or all of the authors. In the Glider Case and Fungi Case, the verification of data presentations was not mentioned.

Case	Verification of Data Presentations
Fungi Case	Not mentioned
Hypoxia Case	Data presentations discussed with all authors
Power Case	Data presentations discussed with all authors
Glider Case	Not mentioned
Stream Case	Data presentations discussed with all authors
Webcam Case	Data presentations discussed with some authors

Table 35: Summary of data presentation verification and how the presentations were verified by research case.

Verification of findings

The findings reported from the results must be supported by the method used and the data collected. Like the data presentations, the findings reported will likely affect how a future user would interpret the data and are the final verification of the research and the data. In five of the six cases, the participants mentioned discussing findings between some or all authors. In the Fungi Case, the verification of findings was not mentioned.

Case	Verification of Findings
Fungi Case	Not mentioned
Hypoxia Case	Research findings discussed with all authors
Power Case	Research findings discussed with all authors
Glider Case	Research findings discussed with all authors
Stream Case	Research findings discussed with all authors
Webcam Case	Research findings discussed between some authors

Table 36: Summary of research finding verification and how the findings were verified by research case.

Storage Tasks

Where the data are kept, who has access to the data, and whether the data are discoverable are all components of long-term integrity. Storage tasks mentioned by participants included the saving data to personal and lab-shared machines, whether machines were being versioned or backed-up, and the file-naming conventions and folder structures they used to support access.

Data copies maintained on personal machines

Maintaining copies of data on personal machines was certainly the most prevalent data storage task. In many of the cases the distribution of data copies allowed co-authors to use the same data in different ways. In the Stream Case, the proliferation of data copies on personal machines during data collection in the field ensured that the data would make it back to the lab to be uploaded to their data repository. In all of the cases the raw data were maintained on somelab-shared server space and downloaded to personal machines for processing and analyses.

Case	Copies of Data Maintained on Personal Machines
Fungi Case	Each author maintained their own copies of data
Hypoxia Case	Everyone maintained their own copies of the data
Power Case	Copies of processed data versions were maintained by each author
Glider Case	The first author maintained raw data sent from the glider to email; copies of processed data versions
Stream Case	Data were saved to personal machines in the field prior to upload to data repository; copies of processed data versions
Webcam Case	Copies of processed data versions were saved to workstations

Table 37: Summary of data copies maintained on personal machines and which copies by research case.

Copies uploaded to lab-shared servers

Lab and department servers act as a clearinghouse for research data and documents, supporting sharing between collaborators. Servers have the added benefit of long-term maintenance and back-up plans. Students in the lab or staff in the department are typically tasked with monitoring, updating, and migrating the contents to new servers when the old servers are no longer current. In all six cases, the researchers uploaded data to alab-shared server. In the Glider, Stream, and Power Cases, data were uploaded to repositories maintained by each group and housed on a lab server, in the other three cases the data were uploaded directly to a lab server.

Case	Copies of Data Uploaded to Lab-Shared Server
Fungi Case	Copies uploaded to lab server
Hypoxia Case	Copies uploaded to a lab, ftp server
Power Case	Shared data repository lives a lab, svn server
Glider Case	Shared data repository lives on a lab server
Stream Case	Shared data repository lives on a departmental server
Webcam Case	Copies uploaded to lab-shared space on CENS servers

Table 38: Summary of data copies maintained on alab-shared server and what type of server by research case.

Copies deposited in lab repository

Data and file repositories are a structured collection of resources that include some metadata fields to make the collection searchable. Unlike a database, the resources are not similar enough to be integrated with one another. In three of the six cases, researchers deposited their data in a lab-shared repository and in another, the data were saved directly to a database built specifically for the research project. In the Glider Case, scripts are used to parse the location data from emails sent by the glider and deposit them in the shared repository along with the metadata collected by the glider during the mission. When the glider returns to shore, the application science data are pulled off the glider and uploaded to the repository. In the Stream Case, the second author deposited the raw data and calibrated versions on the shared lab file repository, along with metadata about the data collected. In the Power Case, raw and processed data versions were uploaded to a software, file, and data repository shared in the sixth author's lab. In the Webcam Case, the image data were collected and indexed in a lab-shared database built for maintaining access to the data for the research project. The database has structure and metadata, but was not as general as a data repository, which serves as a location where data from multiple projects can reside together. In the Fungi and Hypoxia Cases, the labs have no shared data or file repository.

Case	Copies of Data Uploaded to Lab Data Repository
Fungi Case	No shared data repository
Hypoxia Case	No shared data repository
Power Case	Raw data and processed versions were uploaded to a lab, Apache Savan repository
Glider Case	Scripts parse emailed location data directly into the lab data repository, with internal metadata
Stream Case	Raw data, processed versions, and internal metadata were uploaded to a lab, Plone data repository
Webcam Case	Images were uploaded to a lab, MySQL database, but one author was the gatekeeper to the images

Table 39: Summary of data copies maintained in a data repository and what type of data repository by research case.

Copies uploaded to lab cloud space

In addition to the locally shared copies of data, researchers will rely on cloud services to maintain copies of their data. The service mentioned by researchers is DropBox, which allows copies of data to be stored on local machines that are linked to the DropBox account as well as on the DropBox servers. DropBox supports the sharing of folders between accounts. Changes to one local copy will propagate out to the cloud copy as well as any copies on the machines of other users sharing the same file. This service is convenient for individuals to process the data and have the processed available to all the co-authors. Data can be downloaded from a web interface as well as through the folder structure. In the Glider Case and Power Case the researchers used DropBox to maintain shared copies of data and other files. In the Stream Case all the authors used DropBox, but did not use the shared folder functionality. In this case the authors used DropBox as a way to synch their files between their home and lab computers.

Case	Copies of Data Uploaded to Lab Cloud Space
Fungi Case	Not mentioned
Hypoxia Case	Not mentioned
Glider Case	Data and other documents shared with other authors through DropBox
Power Case	Data and other documents shared with other authors through DropBox
Stream Case	Each author maintains their own DropBox space, but not shared
Webcam Case	Not mentioned

Table 40: Summary of data copies maintained on a lab cloud space and what cloud service by research case.

Backing up machines

Long-term maintenance of digital data at the very least requires that machines with data are backed-up regularly, to mitigate the complications that ensue when a machine fails and corrupts the data. In five of the six cases, machines where data were stored were being backed up, and in the other case it was not mentioned, but the lab server is likely backed up. In the Hypoxia Case, the lab PI, the sixth author, requires his students to back-up their personal computers. In the Power and Glider Cases, student system administrators maintain the servers, including backing-up. In the Webcam Case, the third author ran the backing-up of the servers. There was a problem with the backing-up process where it was writing to non-existent disks and data was lost, but they were data that could be retrieved again from MODIS.

Case	Machines are backed-up
Fungi Case	Not mentioned
Hypoxia Case	Server and personal machines are backed-up regularly
Power Case	Server is backed up regularly
Glider Case	Server and personal machine are backed-up regularly
Stream Case	Server is backed up regularly
Webcam Case	Servers are backed-up regularly

Table 41: Summary of machine back-up and which machines by research case.

Data stored with descriptive filenames

Researchers mentioned incorporating metadata into data file names to support access of the data by informing the user what data were contained in the file. Descriptive metadata included the research project, research site, piece of equipment, status of the data, and other information. In four of the six cases, researchers mentioned using descriptive file names. In the Fungi Case, the individual researchers had personal standards for naming files. In the Stream Case and Four, the researchers interviewed described personal standards for naming files that were consistent across the group members. In the Webcam Case, the database and scripts run by the third author named all the data files as they were acquired. The third author designed the database and data files to support access even if the database ceased functioning, precisely through the file naming standard. The other authors from this case each had personal standards for naming files. In the Glider Case, the data repository organization stood in the place of file naming conventions in the support of access to data, and in the Power Case, no file naming conventions were mentioned.

Case	Stored Data with Descriptive Filenames
Fungi Case	Personal standards for file name metadata
Hypoxia Case	De facto lab standard for file name metadata
Power Case	Not mentioned
Glider Case	Not necessary because of data repository organization
Stream Case	De facto lab standard for file name metadata
Webcam Case	Database named image files with metadata; personal standards for file name metadata

Table 42: Summary of descriptive filename use and extent of naming standards by research case.

Folder structures to support access

The organization of files can also support access, both aggregating and differentiating versions of data or data by site. In four of the six cases, folder structures were used to support access to data. The authors interviewed in the Fungi and Stream Cases mentioned using folders to separate various versions of the same data. In the Hypoxia Case, the authors mentioned nesting of the various versions of data, leading to very deeply hierarchical file structures. In the Webcam Case, the database and scripts generated a new folder for each webcam feed, collecting all the images for that feed in the folder. In the Glider and Power Cases, the data repository had structure to support access, making folder structures unnecessary.

Case	Data Store in Folder Structures to Support Access
Fungi Case	Folders are used to separate data versions
Hypoxia Case	Folders are used to nest data versions
Power Case	Not necessary because of data repository organization
Glider Case	Not necessary because of data repository organization
Stream Case	Folders are used to separate data versions
Webcam Case	Database created new folders for each webcam feed

Table 43: Summary of folder structure use and how folder structures were used by research case.

Distinguishing data versions

Descriptive file names informed the user about what was in the file, but the content changes through the various steps in processing the data. In order to distinguish one version of the data from another, file naming standards or a version control system are used. In the Fungi, Hypoxia, Stream, and Webcam Cases, the participants mentioned using datestamps in the file names to distinguish data versions. In the Glider and Power Cases, a version control system, svn, is used to capture snapshots of the files over time, allowing the researchers to return to prior versions.

Case	Distinguishing Data Version to Not Overwrite
Fungi Case	Versions of scripts get datestamped
Hypoxia Case	File versions are datestamped
Power Case	All files versioned using svn
Glider Case	All files are versioned using svn
Stream Case	Lab de facto standard involves datestamps
Webcam Case	File versions are datestamped

Table 44: Summary of file versioning and how files are distinguished by research case.

Maintain server and lab repository

Long-term support of the server and lab repository where data and metadata were deposited must be provided in order for the resources to persist. Servers must be upgraded and content backed-up. In five of the cases, the individuals who maintained the server or repository were not authors on the publication. Only in the Webcam Case was an author in charge of server maintenance. He set back-up schedules and swapped in new machines when old machines broke or ran out of room. In this case the data have been moved to a server maintained by the department.

Case	Server and lab Repository Maintenance
Fungi Case	Research staff member maintains lab server
Hypoxia Case	Server maintained
Power Case	2-3 Graduate students maintain server and repository
Glider Case	2-3 Graduate students maintain server and repository
Stream Case	Server maintained by department
Webcam Case	Servers maintained by third author and department

Table 45: Summary of server and repository maintenance and who performed maintenance by research case.

Data Deposited in Discipline Repository

Storing data on personal machines and lab-shared spaces supports their own access to data, but discipline repositories are a way to make data available to researchers in the discipline. A data repository is a digital library for data, many of which are maintained by discipline or object of study. For instance IRIS is a repository of seismic data, and as mentioned earlier Genbank is a repository of genomic data that supports users across a variety of disciplines. Repositories are trusted sources of vetted data.

All of the participants were specifically asked if they deposited their data in a discipline repository. In none of the cases were data from the research deposited in a discipline repository. Only do the authors from Hypoxia Case ever deposit any data in discipline repositories, but the data reported in the Hypoxia Case publication were not deposited. The reasons provided by members of each case for not depositing data in discipline repositories were as follows. Glider Case authors indicated that deposit of data in repositories was not done by members of their discipline. According to the first author in the Fungi Case, sharing Excel files of data is sufficient, although the format of data sharing is a separate issue from repository use. The researchers from Stream Case indicated that their data were not in the correct format to be included in a hydrology repository, which likely used CUAHSI HIS, the standard format for hydrological data. In the Webcam Case the authors indicated that their data were quite different from those normally collected in Ecology, and as such would not fit in a discipline repository. According to the Power Case authors, the lab-shared repository suffices for their data sharing needs.

Case	Deposit of Data in Discipline Repositories
Fungi Case	Sharing Excel files suffices
Hypoxia Case	Algal genome data deposited in discipline repository, although not for the Hypoxia Case research
Power Case	Shared repository suffices
Glider Case	Not something the discipline does
Stream Case	Their data are not in the hydrology standard format for deposit
Webcam Case	There is no repository that would take the data

Table 46: Summary of data deposited in a discipline repository and if not, why not, by research case.

Documentation Tasks

The archival task of description creates documentation supporting access and interpretation of items. Researchers also capture documentation in a variety of formats and at different points of the acquisition, processing, and analysis phases to support their own access and interpretation of data. The following are tasks of documentation and documentation deposit that ensure future users can access and interpret the data, the ultimate goal of data management.

Annotation in lab/field notebooks

Paper notebooks are used in the field and the lab to capture notes, calibration offsets, initial conditions, tasks, and reflections about the research. These rich resources of annotation have been used for millennia, but can be difficult to access because they are tied to a physical location. In four of the six cases, authors indicated the use of notebooks. They were referred to as field or lab notebooks depending on where they were used. In the Glider, Fungi, and Hypoxia Cases, data acquisition was annotated in notebooks. In the Hypoxia and Webcam Cases, data processing was annotated in notebooks. In the Fungi Case and Power Case none of the authors mentioned annotation in notebooks.

Case	Annotation in lab or field notebooks
Fungi Case	Annotation of data collection efforts
Hypoxia Case	Notebooks used to capture sensor locations during deployments; annotation in the field; annotation in the lab during analysis
Power Case	Not mentioned
Glider Case	Documentation of mission parameters and annotation of mission progression
Stream Case	Not mentioned
Webcam Case	Annotation of which methods worked and which did not work

Table 47: Summary of annotation in notebooks, and what is annotated by research case.

Annotation in separate, digital file

While paper notebooks are extremely functional under a wide variety of conditions, digital files have the advantage of being searchable. In three of the six cases, researchers mentioned capturing annotation in a digital file separate from the data. In the Fungi Case, the first author described a spreadsheet he fills out during data processing and analysis, where he tracked data permutations throughout research. In the Glider Case, mission files were created to capture the initial conditions of the experiment, essentially the settings of how frequently the glider should surface, where the experiment began, etc. In the Webcam Case, the second author indicated that she recorded her methods and notes about methods she tried in a digital notebook in addition to her paper notebook. In the Stream Case and Power Case, no one mentioned annotation in a separate, digital file.

Case	Annotation in a Separate, Digital File
Fungi Case	Meta-analysis spreadsheets captured processing annotation
Hypoxia Case	Did not create separate, digital files to annotate data
Power Case	Not mentioned
Glider Case	Mission file generated for the glider mission with initial conditions
Stream Case	Not mentioned
Webcam Case	Methods and notes captured in digital files

Table 48: Summary of annotation in separate, digital files, and what is annotated by research case.

Filling out a protocol sheet

Protocol sheets clearly delineate protocol steps, and prompt the recording of crucial metadata at specific points during the protocol. In two cases following protocol sheets was mentioned by authors. In the Stream Case, the protocol sheet contained the data collection protocol for the usual sensors and other equipment this group typically deployed. In the Hypoxia Case, protocol sheets were developed to standardize the counting of algal species in physical samples. The sheets listed typical species and spaces to record counts. In the Glider and Fungi Cases, the researchers mentioned not having a protocol sheet. In the Webcam Case, protocol sheets were not mentioned. In the Power Case, the protocol was encoded in the scripts used to run the measurement equipment, so there was no need for a protocol sheet.

Case	Filling Out Protocol Sheets
Fungi Case	Not used
Hypoxia Case	Instructions for counting algal species, with prompts for values
Power Case	Protocols captured in scripts that run measurement equipment
Glider Case	Not used
Stream Case	Instructions for data collection, with prompts for annotation
Webcam Case	Not mentioned

Table 49: Summary of protocol sheet use; when they are used and what documentation is captured by research case.

Ingest of internal metadata with lab-shared data

Some metadata are captured by the equipment and included along with the data. These metadata are ingested along with the data, supporting the access and interpretation of the data. In four of the six cases the researchers mentioned ingesting internal metadata with the data available on shared storage. These are the same four cases where lab repositories have been implemented. The data ingest task prompts ingest of metadata as well.

Case	Ingest of Internal Metadata with Lab-Shared Data
Fungi Case	Not mentioned
Hypoxia Case	Not mentioned
Power Case	Header rows metadata parsed into the repository
Glider Case	Datestamps and geolocations are parsed into the repository along with data
Stream Case	Descriptive metadata deposited along with raw data; inherits to any derivative data products
Webcam Case	Webcam metadata pulled from websites and captured in database

Table 50: Summary of internal metadata ingest, what metadata, and when they are ingested by research case.

Documentation of scripts/code

Scripts and code are used for the acquisition, parsing, processing, and analysis of data. As such they are a remarkable resource for understanding how data were collected and processed throughout the course of research. Scripts and code written by someone else are not always straightforward to interpret. Documentation of scripts and code mitigate that distance. In four of the six cases the documentation of code was mentioned, ranging from minimal documentation, such as an explanation of the variables declared, to documentation of every line of code. In the Hypoxia Case and Power Case, the documentation of code was not mentioned.

Case	Documentation of Scripts and Code
Fungi Case	Minimal header description of script functionality
Hypoxia Case	Not mentioned
Power Case	Not mentioned
Glider Case	Minimal documentation
Stream Case	Varies across individuals; variable description at the beginning to comments on every line
Webcam Case	Minimal documentation, but getting better

Table 51: Summary of script and code documentation, and documentation method by research case.

Including research documentation in publication

The first place most prospective data users are exposed to data are the publication. Thorough documentation of methods, processing, and analyses can support the data users' ability to evaluate a potential data resource. The publication also serves as a starting point for trusting and interpreting the data for the potential user. In all six publications the research was documented in the publication. In the Fungi, Glider, and Stream Cases the research sites and experiments were clearly described. In the Hypoxia Case, the documented research was remarkably thorough, clearly delineating between different types of data collected and how the data were processed and analyzed. In the Webcam Case and Power Case, the documented research had a thin description of the method and tasks applied to data, but in both these cases more thorough documentation existed elsewhere.

Case	Documentation of Research in Publication
Fungi Case	Sites and experiments clearly described
Hypoxia Case	Thorough documentation of data collected and how each data type was tasked
Power Case	Thin description in the publication, but more thoroughly described in companion publication
Glider Case	Sites and experiments clearly described
Stream Case	Sites and experiments clearly described
Webcam Case	Thin description of all the processing used on data, but more thoroughly documented in slides from a method workshop developed by authors

Table 52: Summary of documentation in publication and what was documented by research case.

Including external documentation with lab-shared data

The researchers described three different ways they captured documentation that is not directly linked to the data. These included protocol sheets, digital files, lab, and field notebooks. These sources capture a much greater depth of documentation than is captured in the publication and would greatly assist in data interpretation by other researchers, but only if they are accessible. In five of the six cases the external documentation sources were stored at least in physical proximity to where data were stored in the lab. In the Stream Case and Webcam Case, the external documentation is stored directly with the data in the shared repository and database respectively. In the Glider Case and Fungi Case, the external documentation are stored on the same server as the data, but not necessarily integrated. In the Hypoxia Case, the physical copies of external documentation are stored in the lab. In the sixth case no external documentation was captured, so there was no external documentation to store with the data.

Case	Deposit of External Documentation with Lab-Shared Data
Fungi Case	Meta-analysis spreadsheets were included with the data files on the lab server
Hypoxia Case	Paper copies of notebooks and protocol sheets are deposited in lab binders, but not digitized and included with the data on the server
Power Case	No external documentation captured to include with data
Glider Case	Mission files included on the server, but not specifically with the mission data in the repository
Stream Case	Protocol sheets are digitized and deposited in the repository with the data
Webcam Case	External documentation added to database

Table 53: Summary of external documentation deposit, which external documentation, and where they are deposited by research case.

Including scripts/code with lab-shared data

As mentioned earlier the scripts and code used by researchers to acquire, task, and analyze data contain vital documentation of how the data were handled throughout the research. The deposit of these resources with the data allows them to be maintained and accessible together. In the Stream Case, the scripts were deposited with the data in the repository. In the Power Case, the lab repository is also a code repository, so code in deposited in the repository. Whether the code and data are linked to one another was not mentioned. In the Webcam Case, scripts were stored and maintained with workshop materials developed from the research. In the Glider and Fungi Cases, the scripts and code are maintained on personal computers, along with personal copies of data. In the Hypoxia Case, the deposit of scripts and code was not mentioned.

Case	Deposit of Scripts and Code with Lab-shared Data
Fungi Case	Maintained on personal computer; only provided when asked
Hypoxia Case	Not mentioned
Power Case	Lab repository is also a code repository
Glider Case	Maintained on personal computer; only provided when asked
Stream Case	Deposited in repository with data
Webcam Case	Maintained with method workshop materials

Table 54: Summary of scripts and code deposit with lab-shared data, and method by research case.

R1.4 What data management tasks are associated with each stage?

From the list of selection, verification, storage, and documentation tasks delineated above, we can see that management tasks happen throughout the life cycle. The following table plots the various data management tasks by class along the generalized data Life Cycle Stages mentioned in R1.1. The tasks do not necessarily happen in the order listed within a stage and, as seen above, the tasks performed in each case vary. The numbers in square braces indicates the number of cases in which the task was mentioned or observed.

Life Cycle Stage	Selection Tasks	Verification Tasks	Storage Tasks	Documentation Tasks
Planning	Selecting site/s [4] Selecting appropriate variables, equipment, sampling frequencies [6]	Verification of methods [6]		
Calibration		Equipment calibration [4]		Annotation in lab notebooks [2]
Collection		In-field verification [5]	Saving data to personal machines [6] Deposit of data in lab-shared space (server, repository, cloud storage) [6] Backing up machines [5]	Annotation in separate, digital file [1] Annotation in field notebooks [3] Collection protocol sheet [1] Ingest internal metadata [4]
Processing	Discarding outliers [5]	Data calibration [3] Automated tasks for removal of outliers [3] Verification of outliers [5]	Descriptive file-naming [4] Folder structures to support access [4] Distinguishing data versions [6]	Annotation in lab notebooks [2] Processing protocol sheet [1] Annotation in separate, digital file [2] Documentation of scripts/code [4]
Analysis		Verification of findings [5]		
Publication	Selecting data to tell a story [4]	Verification of data presentation [4]		Including research documentation in publication [6]
Preservation	Discarding samples or data no longer of use [2] Discarding intermediary data versions [3]		Server/repository maintenance [6] Deposit of data in discipline repository [1]	Including external documentation with data [5] Including scripts/code with data [2]

Table 55: Generalized model of data management tasks by type plotted along the data life cycle stages.

Notice that selection tasks happen largely at the beginning and end of the data life cycle. Storage tasks only start when there is something to store and, once initiated, these tasks, such as backing-up machines, persist beyond the stage in which they are listed. Verification and documentation tasks happen more regularly throughout the life cycle, with a different management task for each stage. Verification tasks only need to happen occasionally to assure the research is progressing appropriately; they are essentially an opportunity for the researchers to check in with the research and ensure that it is on track. Documentation tasks proliferate by mode, so for every stage of research there are multiple ways the researchers could be documenting the activities from that stage. For instance, there are three different external documentation tasks mentioned in these cases, all of which could be used to document the data acquisition stage. The earlier documentation tasks affect whether other documentation tasks occur later in the life cycle. For instance, if no external documentation is collected in a notebook, protocol sheet, or external file during the data acquisition stage, then there is no way the researchers could deposit these sources of documentation during the preservation stage.

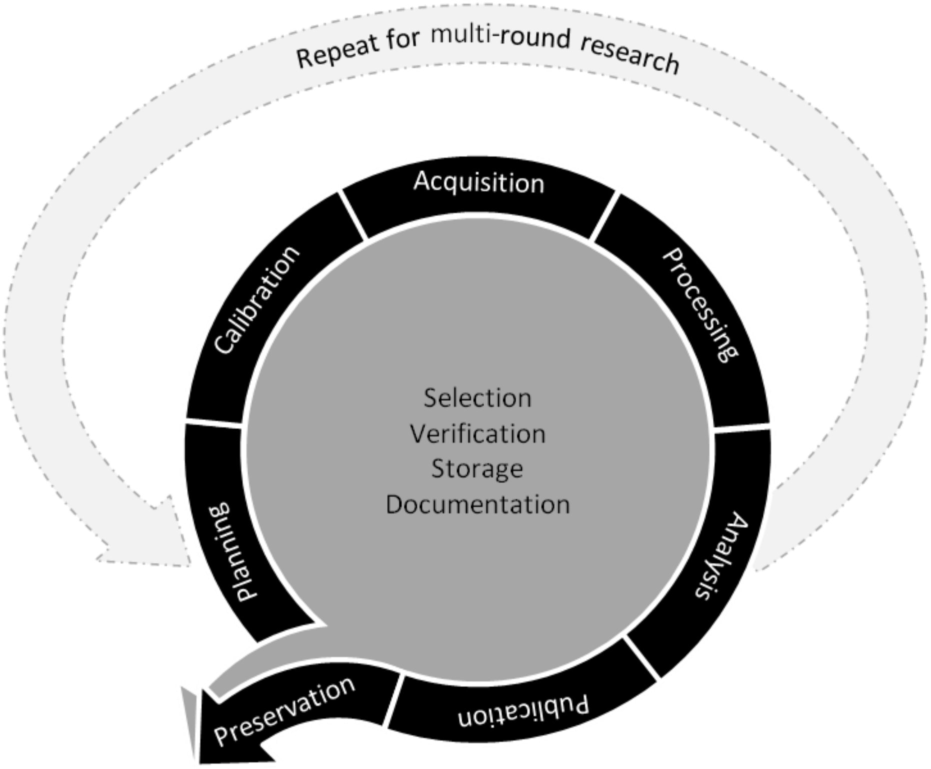


Figure 16: The life cycle of data, from the steps leading to acquisition to preservation, updated to include data management task categories.

R1.5 For whom are the data managed?

What has been described up until this point are the many tasks data are subjected to during scientific research that “add value” to the data. These tasks look very similar to the regular scientific practices that make up performing research, rather than data management. The act of performing research not only results in the creating of scientific knowledge, but also in the creation of data to which value has been added.

Tasks performed by researchers can be grouped into two stages: pre-publication and post-publication. For the research reported in the Glider, Stream, Webcam, and Power Cases, a research question led directly to an experiment design, to data collection, and eventually to a publication. Whereas for longer-term monitoring projects, like the minirhizotron deployment in the Fungi Case or the harbor deployment in the Hypoxia Case, the pre-publication stage lasted many years and the data collected was used to answer multiple research questions. While the pre-publication stage has a variable beginning and a well defined end, the post-publication stage has a well defined beginning and a poorly defined end. This stage can extend from the point of publication to when the student who is first author leaves their PI’s lab or beyond.

Each of the two stages has a different end and means, and a different priority for the research groups. Pre-publication the researchers attempt to make the data usable for the research being reported. Post-publication the researchers make the data usable for future research. The first stage is the most important to the researchers collecting the data, as this is their contemporary purpose, which explains the many data management tasks the participants mentioned during the interviews. This bias could be an artifact of framing the interview in the preparation of the publication. To mitigate the limited view afforded by the publication-frame, the interviews were supplemented with questions about reproducing results five or ten years post-publication, and specifically soliciting management tasks after publication. Very few of the participants mentioned any tasks being performed after publication, but for those who did there was a different strategy than in preparation of publication. The researchers selected data for disposition rather than for acquisition, to make those data in the collection easily locatable. For physical samples, researchers were under space restrictions that are no longer an issue with digital data.

Participants were asked if they would be able to come back to the data in five or ten years, and be able to use them again. Everyone who answered the question indicated that they would be able to do so in the short term, but on a longer timescale there was more variability. The first author from the Fungi Case, specifically generates his "meta analysis spreadsheets" so that he can come back to the data and know what has been done to them. The other participant from the Fungi Case indicated that a member of the lab could easily understand the data, but someone from outside the lab would find there was not enough documentation to be able to access and use the data. The second author from the Hypoxia Case was confident that she could replicate her portion of the results five years later.

The participants from the Power Case were positive that the work could be reproduced, even a decade later. The shared repository has been running for a decade now, and they have shared data from the very beginning. According to the PI, the repository was built to support reproducibility. One participant from the Glider Case indicated that he could easily recreate his work using his scripts and that the data are in stable storage. The other participant from the Glider Case does not see access and documentation of the data to be an issue. He indicated that rapidly changing tools would become a problem after three or four years.

The fourth author from the Stream Case indicated the lab repository captured enough metadata that it would be possible to come back to the data. The one piece missing for someone who had not participated in the research would be a better understanding of the research site, which would need to be more detailed than was captured in the publication. Two of the authors from the Webcam Case indicated that the results could be reproduced, but one of the authors would need to access the data for the others. The first author thoroughly documented his tasks and scripts because he taught a workshop on the method. The third author acted as gatekeeper to the database andlab-shared server, and in order to interface with the webcam data that had been collected, he would need to act as the interface.

Case	Participant	Conditions/Limitations	How Long?
Fungi	1st Author (Lead)	Meta-analysis spreadsheets generated to support reuse	-
	2nd Author	Enough documentation for someone from lab	-
Hypoxia	1st Author (Lead)	-	
	2nd Author	She could come back to her own data	5 yrs
	6th Author (PI)	-	
Power	1st Author (Lead)	Lab repository has already supported reuse	10 yrs
	6th Author (PI)	Lab repository built for reuse	-
Glider	1st Author (Lead)	Scripts and data are stored stably	-
	6th Author (PI)	Tools become a limitation	2-3 yrs
Stream	1st Author (Lead)	-	
	2nd Author	-	
	3rd Author	-	
	4th Author (PI)	Lab repository would support reuse for someone familiar with the research site	-
Webcam	1st Author (Lead)	Documentation and scripts would support reuse	-
	2nd Author	-	
	3rd Author	Data is reusable, but only with his assistance	-

Table 56: Summary of future data usability by participant by case.

R2 How are data management tasks distributed among members of the research group?

How data management tasks were distributed and who performed what data management tasks, are different questions. The following section breaks these two questions apart, first presenting who performed which data management tasks, and then how the research group members came to perform the tasks they performed.

R2.1 Who performed which data management tasks?

The same data management tasks were not performed by everyone. Some tasks were performed by a single author, some were performed by all the authors, and some by people beyond the author group. In plotting who performed which data management task by case, as they are below, patterns of how data management tasks are performed emerge.

Fungi Case

In the Fungi Case the data management tasks were relatively evenly distributed between all three authors, and a few tasks were performed by members of the research lab not included in the author list. The first and second authors split the data handling, so that the second author handled the data that were collected from the field site, and the first author handled the structural equation modeling. This split is visible in the data management tasks they each performed. The second author deployed the data collection equipment and while the equipment collected data, developed scripts to calibrate and clean the data. In the reported research, he performed most of the data processing, including calibration and removing outliers, and documenting his scripts. The first author performed all of the analyses to fit the data into the reported model, he developed the method, documented the task, and statistically verified the method. The third author, as the head of the lab, was occasionally brought in to consult on the research and how it was proceeding. The third author has staff members who handle the research equipment and server maintenance, so in depositing data and documentation the first and second authors are passing off the long-term maintenance of those resources to the staff members.

Life Cycle Stage	Data Tasks Performed by Authors	1st	2nd	3rd	Beyond
Planning	Select research site			X	
	Select appropriate variables		X	X	
	Select appropriate equipment		X	X	
	Select appropriate sampling frequencies	X	X		
	Verify Methods	X			
Calibration	Equipment calibration		X	X	X
	Annotation in lab notebooks				
Acquisition	In-field verification	X		X	
	Data copies saved to personal machine	X	X		
	Data copies saved to lab-shared space	X			
	Backing-up machines				X
	Annotation in separate, digital file				
	Annotation in field notebook		X		
	Collection protocol sheet				
	Ingest internal metadata		X		
Processing	Discarding outliers		X		
	Data calibration		X		
	Automated tasks for outlier removal		X		
	Verification of outliers	X	X	X	
	Descriptive filenaming	X	X		
	Folder structure to support access		X		
	Creating separate versions of files to differentiate		X		
	Annotation in lab notebook	X	X		
	Processing protocol sheet				
	Annotation in separate, digital file	X			
	Documentation of scripts/code		X		
Analysis	Verification of findings				
Publication	Selecting data to tell a story	X			
	Verification of data presentation				
	Including research documentation in publication	X		X	
Preservation	Discarding samples or data no longer of use				
	Discarding intermediary versions of data				
	Server/repository maintenance				X
	Deposit of data in discipline repository				
	Including external documentation with data	X			
	Including scripts/code with data				

Table 57: Matrix of which author or individual beyond the author group performed which data management task during the research reported in the Fungi Case publication.

Hypoxia Case

In the Hypoxia Case the data management tasks were mainly performed by the members of the sixth author's lab. For the initial selection and verification tasks, authors across both labs participated. Towards the end of the data life cycle the first author performed tasks on her own in the preparation of final data for publication. She participated in all of the data management tasks, except the site selection, which happened before her involvement. Like the Glider and Fungi

Cases, the lab maintains its own server and it was unclear who specifically maintained the server. Within the lab, binders of protocol sheets are kept.

Life Cycle Stage	Data Tasks Performed by Authors	1st	2nd	3rd	4th	5th	6th	Beyond
Planning	Select research site		X			X	X	X
	Select appropriate variables	X				X	X	X
	Select appropriate equipment	X			X	X	X	X
	Select appropriate sampling frequencies	X				X	X	X
	Verify Methods	X	X	X	X	X	X	
Calibration	Equipment calibration	X	X	X				X
	Annotation in lab notebooks	X	X	X				X
Acquisition	In-field verification	X					X	
	Data copies saved to personal machine	X	X	X				
	Data copies saved to lab-shared space	X	X	X				
	Backing-up machines							X
	Annotation in separate, digital file							
	Annotation in field notebook	X						
	Collection protocol sheet							
	Ingest internal metadata	X	X	X				
Processing	Discarding outliers	X	X	X				
	Data calibration	X	X	X				
	Automated tasks for outlier removal							
	Verification of outliers	X	X	X			X	
	Descriptive filenaming	X	X	X				
	Folder structure to support access	X	X	X				
	Creating separate versions of files to differentiate	X	X	X				
	Annotation in lab notebook	X	X	X				
	Processing protocol sheet	X	X	X				
	Annotation in separate, digital file							
	Documentation of scripts/code							
Analysis	Verification of findings	X	X	X	X	X	X	
Publication	Selecting data to tell a story	X						
	Verification of data presentation	X	X	X			X	
	Including research documentation in publication	X						
Preservation	Discarding samples or data no longer of use	X						
	Discarding intermediary versions of data	X						
	Server/repository maintenance							X
	Deposit of data in discipline repository							
	Including external documentation with data	X	X	X				
	Including scripts/code with data							

Table 58: Matrix of which author or individual beyond the author group performed which data management task during the research reported in the Hypoxia Case publication.

Power Case

In the Power Case, there were a number of stages of research delegated to authors, depending on their area of expertise. The fourth author used the measurement equipment, the second and third authors used the software and software performance measures, and the first author was involved in almost everything. The project was relatively new at the time, and the faculty members were involved in all the planning and final publication stages, to ensure validity. Like most of the other cases, the repository and server were maintained by other members of the lab beyond the author group.

Life Cycle Stage	Data Tasks Performed by Authors	1st	2nd	3rd	4th	5th	6th	Beyond
Planning	Select research site							
	Select appropriate variables	X	X	X	X	X	X	
	Select appropriate equipment							
	Select appropriate sampling frequencies	X			X			
	Verify Methods	X	X	X	X	X	X	
Calibration	Equipment calibration	X			X			
	Annotation in lab notebooks							
Acquisition	In-field verification	X			X			
	Data copies saved to personal machine	X	X	X	X			
	Data copies saved to lab-shared space	X	X	X				
	Backing-up machines							X
	Annotation in separate, digital file							
	Annotation in field notebook							
	Collection protocol sheet							
	Ingest internal metadata							
Processing	Discarding outliers	X			X			
	Data calibration				X			
	Automated tasks for outlier removal							
	Verification of outliers	X			X			
	Descriptive filenaming	X	X	X				
	Folder structure to support access	X	X	X				
	Creating separate versions of files to differentiate							
	Annotation in lab notebook							
	Processing protocol sheet							
	Annotation in separate, digital file							
	Documentation of scripts/code							
Analysis	Verification of findings	X	X	X	X	X	X	
Publication	Selecting data to tell a story							
	Verification of data presentation	X	X	X	X	X	X	
	Including research documentation in publication	X	X	X	X	X	X	
Preservation	Discarding samples or data no longer of use							
	Discarding intermediary versions of data							
	Server/repository maintenance							X
	Deposit of data in discipline repository							
	Including external documentation with data	X						
	Including scripts/code with data	X	X	X				

Table 59: Matrix of which author or individual beyond the author group performed which data management task during the research reported in the Power Case publication.

Glider Case

Glider Case was characterized by the first author being the main contributor. He performed the majority of the data management tasks, because he was the only one handling the data in any capacity. He worked with the head of his lab for initial selection tasks. He consulted with the authors beyond the sixth author’s lab for verification tasks. In this case, the lab-shared repository and server were run by other students from the lab and the automated ingest scripts they developed were used to deposit the raw data and internal metadata in the repository.

Life Cycle Stage	Data Tasks Performed by Authors	1st	2nd	3rd	4th	5th	6th	Beyond
Planning	Select research site	X	X					
	Select appropriate variables	X	X					
	Select appropriate equipment							
	Select appropriate sampling frequencies	X	X					
	Verify Methods	X	X	X	X	X	X	
Calibration	Equipment calibration							
	Annotation in lab notebooks							
Acquisition	In-field verification	X						
	Data copies saved to personal machine	X						
	Data copies saved to lab-shared space							X
	Backing-up machines							X
	Annotation in separate, digital file	X						
	Annotation in field notebook							
	Collection protocol sheet							
	Ingest internal metadata							X
Processing	Discarding outliers							
	Data calibration							
	Automated tasks for outlier removal							
	Verification of outliers							
	Descriptive filenaming							
	Folder structure to support access							
	Creating separate versions of files to differentiate	X						X
	Annotation in lab notebook	X						
	Processing protocol sheet							
	Annotation in separate, digital file							
	Documentation of scripts/code	X						
Analysis	Verification of findings	X	X	X	X	X	X	
Publication	Selecting data to tell a story	X						
	Verification of data presentation							
	Including research documentation in publication	X						
Preservation	Discarding samples or data no longer of use							
	Discarding intermediary versions of data							
	Server/repository maintenance							
	Deposit of data in discipline repository							
	Including external documentation with data	X						
	Including scripts/code with data							

Table 60: Matrix of which author or individual beyond the author group performed which data management task during the research reported in the Glider Case publication.

Stream Case

The authors from Stream Case worked closely together throughout the research, but they also had well-defined roles, which can be seen in the figure below. For tasks where the group needed to select or verify, the whole group participated. Otherwise, the second author, in handling the processing stage, performed the data management tasks associated with that stage. He calibrated the data, deposited data and internal metadata in their shared repository, and discarded outliers. Similarly, the first author handled the majority of the analysis and the writing of the publication, so she performed the data management tasks associated with those stages. The server was maintained by the department, rather than someone specific to the lab.

Life Cycle Stage	Data Tasks Performed by Authors	1st	2nd	3rd	4th	Beyond
Planning	Select research site	X	X	X	X	
	Select appropriate variables	X			X	
	Select appropriate equipment	X		X	X	
	Select appropriate sampling frequencies	X	X	X	X	
	Verify Methods	X	X	X	X	
Calibration	Equipment calibration	X				
	Annotation in lab notebooks					
Acquisition	In-field verification	X	X	X	X	
	Data copies saved to personal machine	X	X	X	X	
	Data copies saved to lab-shared space		X			
	Backing-up machines					X
	Annotation in separate, digital file					
	Annotation in field notebook					
	Collection protocol sheet	X				
	Ingest internal metadata		X			
Processing	Discarding outliers					
	Data calibration		X			
	Automated tasks for outlier removal		X			
	Verification of outliers	X	X	X	X	
	Descriptive filenaming	X	X	X		
	Folder structure to support access	X	X			
	Creating separate versions of files to differentiate	X				
	Annotation in lab notebook					
	Processing protocol sheet					
	Annotation in separate, digital file					
	Documentation of scripts/code	X	X			
Analysis	Verification of findings	X	X	X	X	
Publication	Selecting data to tell a story					
	Verification of data presentation	X	X	X	X	
	Including research documentation in publication	X				
Preservation	Discarding samples or data no longer of use					
	Discarding intermediary versions of data					
	Server/repository maintenance					X
	Deposit of data in discipline repository					
	Including external documentation with data	X				
	Including scripts/code with data	X	X			

Table 61: Matrix of which author or individual beyond the author group performed which data management task during the research reported in the Stream Case publication.

Webcam Case

The first three authors of Webcam Case worked closely with one another, only bringing the other authors in for verification of the research plan. The first and second authors worked in parallel on two different data sources that were compared during analysis. This parallel work is reflected in the data management tasks they performed, which were roughly the same throughout the data life cycle. The third author, as technical staff performed data management tasks that relied on his expertise, and at the time of writing this dissertation, he continues to collect data and maintain the server.

Life Cycle Stage	Data Tasks Performed by Authors	1st	2nd	3rd	4th	5th	Beyond
Planning	Select research site	X	X				
	Select appropriate variables	X	X				
	Select appropriate equipment						
	Select appropriate sampling frequencies	X	X	X			
	Verify Methods	X	X	X	X	X	
Calibration	Equipment calibration						
	Annotation in lab notebooks						
Acquisition	In-field verification	X	X				
	Data copies saved to personal machine	X	X				
	Data copies saved to lab-shared space			X			
	Backing-up machines			X			X
	Annotation in separate, digital file						
	Annotation in field notebook						
	Collection protocol sheet						
	Ingest internal metadata		X	X			
Processing	Discarding outliers	X	X	X			
	Data calibration			X			
	Automated tasks for outlier removal			X			
	Verification of outliers	X	X				
	Descriptive filenaming	X	X	X			
	Folder structure to support access			X			
	Creating separate versions of files to differentiate	X	X	X			
	Annotation in lab notebook		X				
	Processing protocol sheet						
	Annotation in separate, digital file	X	X				
	Documentation of scripts/code	X					
Analysis	Verification of findings	X	X				
Publication	Selecting data to tell a story	X	X				
	Verification of data presentation	X	X				
	Including research documentation in publication	X	X				
Preservation	Discarding samples or data no longer of use	X					
	Discarding intermediary versions of data	X					
	Server/repository maintenance			X			X
	Deposit of data in discipline repository						
	Including external documentation with data			X			
	Including scripts/code with data	X	X				

Table 62: Matrix of which author or individual beyond the author group performed which data management task during the research reported in the Webcam Case publication.

R2.2 How were roles distributed?

For all of the participants, when asked how tasks were distributed among the co-authors, indicated that distribution happened organically, and were more understood than assigned. The question was so diffuse that the researchers could not give an answer that would reflect the specifics of everyday interactions that ended with person A performing task 1, and person B performing task 2, etc. Some patterns emerge through answers to questions about author contribution and the data practices from planning to publication,.

The tasks varied by how many people were required to accomplish them. Some data management tasks were typically performed by the group and some were performed by individuals. Many of the storage and documentation task were performed by an individual or a few people, but many of the planning and verification tasks were performed by the group. From a high-level, this observation holds true, but there are changes in the number of people performing each task throughout the life cycle. Selection tasks begin as group tasks and become more solitary as the research progresses. Verification tasks began as a group task and ended as a group task, but in the middle of the data life cycle they were more solitary. At publication, storage tasks transition from tasks performed by individuals to tasks performed by those beyond the author group. Documentation tasks are handled throughout the data life cycle exclusively by the non-faculty authors.

Trends in task distribution

There were three trends observed in the distribution of data management tasks among research group members: which members were from the publication discipline, who took ownership of the research problem, and the authors' area expertise.

Matching the publication discipline

As mentioned in the prior chapter, the publication from each research group could only be submitted to a journal of any one discipline at a time, regardless of whether more than one discipline was represented in the research group. Data acquired as part of the research may have been meaningful to all of the group members, but was presented in the context of a prior literature and using methods from one discipline. Those group members from the publication discipline handled the data, as they were trained in that context and those methods. Co-authors representing other disciplines mainly contributed in an advisory capacity.

The majority of the data management tasks were performed by the co-authors whose discipline matched the discipline of the publication. In the Power, Glider, Hypoxia, and Webcam Cases this included anyone from the last author's lab, as the last author was the PI from the publication discipline. In the Fungi and Stream Cases, all of the authors are from the same lab, and the discipline of all the authors matches the discipline of the publication.

Ownership of the research problem

The author who "took ownership of the problem," was the person who led the research, including developing methods, acquiring data, soliciting assistance from other authors, and the writing of the publication. In all six cases, the person who owned the problem was positioned as the first author, but how they become the lead author varies. In the Glider and Webcam Cases, the person who posed the research problem was the de facto owner of the problem. In the Fungi, Stream, and Hypoxia Cases, the problem was posed by the group, and a person became the lead by taking ownership of the problem. In the Power Case, the phenomenon of study was so new that all authors were significantly involved throughout the research, in order to define the research trajectory they will follow over the next four years on this project. In this case, the lead author participated in the most stages of the initial research, so was the lead author for at least this publication.

Area of expertise

The granularity of area of expertise varies depending on single or multi-disciplinary research groups. In multi-disciplinary research groups, the area of expertise is the discipline of practice.

For instance, in the Hypoxia Case the robotics researchers were only part of the initial equipment selection, because they knew the technology. Robotics researchers are unaware of the data handling and analysis appropriate for marine biology research, and as such they did not participate in many of the data management tasks associated with those processes. In the Power Case, rounds of research were broken up by whether they fell within electrical engineering or computer science, and the researchers who participated in each fell along those lines. The Glider Case is the mirror image of the Hypoxia Case; the marine biologists participated in the research design and some verification, but none of the data handling and analysis that would lead to data management tasks.

In cases where the co-authors all come from a single discipline, the area of expertise was more fine-grained. In single-discipline groups the area of expertise may be tied to a role, such as graduate student researcher or technical staff, or to the skill-sets of the individuals. Across all cases, the area of expertise of the faculty PI was the broader view of the research field and the ability to evaluate the research. Because of their role, the PIs participated in all of the selection and verification discussions, but did not handle research data and did not perform data management tasks related to storage and documentation.

Demonstrating the Trends by Case

The three trends described above are difficult to tease apart, and without having been a part of the research it is difficult to know which factor came first, or to identify any causal relationships. Below, I will demonstrate how the set of trends can be used to explain who performed which data management tasks.

The Fungi Case was a single-discipline group, so the area of expertise was tied to who knew what. The first author knew the modeling method. He performed the data management tasks associated with analysis, publication, and preservation while modeling the ecological system. The first author may have come up with the research plan, implemented the model, and written up the publication, but he relied heavily on the expertise of the second author. The second author knew the data. He was intimately connected with the data collection effort from the initial equipment deployment. The second author performed many of the data management tasks associated with planning, acquisition, and processing. The third author, as the PI, consulted on some of the selection and verification tasks.

The Hypoxia Case was a multi-disciplinary research group, so the area of expertise was tied to discipline, with those authors who were members of the publication discipline, marine biology, contributing more to the data management than those from the collaborating robotics lab. The first author took ownership of the problem, and had been working at the research site, with the equipment since she joined the research group in 2006. She participated in every data management task, because she was involved in all of the data handling. The second author began studying the harbor before the first author, and was specifically familiar with the analyses surrounding community dynamics. The data management tasks she performed were consistent with her involvement during data acquisition and analysis of community dynamics. The fourth and fifth authors, as the collaborating robotics researchers, were consulted for the initial equipment deployment and on-going maintenance. The participation in data management of the fifth and sixth authors, as PIs, was limited to initial selection tasks and verification tasks throughout.

The Power Case was a multi-disciplinary research group, so the area of expertise was tied to the discipline of each member. The fourth author knew the measurement equipment, so he was responsible for the measurement stage and the attendant data management tasks. The second and third authors were familiar with simulation and testing of software performance, so they were responsible for that stage and the attendant data management tasks. The first author was involved in all of the stages of research, in addition to developing the software solution presented in the publication. Also as the lead author, he performed the majority of the data management tasks. As PIs, the fifth and sixth authors only participated in initial selection tasks and verification tasks throughout.

In the Glider Case, as a multi-discipline research group, the area of expertise is tied to discipline. The first and sixth authors knew the robotics, the second and third authors knew the ocean model, and the fourth and fifth authors knew the marine biology. The discipline of the publication is robotics, so the first and sixth authors performed the majority of the data management tasks. The first author was the lead author and, as the only member of the publication domain who was also not a PI, he performed the majority of the data handling during research and the attendant data management tasks. The sixth author is the PI, and only performed initial selection tasks and some verification tasks throughout.

The Stream Case was a single-discipline research group, so the area of expertise was tied to the specific skill sets of the members. For the first author, the publication was part of her dissertation research, so her area of expertise was the research problem. She was the lead author, because she had taken ownership of a problem generated by the group, and participated in data management tasks throughout the research. The second author was described as a “whiz” at the statistical package R and graphing. He handled the data during the acquisition and processing stages, and performed the accompanying data management tasks. The third author was in charge of the equipment, so his contribution was the deployment of the equipment. The data management tasks he performed were those surrounding the equipment, such as selecting and calibrating equipment prior to data collection. The fourth author, as PI, helped with the initial selection and verification tasks, as well as some data collection.

The Webcam Case was a multi-disciplinary research group, so the area of expertise was tied to the respective disciplines of the authors. The first, second, and fifth authors were from ecology, and third and fourth authors were from computer science. The first and second authors split the ecology work between them, performing many data management tasks throughout. The third author performed tech support for both of the first two authors, and the data management tasks he performed surrounded the server and database. The fourth and fifth authors were both PIs and their interactions were limited to planning and verification of the research.

R3 For what data management tasks do researchers perceive they are responsible?

In prior sections, the data management tasks researchers performed during research were explored. The following addresses what researchers perceive they are responsible for with regards to their data. This section describes who the researchers indicated were responsible for the data, what that responsibility entails, whether these responsibilities were fulfilled, to whose standard they were fulfilled, and who held whom responsible.

R3.1 Who is responsible for data and what does responsibility entail?

When asked who was responsible for the data produced in each case, four parties of were identified as responsible: the lead author, the PI, all of the authors, and the lab server administrator. The majority of the authors interviewed were either the lead author or the lab PI, both of which were the answers frequently provided by the participants. Seven of the participants indicated that the lead author was responsible for the data, of these participants four were lead authors. Five of the participants indicated that the PI was responsible for the data, of which three were PIs. In total, nine of the sixteen participants indicated that at least they themselves would be responsible for the data. Three of the participants saw the responsibility changing hands, starting with the lead author and then transferring to either the PI or the lab administrators.

In four of the cases, the authors interviewed gave conflicting answers as to who was responsible for the data. In the Stream Case, the first three authors indicated the lead author was responsible, but the fourth author indicated the PI was responsible. Notice that in this case the first author and PI both claimed that they themselves were responsible. In the Power Case, the authors interviewed appeared to abdicate responsibility to one another. The first author indicated the PI as responsible, and the PI indicated the lead author as responsible. In the Fungi Case, the first author identified the lead author as responsible, and the second author indicated the lab administrator was responsible. In the Glider Case, the first author indicated the lab administrator was responsible for the data, and the PI indicated he was responsible for the data.

In the other two cases, the authors were in slightly better agreement as to who was responsible for data. In the Webcam Case, all three authors interviewed thought they were each responsible in some part. In the Hypoxia Case, the first author and the PI are agreed that the lead author was initially responsible and that responsibility transferred to the PI over time. The second author offered herself as responsible for her part, which indicates that she believed all of the authors were responsible for their own parts of the data handling process.

Case	Participant	Lead	PI	All Authors	Lab Admin	Self
Fungi	1st Author (Lead)	X				X
	2nd Author				X	
Hypoxia	1st Author (Lead)	X	X			X
	2nd Author			X		X
	6th Author (PI)	X	X			X
Power	1st Author (Lead)		X			
	6th Author (PI)	X			X	
Glider	1st Author (Lead)				X	
	6th Author (PI)		X			X
Stream	1st Author (Lead)	X				X
	2nd Author	X				
	3rd Author	X				
	4th Author (PI)		X			X
Webcam	1st Author (Lead)			X		X
	2nd Author			X		X
	3rd Author			X		X
Total Mentions		**7**	**5**	**4**	**3**	
Total Self-mentions		**4**	**3**	**3**	**-**	

Table 63: Summary of who was mentioned as responsible by each participant by case.

"Responsibility for data" is a vague concept, so to clarify, participants were asked what responsibility entailed after they indicated who was responsible for data. The tasks mentioned were as follows, from most to least number of mentions: data storage, documentation, data veracity, backing-up storage locations, maintenance, dissemination of the data, answering questions about the data, data quality, accessibility of the data, dissemination of the results, and appropriate methods. The participants indicated between one and four tasks entailed in responsibility. On average, each participant indicated two tasks.

Uniting the parties indicated as responsible with all the tasks entailed in responsibility, yields the matrix found in table 63 below. The tasks entailed in responsibility vary by the party indicated as responsible. Some tasks are only applied to certain types of responsible parties and certain responsible parties are indicated as responsible for a subset of the tasks. The seven participants who indicated that the lead author was responsible for data indicated that responsibility entailed a set of tasks that as the lead author they are in the position to perform, such as documentation and dissemination of data. The three participants who indicated that the lab administrators were responsible for data, also indicated that responsibility entailed storage, backing-up machines, and long-term storage. All of these tasks are in line with the expected role of a server or repository administrator. The four participants who indicated that all authors were responsible for data, also indicated that responsibility entailed a set of diffuse tasks.

Responsibility Entails	Lead(7)	PI (5)	All Authors (4)	Lab Admins (3)	Total mentions
Data storage	5	3	2	3	11
Documentation	5	2			6
Data veracity	3		1		3
Backing-up	1	1		2	3
Maintenance	1	2		1	3
Dissemination of data	2				2
Answering questions		1	1		2
Data quality		1	1		2
Accessibility			2		2
Dissemination of results			1		1
Appropriate methods			1		1

Table 64: The tasks each author mentioned as being entailed in responsibility, for whom they were entailed, and the total number of times they were mentioned.

For a given task, the above matrix is likely to display overlaps between who are indicated as the responsible parties: storage is the responsibility of everyone indicated; documentation is the responsibility of the lead author and the PI; etc. This can be explained by the timing of the tasks, and the imprecision of language. If we split this out by time, and unravel the implications we see that these parties are actually responsible for different aspects of the same responsibility task. As an example, data storage is a task applied to all of the possible responsible parties. The lead author and other authors were responsible for storing data during the research, and putting the data in longer-term storage post-publication. The lab administrator was in charge of maintaining the server where the data has been stored by one or more of the authors. The PI was responsible for purchasing servers, and hiring or designating a lab administrator to maintain the servers. Responsibility for storage thus has different meanings depending on to whom the term is applied and when the storage is happening with relation to publication. In cases where participants gave conflicting answers of who was responsible the conflicts disappear when the variation in interpreting responsibility is taken into account.

Let us now return to Power Case, where the first author and the PI appeared to abdicate responsibility to one another. Here the first author and PI both indicated that the other was responsible for the storage and maintenance of the data. In this case the participants referred to different aspects of storage and maintenance in time. The PI indicated that the lead author was responsible for storing the data and that the students he has appointed as lab administrators are responsible for maintaining the server. The first author indicated that ultimately the PI is responsible for storage and maintenance, which was fulfilled through setting lab policies and designating students to act as lab administrator to maintain the server. With this clarification of how responsibility was interpreted by each participant, we see that the participants were not abdicating responsibility, but considering responsibility over different time-scales.

R3.2 Were the responsibilities fulfilled and to what standard?

Participants were asked whether they had written a data management plan for grant proposals submitted to the NSF, to prime them for a series of questions about data management. When asked if they had fulfilled their data management responsibilities the majority of the participants affirmed that they had. Seven of the participants were confident that they were fulfilling their responsibilities. Four were less confident about their fulfillment, or as one participant described his fulfillment, “okay, not great” (Webcam Case, First Author). Five of the participants did not think they had fulfilled their data management responsibilities.

Case	Participant	Positive	Neutral	Negative	Role
Fungi	1st Author (Lead)	X			
	2nd Author			X	
Hypoxia	1st Author (Lead)		X		
	2nd Author		X		
	6th Author (PI)	X			
Power	1st Author (Lead)	X			
	6th Author (PI)	X			
Glider	1st Author (Lead)			X	
	6th Author (PI)	X			
Stream	1st Author (Lead)			X	
	2nd Author			X	
	3rd Author			X	
	4th Author (PI)		X		
Webcam	1st Author (Lead)		X		
	2nd Author	X			
	3rd Author	X			
Total		**7**	**4**	**5**	

Table 65: Assessment of data responsibility fulfillment by participant by case.

The participants were then asked to compare their data management practices to their peers. Participants who were PIs brought up their experiences on NSF review panels where they had read the data management plans of some of their peers. Eleven of the sixteen participants provided some comparison of their data management practices, as seen in table 66 below. Six of the participants indicated that their data management practices were on par with their peers, three of these participants indicated that they were doing better than their peers. It is important to note that all six of these participants were either PIs or recent PhD graduates, who have a broader perspective of the other researchers’ practices.

Case	Participant	Comparison	Better	On par	Not well
Fungi	1st Author (Lead)	Long-tail		X	
	2nd Author	Big Sci			X
Hypoxia	1st Author (Lead)	Long-tail		X	
		Big Sci			X
	2nd Author	-			
	6th Author (PI)	-			
Power	1st Author (Lead)	-			
	6th Author (PI)	Long-tail	X		
Glider	1st Author (Lead)	Big Sci			X
	6th Author (PI)	Long-tail		X	
Stream	1st Author (Lead)	Big Sci			X
	2nd Author	Big Sci			X
	3rd Author	Big Sci			X
	4th Author (PI)	Long-tail	X		
		Big Sci			X
Webcam	1st Author (Lead)	Long-tail	X		
		Big Sci			X
	2nd Author	-			
	3rd Author	-			
Total		**Long-tail**	**3**	**3**	
		Big Sci			**8**

Table 66: Assessment of data management fulfillment in comparison with long-tail peers and big science organizations.

Eight of the participants, including three who compared themselves to their long-tail peers, compared their data management practices to those of observatory, big science-type research project. In all of these comparisons, the participants indicated that they were not performing up to the standard set by these projects. In the Glider Case, one of the authors compared his performance to that of a marine biology research station, where researchers use the same gliders, generate documentation through a workflow tool, and maintain data for each run in a an integrated database. The research station has a dedicated data management team, and there is very little variation in the runs performed each day. This comparison was unfair to the Glider Case author, because his runs vary significantly, and as a result the documentation and workflows cannot be reused from one run to another. In the Fungi Case, one of the authors compares his performance to the data management of NASA's Earth Observation Satellite data, specifically referencing the standard description of processing levels. Like the author in the Glider Case, this was not a fair comparison. All four of the authors in the Stream Case compared themselves to a forestry observatory that was run by other members of the lab. In the Hypoxia Case, one of the authors compared her performance to a large oceanographic repository. She was the only participant to point out that this comparison was not fair, and explained that her work was messier, since she was not performing standard oceanographic surveys that would even fit into the repository. In the Webcam Case, one of the authors compared his performance to researchers who use workflow tools. He has tried to use them, but found they did not support his workflow, because he needed to change the workflow from experiment to experiment, where the benefit derived from workflow tools comes from the ability to reuse the same workflow. It is important to note that the all five participants who indicated that they had not fulfilled their data management responsibilities in table 65, were comparing themselves to big-science projects.

R3.3 Who holds whom responsible?

Like distribution, participants indicated that everyone holds themselves responsible. In answering other questions, participants brought up lab policies that encouraged responsible data management and were upheld by the PIs. In the Power, Glider, and Stream Cases, lab repositories were maintained and lab policies encouraged deposit of data in the lab repositories. In the Power Case, the sixth author implemented an automated prompt that pops up when his students submit monthly progress reports, to encourages them to upload their data to the lab repository. In the Glider Case, scripts parsed data broadcast from the AUV, meaning that data transferred directly into the lab repository without much human effort. In the Stream Case, the lab repository served not only as a record of the research data, but a means for active sharing between the group members. There was a policy in place for deposit in the lab repository, but the repository used was chosen specifically because of the low barrier to deposit, the drag-and-drop interface makes it easy for the members of the lab to deposit their data.

In the Hypoxia Case, lab policies required copies of data and documentation be left with the lab. Students from this lab were expected to select data worth long-term maintenance and to discard extraneous, intermediary versions. Notebooks and protocol sheets were photocopied and stored in binders. The second author mentioned that when they find a wonky data point, they are able to pull down a protocol binder and look at the context for that data point. Digital data were stored on a shared lab server, and physical data resided in refrigerators and freezers. All three authors interviewed from this case mentioned lab policies that apply to documentation and discarding data, and two of the authors mentioned the slap on the wrist students received when not complying with the lab policies.

CHAPTER 7: DISCUSSION

The following is a brief summary of the prior two results chapters, and discussion of major findings from the research.

Summary of Results

In the last two chapters, six cases were examined. A research publication provided a frame to view the practices of each group. The research reported in the publication fell across a variety of domains, from computer science, electrical engineering, and robotics, to ecology and marine biology. Despite the variety of disciplines, all of the research performed fell within the purview of CENS, a research center focused on the development of embedded networked sensing equipment. The cases are ordered by whether the findings reported were of interest to a science discipline, technology discipline, or some mix of science and technology disciplines.

Authors from each group were interviewed about the publication, the data, and their data practices. The results present the data practices as reported by the lead author, and one or more additional authors. For each of these publications, the lead author, the author who had taken ownership of the research, occupied the first author position. In the Glider Case the lead was a postdoc, in the Webcam Case the lead was a research staff member, and in the other four cases the lead authors were students. The PI from the publication discipline occupied the last author position, and all other contributors were list in order of contribution between the first and last authors. This author ordering is common for experimental research, but by no means universal for academic publishing.

To provide context for each case, the disciplines of each researcher and the relationships between researchers in each group were mapped. Each group was largely formed from a subset of one research lab, including the lab PI, who was positioned as the last author in each case. For those cases where interdisciplinary collaborators were involved, the group would include one or two representatives from another lab. Representatives would always include a PI and possibly another person from that PI's lab.

The data collected and generated by each group were described. Although none of the data have been shared with data users beyond the research group at this point, there was some perceived reusability value for the data collected in each case. Researchers indicated that other members of their own field would be interested in using the data or that the data would be useful in other contexts, such as environmental monitoring by state agencies or reference datasets for testing methods.

The division of labor among the contributors was described for each case. The researchers from each case reported that there was no formal allocation of tasks, instead the distribution happened organically. Several factors associated with the division of labor were: position in the author order, such as first author or last author, and area of expertise. For all of the cases, except the Power Case, the researchers had worked together before, so they already had an understanding of who would do what. In the Power Case, although the authors had not yet all worked together, subsets of the authors had worked together prior to the research.

By comparing the six cases, a model of the data life cycle was developed. The model was general enough to describe the data handling stages from five disciplines, and a very large variety of data and data types. The only variation across cases was the relationship between the cycle of data described in the publication and data life cycles encompassing a longer-term data collection effort, and how many time the first five stages, from Planning to Analysis, were cycled through before progressing on to Publication and Preservation. Both of the technology research cases went through three rounds of the first five stages, whereas the application science and the mixed science and technology cases went through a single round. The purely sciences cases were sub-cycles of a much longer-term data collection effort, and either pulled from existing data or contributed back to a longitudinal data collection.

Where the data resided at each stage of the data life cycle was mapped, using the descriptions of data handling. From this map the proliferation of data versions can easily be discerned. Multiple people worked on copies of the same dataset and very little removal of intermediary versions might make finding important data difficult in the future. The researchers value the raw data over processed versions, because they could always run the same scripts again to recreate the processed versions.

A taxonomy of data management tasks was constructed from accounts of how the data were handled, answers to specific questions about selection, verification, storage, and documentation practices. Researchers were asked about these categories of tasks based on prior research, how earlier interview participants answered questions, and reflexive adjustments made to the interview instrument. These categories were then used to organize tasks based on the function they served: selecting for quality, verifying for veracity, storage for accessibility, and documentation for interpretability. The task categories held across the six cases.

The data management tasks performed by researchers were similar across all six cases, and were grouped into generalized data management functions based on the outcome of the tasks. The functions were fulfilled using different methods so long as the outcome of performance was the same. Tasks were performed using social processes, such as discussing the results to verify the findings. Similarly, tasks were performed using highly technical processes, such as the use of scripts to automatically parse and ingest data to a lab repository. Each group reported performing some subset of the tasks, but the tasks listed may not be exhaustive of all of the data management performed by members of the groups.

The generalized data management tasks were linked to the stages of the data life cycle during which they occurred. This demonstrated that researchers performed data management tasks throughout the data life cycle. Some variation was observed in the distribution of data management tasks through the life cycle between categories of data management tasks.

At different stages of the life cycle, data are being managed for different users. As would be assumed, the researchers were managing the data for their own use during the majority of the data life cycle. After the results from the data were published, the user that would benefit from the data management tasks shifted beyond the authors to potential future users. For instance, final data were deposited with the lab, so that others in the lab could use them, and documentation was stored with the data to support interpretation by someone who had not necessarily been a part of the data collection. This is not to say that a future user would not benefit from the tasks performed prior to publication, rather that these tasks were not performed with the future user in mind.

The data management tasks were performed by different people. Some tasks were performed by a single author, by a few authors, by all of the authors, and by people who were beyond the author list. Using the descriptions of individual contributions and the tasks performed, a task distribution matrix was constructed for each case. From these matrices, who performed which task was identified.

Patterns were first identified by data management task category. Selection tasks were performed by the entire author group at the beginning and by individuals towards the end of the life cycle. Verification tasks were performed by the whole group or by some subset. Storage tasks were performed by individuals from the author group as well as beyond the author group. Documentation tasks happened throughout by those handling the data, as they handled the data.

Patterns were also identified by who performed which tasks, specifically by the author position and area of expertise. The lead author performed the majority of the tasks. The PI from the publication discipline consulted on all of the group selection and verification tasks. PIs from other disciplines only consulted on the group selection and verification tasks during the Planning stage and sometimes during the Publication stage. Other students and staff performed research tasks according to their area of expertise, and as a result performed the data management tasks that accompanied their data handling.

Researchers were solicited for parties they believed to be responsible for the data, and opinions on what that responsibility entailed. Four parties were identified as responsible: all authors, the first author, the PI, and lab server administrators. Of the responsibilities, some are data management tasks, such as data storage and documentation. Other responsibilities listed were not related to data management, such as dissemination of results. On the whole, the reported distribution of data management responsibilities is not as strong as the reported participation in data management tasks captured in the earlier sections. There is some indication of the transference of data management responsibility from those in the author group to parties beyond the author group post-publication, such as technical support staff or students who have been designated as lab administrators.

Researchers were also solicited for whether they believed they were fulfilling their responsibilities. When comparing themselves to their peers, the researchers assessed their fulfillment as on par or better. When comparing themselves to much larger-scale observatories and repositories, the researchers were assessed their fulfillment as poor.

Discussion of Findings

The following findings are significant contributions to the data management literature. These findings open up the black box of science, and make the data management contributions of the data producers more visible.

Data are diverse

Treating data as a uniform category in policy oversimplifies the diversity of data collected and how they transform over time. Within these six cases a wide array of data and data types have been identified. The research reported in both the Fungi and Webcam Cases fell within the discipline of ecology, yet both cases reported radically different data being collected and generated. The research reported in the Glider and Hypoxia Cases had overlapping authors, and again radically different data were being collected and generated in both cases. The research reported in the Stream and Hypoxia Cases were both studying water phenomena, but as Ribes and Finholt (2007) noted in studying environmental engineers and hydrologists there were few overlapping variables collected by these groups, due more to the nature of what data could be collected within an aqueous environment than to some common phenomenological interest.

In these cases the raw data and final data were identified as resources that were valuable to maintain. The proliferation of data versions during processing complicates access. As we have previously observed with simulated data, there is little value in maintaining processed data versions if they can be derived again, as long as the raw data and the data processes are preserved. Not all raw data need to be maintained indefinitely. Physical samples are no longer necessary to keep once validated through peer review by the larger scientific community. Raw observation data can be discarded as long as they can be downloaded again, as seen with the satellite data in the Webcam Case. Arguably the satellite data were not raw from the perspective of NASA, where the data have moved through multiple levels of processing before being made available as MODIS products, but the data were raw from the perspective of the research group.

In these six cases, three of the four types of data identified in the Long-lived Digital Data Collections report (2005) are specifically acquired by the researchers: observational, computational, and experimental data. Observational data were collected by all six groups, and include variables such as dissolved oxygen, chlorophyll, and temperature. The raw observational data and perhaps the final data that come from these data are maintained. Computational data were generated in the Power and Glider Cases, and included responses to simulated environmental input and generating simulated glider paths. The Fungi, Stream, and Webcam Cases also rely on modeling to work with their data, which produced coefficients that are reported in the publications themselves. In all five of these case the raw data, besides the MODIS data from the Webcam Case, that served as model inputs and the scripts/code used to run the models are maintained. Experimental data was collected in the Power, Glider, Stream and Webcam Cases. In the Power and Glider Cases, the equipment and algorithms were being tested. In the Stream and Webcam Cases the methods were being tested. Because most of this data is derived, there is little priority to maintain them, but they are none the less. I would argue that records are also being created by some of the groups, such as the protocol sheets from the Hypoxia and Stream Cases and the meta-analysis spreadsheet generated in the Fungi Case. These records capture the various experiments performed by the research group over time if viewed as a corpus, but serve only as metadata. Records are very important and are maintained by groups and individuals. The Long-lived Digital Data Collections typology is difficult to apply to the varied research that is performed even within a single publication. There are very blurry boundaries between simulated, experimental, and observational data, especially in the two technology cases, and between experimental and observational data in the mixed science and technology cases.

This report also distinguishes between levels of data collections: research, resource, and reference collections (CODATA-CENDI Forum on the National Science Board Report on Long-Lived Digital Data Collections, 2005). The collections maintained by each of the research labs and in some cases the lab collaborations fit within the definition of research collections. They are maintained by small communities and have local standards which may not conform to broader standards. In the Stream Case the authors even mention that their data could not be put in a domain repository, a type of resource collection, because their data did not conform to the hydrologic data standard. The use of local description methods are common to environmental sciences (Estrin, Michener & Bonito, 2003; Zimmerman, 2003; 2007; Zimmerman & Nardi, 2006). Given the way their data come to them off the instruments, without being marked-up automatically in the hydrologic data standard, it is not a surprise that they do not conform to the standard. There would need to be some intermediary process that put the data into the standard for them that would make use of the standard feasible. As Cragin and Shankar (2006) note the complexities of these research collections are obscured by so broad a classification. For the Stream Case, the researchers would like to be able to move up to contributing to a resource collection, but the infrastructure is not there to support them. In other cases, such as the Power and Glider Cases, the current research collection is sufficient for their needs and the needs of those who ask for their data.

The research groups from each case are already experiencing the data deluge (Hey & Trefethen, 2003). In the Fungi, Hypoxia, Stream, and Webcam Cases, the groups are using new forms of instrumentation, embedded networked sensors, and capturing significantly more data than their peers outside of CENS. With these data they were able to develop new methods, as in the Fungi, Stream, and Webcam Cases, and capture a rare phenomenon with enough depth to overthrow the dominant theory on fish kills, in the Hypoxia Case. What was not captured in this round of interviews that had been seen in prior interview rounds, was the problem of having too much data for their existing tools and methods. During the last four years they have developed new tools and methods that not only allow them to deal with the data deluge, but to start "playing" with the data, as one author from the Fungi Case described. In the Power Case no one ever mentioned the volume of the data collected being particularly big, and in the Glider Case, the researchers were incorporating big data into the algorithms being tested.

Although they are collecting large quantities of data that require new methods to work with, these would not yet be considered "big data" (Lynch, 2008). The data collected for the technology and mixed science and technology cases lack a spatial and temporal coverage that would tip them into the big data category. Lynch argues that through proper data management these data could be aggregated with other data to become "big."

Despite the potential utility of the data collected by these groups as described by Heidorn (2008), there is a real question of how valuable these data would be for reuse. Because they are stored in ad hoc data structures, with semi-standard metadata, they would not lend themselves well to being included in data-driven research (Hey, Tansley & Tolle, 2009). The question remains whether this is because the data are themselves not valuable or that the lack of data standards surrounding the data storage and mark-up render potentially valuable data invaluable. The data collected during the long-term deployments described in both the Fungi and Hypoxia Cases could be considered big, and have already been shared beyond these research groups as reported in (Wallis et al., in progress).

The life cycle of data is generalizable

The data life cycle presented here is flexible enough to describe research performed in all six cases, which span diverse data, six disciplines, and at least four institutions. At the same time, the model provides enough structure that variation in data and research practices can be systematically identified. Two cases displayed a significantly different structure than the others; the two technology cases required multiple rounds of the first five life cycle stages, while in the other cases the researchers went straight through the life cycle.

The information resource life cycle model from Levitan (1982) was developed for the describing information resources generally and had five stages that added-value to the resource for users: generation, institutionalization, maintenance, enhancement, and dissemination. Although scientific research data are an information resource, these stages do not map to the data life cycle model presented here. Generation maps to the planning, equipment calibration, and acquisition stages, dissemination maps to publication, and enhancement is most likely data processing. The other two stages map better to the data management tasks. Institutionalization can be considered the process of selection and validation, where the data become truth representations of the phenomenon being studied. Maintenance can be considered all of the storage and documentation that are recorded to support access.

Prior data life cycle models developed from observations of CENS research were focused on collaborative application science and technology sensor deployments (Wallis, Borgman et al., 2008; Wallis, Pepe et al., 2008). At the time, collaborative sensor deployments were a dominant mode of research within CENS, but this is no longer true. The collaboration came together to adapt extant instrumentation to improve science or to design and construct unprecedented instruments, to use the Shrum et al (2007) classification. Even in the collaborative cases, the collaborators are working independently of one another, and we are seeing far fewer dependencies (Borgman et al., forthcoming).

Development of more stable technologies has allowed the application scientists to collect data without the need for technology researchers to participate as actively in the research. Technology researchers are now so well acquainted with the science applications for which they are developing technology that the application scientists may act in an advisory capacity, rather than as full research collaborators. As a result the research across the cases is more diverse than the collaborative sensor deployments, and the data life cycle model presented here was generalized to apply more broadly.

The data life cycle model presented begins before the acquisition of the data, during the Planning stage, rather than at the point of collection, like the Lee, Tibbo et al (2007) model. To use an analogy, this life cycle begins at the conception of data rather than at the birth of data. As we have seen in prior work from Mayernik et al (2011) regarding metadata creation, important contextual information is created during the Planning stage that will become useful for future interpretation. I would argue that there are other tasks performed during this life cycle stage that shape the eventual data product, and therefore deserve attention.

Collins (1998) identified the aftermath of data collection is a most interesting time to study collaborative data practices, because this was is the time when participants come together to explore the meaning of data and phenomena. The results from this research confirm that there are various group interactions occurring after data collection, such as verification of outliers and findings. The group data practices that occur before data collection, when the research group comes together to plan the research, would be another interesting time to study collaborative data practices. Across all the cases, all of the authors came together during the Planning stage to develop and verify the methods. In the interdisciplinary cases, the verification of methods during the Planning stage was the one time when all of the authors would participate. The period during the drafting of the publication, when researchers are verifying their findings and developing data presentations, would be another interesting time to study collaborative data practices.

Researchers are actively managing their data

The NSF Data Management Plan requirement was put into place to encourage the "dissemination and sharing of research results" (NSF, 2011a; 2011b). This description of data management does not address the support necessary to encourage dissemination and sharing, this support must be inferred. In order to support dissemination and sharing the data must by accessible, i.e., locatable and intact, and documented to support interpretation. Storage and documentation are what many researchers associate with "data management" and this comes out in what researchers believed data responsibility entailed. I would also argue that a way to encourage sharing is through the collection of the highest quality data possible, using selection and verification. Selection and verification processes ensure that data are "fit for contemporary purpose," one of the main functions of data curation as defined by Lord and MacDonald (2003). The researchers were observed here performing many tasks while handling their data that contribute to their ability to use the data. These tasks could also be leveraged to support dissemination and sharing.

Prior research of data sharing in the CENS community has demonstrated that researchers are sharing their data, but not through deposit in repositories, instead they conduct personal interactions with the data user (Wallis et al., in progress). The data management practices observed in this community support this method of sharing. If they can find the data to replicate their results then they will be able to share the data beyond the group. Many of the researchers' storage and metadata practices specifically supported this need. Removal of intermediary versions, use of folder structures, including documentation and scripts with the data, and many other tasks all support data access and use beyond the contemporary use.

The data management practices described here are very similar to those best practices set out for ecologists by Cook, Olson et al (2001) in figure 1, specifically the use of descriptive file names, file organization, quality assurance, and documentation, which are four of the seven recommended best practices. The other best practices that were not observed in this dissertation are the use of consistent and stable file formats, defining parameters, and assigning descriptive data set titles. Though these best practices were not observed, this does not mean that they are not performed. I did not ask about the file formats used, whether they defined parameters, or the format of titles applied to datasets. These should be solicited in future research.

Of the data management best practices set out by Martín and Ballard (2010) for the monitoring of bird and diversity data, in figure 2, roughly half were observed in this dissertation. The Hypoxia and Power cases both discuss lab "data policies." In all of the cases, all of the authors ensure "data quality" as a group. The "data documentation and organization" best practices, such as file names and metadata were observed. Of the "data life-cycle control practices," database design and maintenance, as well as data storage and archiving were observed. The practices under "longevity and use" fall outside of the temporal window observed here, and I would need to follow these groups for a longer time period to determine data security, access, dissemination, and publishing. The other best practices not observed such as understood roles and responsibilities, use of data standards, and running data audits provide other challenges to the groups under study. For all six cases, the research varies enough within each case that data standards would be an issue, and similarly, running a data audit may not be useful if the data cannot be integrated from data research project to research project. The need for understood roles and responsibilities is something to be addressed by this dissertation.

The best practices set out in Whitlock (2011) were encouraging the use of domain repositories, which have only been adopted by one of research groups, although not for the research reported in the Hypoxia Case publication. Other groups have not found domain repositories, either because they are not looking or because they have yet to find one to suit their needs. At the same time these groups are not relying on personal websites to act as a long-term data archive, which Whitlock must warn against. These groups exhibit more sophisticated use of servers, lab-shared repositories and data bases, and routine back-up and maintenance handled by someone hired or appointed to perform these tasks.

Unlike the tension between short and long-term concerns identified by Karasti, Baker & Halkola (2006), where the people were concerned about technological solutions, data volume, and metadata in the short-term and scientific inquiry and data sharing in the long-term, the short and long-term tension here is between making data useful for scientific inquiry in the short-term and data sharing in long-term. The data management tasks were mainly performed by themselves for themselves, essentially managing data for the contemporary use. The product of these data management tasks would still support the needs of others, by increasing data quality, veracity, accessibility, and interpretability. As prior studies (Borgman, 2004; 2007) have pointed out, existing data practices could be leveraged to support the needs of a broader set of data re-users. With a better understanding of what tasks are common and which vary by discipline, and the dominant modes of task performance, we could determine which tasks can be encouraged, supported, or even automated to support future leverage of these practices.

There were some potential complications for data reuse. As noted in Kanfer et al (2000), data are deeply embedded in tacit knowledge and local practices that make them difficult to extract. In the Fungi Case one of the authors acknowledged the difference in data management necessary for individuals from the group and individuals beyond the group. He mentioned that the metadata they currently capture is sufficient for someone who knows the site and equipment, but beyond their research group much more depth would be necessary to support reuse. Similarly, in the Stream Case one of the authors mentioned he would need to provide more site description information to support interpretation by anyone who intended to use their data. In the Hypoxia Case, the metadata is stored in the lab in physical binders which support local access, but add another layer of complication to remote data access.

The researchers also admit they would have a hard time understanding the data well enough to reproduce their results 5-10 years later. As found by Mayernik et al (2011), implicit metadata knowledge disperses over time along with the individuals. In some cases, the people themselves would be necessary for locating or using the data, and they may have left the lab. To use Cole’s (2008) term, the data from these groups have not yet differentiated, and as a result these data are not yet mobile.

Data management tasks happen throughout the life cycle

Tasks are performed to ensure the quality and veracity of the data even before the data are collected or produced. Tasks continue throughout the stages of the data life cycle, and culminate in the publications of the research. After publication, there are some tasks that are performed to specifically support future access to the data.

The Digital Curation Centre life cycle model (Higgins, 2008) introduced in Chapter Two describes the processes of curation on the part of the curator. The only role of the data producer within this context is to hand their data off to the curator. The science leading up to the deposit of data are treated like a black box in the curation life cycle model. The entire data life cycle model presented in the results occurs in the first stage of the DCC, Conceptualize. The DCC model only takes into account the data curation work performed by the data curators or librarians. Given how much data management work goes on during the data life cycle, I would argue that data producers have a much bigger contribution to the overall management of research data than is acknowledged in the DCC model, where the data management work of the data producers is all but invisible to the data curators.

The digital resource lifecycle from Lee, Tibbo et al (2007) includes the creation and use activities, as well as the curation activities applied to the digital resource. The data life cycle presented in the results overlaps with the first three stages of their seven stage model, seen in figure 4, pre-creation designing and planning, creation, and primary use. Transfer to archives, archive, transfer copies, and secondary use environment have no place in the science researcher's practices. The data life cycle model presented in the results almost entirely ignores the external curation aspect, because these researchers are performing minimal specific preservation activities, and are not handing their depositing their data with curators. Where is the middle ground between the view of the data producer and the data curator?

Ideally we want both the data producers and the data curators to be seeing the same data curation life cycle, including the curation activities performed by both groups. Like Lee, Tibbo et al (2007), figure 17 shows a more complete picture of what could be happening to the data if the data were being deposited in a data repository. The model below unites the data life cycle including data management tasks performed by the data producers with the digital curation life cycle including digital curation activities performed by the data curators. Both parties are visible, as are their work.

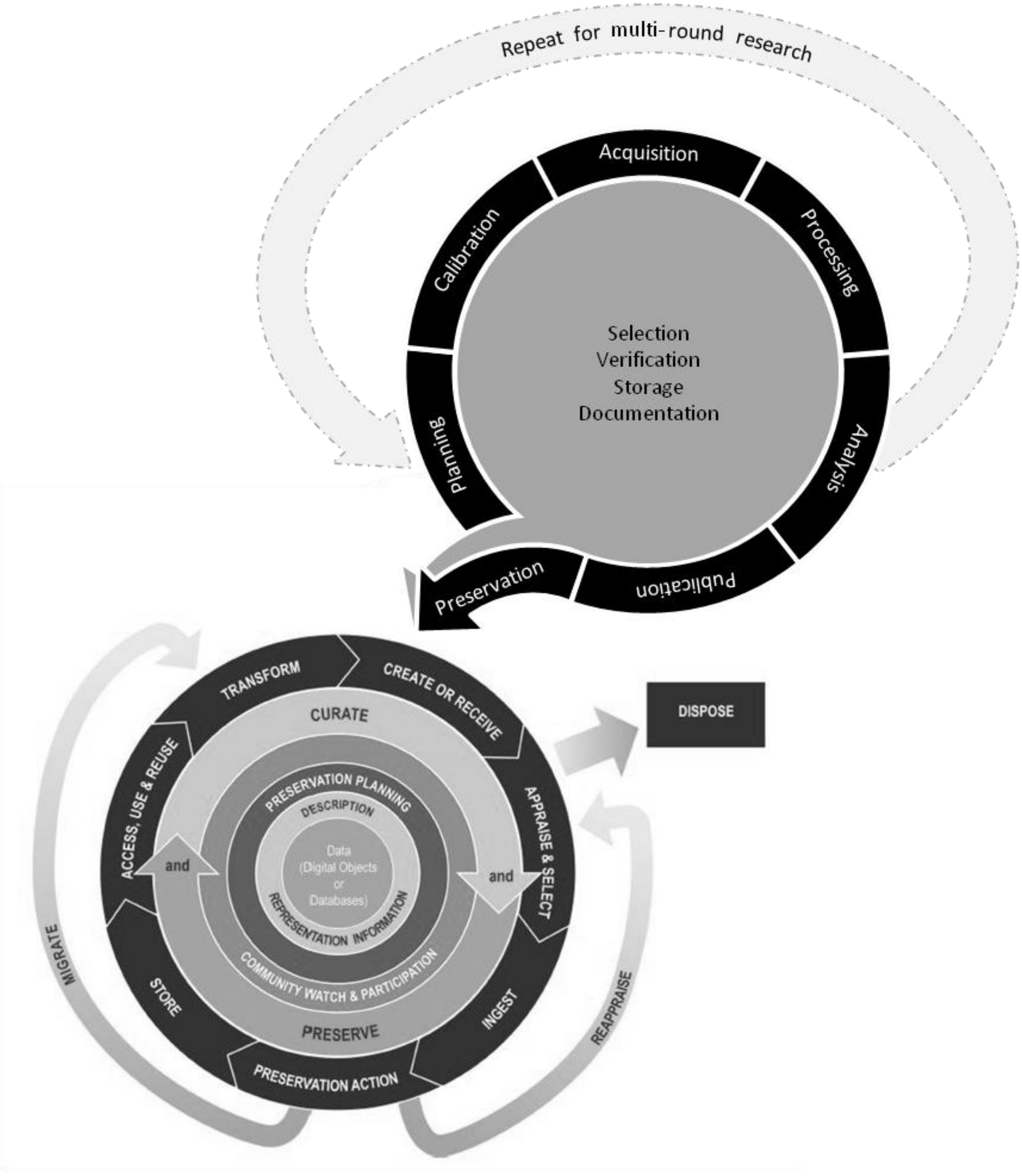


Figure 17: The ideal CENS data curation life cycle, an integration of the data life cycle with the digital curation lifecycle from Higgins (2008).

Researchers not depositing their data is a long-standing issue plaguing digital libraries, data repositories, and other data curation efforts, the “if we build it, they won’t come” problem. There is a critical point after data acquisition, when researchers from all cases were described depositing their raw data in a shared lab server or repository. The introduction of standardized local servers and data repositories, could support the deposit of the raw data in a manner amenable to aggregation.

The Baker, Millerand et al system of data curation sub-cycles (2009), embraces this leverage point. In their model, just after researchers collect the raw data they are deposited with the data manager team employed by the LTER site. These data managers perform the necessary data curation activities on the raw data, while other parties are processing and analyzing the data.

Data management tasks are distributed within a research group

The three data management stakeholder reports presented in the literature review each had a category for the data producer, and tied data management responsibilities to that stakeholder. In the Wellcome Trust Report (Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility, 2003), the data producer was responsible for depositing their data in a data repository. None of the data producers interviewed here mentioned this as their responsibility, except the marine biologists from the Hypoxia Case. These data producers have data that they deposit in Genbank, the repository about which the report was written. Although for this case, they had no genetic sequence data they needed to deposit.

In the UKOLN Report (Lyon, 2007), the data producers were responsible for: managing data for the life of the project, meeting standards for good practice, comply with funder/institutional data policies, and work up data for use by others. The first two of these encompass the majority of the data management tasks listed in the results, and are performed for the research itself, but have benefits for data users. The third responsibility only comes into relevance in the Hypoxia Case, where the researchers are required to deposit genetic sequence data at Genbank, the data repository of the NIH, a major research funder. The fourth responsibility is not something that is necessarily applied to all the data, but can be performed after the fact, provided the data can be found and were properly documented.

In the Long-lived Data Collections report (2005), the responsibilities listed for the data producers, managers, and scientists, as seen in figure 3, end up being performed by one or all of the authors in all six cases. Data producers are responsible for producing data; this was true for all authors in all six cases. Data managers are responsible for database operation and maintenance; this was performed by students or staff members that were appointed at the lab or departmental level and in the Webcam Case, this person was also an author on the publication. Data scientists include many categories of people, such as disciplinary experts, curators, expert annotators, and anyone who is crucial to the successful management of a digital data collection. The lead author and other contributing authors perform these roles during the contemporary use. Data scientists are not given any specific responsibilities in this report.

All of the responsibilities from these reports discussed above were or could be applied to members of the data producer category. When the data is produced by one person, it is easy to understand who the responsible party is. But, when a group of people are producing the data, this becomes more difficult. Are all data producers equally responsible, or are individuals responsible for one responsibility and not another? Arzberger et al (2004b) provided some ways to scale-up the local practices in a way that would benefit reusers. Of the ways listed, “the assignment and assumption of formal responsibilities,” and “accountability” are the easiest to adopt, but they require making the data management task distribution process explicit. The research presented here was able to attribute the performance of specific data management tasks to individual researchers. In all six cases the task distribution is largely informal, and based on entrenched roles and areas of expertise.

The perceived data management responsibilities aligned well with the data management tasks performed by the researchers. PIs, postdocs, graduate students, undergrads, technical staff, research staff, and extra-disciplinary collaborators were all part of the research groups captured. The roles performed by each were different, as were the data management tasks they performed. Graduate students lead the research in four cases, a postdoc lead one, and a research staff member lead another. The lead author occupied the first author position and performed the bulk of all management tasks. The lead author's PI assisted with selection and verification tasks. Extra-disciplinary collaborators consulted during the planning stage of the life cycle, in either the selection tasks or the verification tasks, depending on the discipline of the publication. If the publication fell within application science, then the extra-disciplinary collaborators participated in selection tasks during planning, as they consulted on the types of equipment were necessary and other collection parameters. If the publication fell within technology research, then the extra-disciplinary collaborator consulted during verification tasks throughout to ensure the technology researchers "got the science right."

Table 67 below was constructed using the parties named as responsible for data management and what those responsibilities entailed, organizing them into roles and responsibilities for each research group member. Although there are different job titles for the members of the research labs, when working on a research project together, the roles are reduced to lead author, PI, author, and the lab admin. These first three roles are analogous to the roles of authors contributing to a publication, as set out by Baerlocher et al (2007): primary, supervisor, and contributing authors.

Role	Responsibilities
All authors	Ensure data validity Answer questions from data user
Lead author	Document data to support interpretation throughout research Store data in a stable location (multiple personal machines, lab server, cloud) Deposit data in a domain repository (if applicable)
PI	Ensure data quality Appoint students or staff to serve as lab admin Set and uphold lab policy to encourage documentation and storage practices
Lab administrator	Maintain and back-up data storage location

Table 67: Data producer roles and data management responsibilities.

All authors are responsible for ensuring the validity of the data and answering the questions of people who want to use the data. The lead author and the PI inherit these responsibilities in addition to responsibilities that are uniquely their own. The lead author is also responsible for documentation, data storage, and deposit in a domain repository. In addition to ensuring validity, the PI is responsible for ensuring data quality, hiring or appointing lab administrators, setting and upholding lab policies. Lab administrators are responsible for maintenance and back-up of data storage.

The link between data management contribution and author order

That the various data producers know their data management roles and responsibilities in order to fulfill them is of benefit to the data users because this encourages good data management. That the data users would also benefit from knowing the data management roles and responsibilities of the various data producers, so that they know to whom they should address data quality, validity, storage, and documentation questions. The data management tasks performed by a given individual, reflects the individual's role in the research. For instance, a technical staff member's role is to assist with data collection equipment or running servers, and their data management contributions are likely to be in storage.

Just as promotion and tenure committees find it difficult to use author order to identify the relative contributions of the authors for evaluation (Rennie, Yank et al. 1997), it is difficult to use the author order to identify the data management functions performed by the individuals. In all of cases, the authors are ordered by contribution, with the lab PI as the last author. Even in cases where the project had multiple PIs, the PI from the lab rather than the project held the last position. Although the author order was uniform within the cases described here, the author order is not uniform for all of CENS. Within the CENS Data Practices research group, we vacillate between order by contribution and first/last. This view of author contribution is slightly skewed, and a larger sample, including more papers where a PI was the lead author would change the balance, as well as sampling publications where the authors are listed alphabetically.

In the cases where multiple disciplines were represented, only the authors who were from the publication discipline occupied the first and last author position. First/last author order was used in all of the cases and all of the cases were from the experimental sciences, so there is reason to believe that the authors did not need to come to an agreement on the type of author order to adopt (Maciejovsky, Budescu et al. 2009). Although it is likely that the authors did come to an agreement as to which author occupied each position and who was considered an author.

The lead author in each case performed the bulk of the data management tasks, and the PI performed a somewhat consistent subset of the data management tasks. Where the author order does not clarify data management contribution is with all of the other authors. Unlike the assumption reported in Burman (1982), where researchers assume that only the first, second, and last author positions have any significant contribution, in the cases with more than three authors the researchers in the other positions ranged from having very little to significant data management contributions. In the Glider Case, the third author provided some access to the ocean model, the fourth author contributed significantly to the quality and validity of the data, and the fifth author helped acquire funding and shaped the current research trajectory from the application science side. In the Stream Case, the third author provided all of the technical support, assisted with design of the experimental setup, was part of the data collection effort, and has provided feedback on the data visualizations and documentation in the publication. In the Hypoxia Case, the third author assisted with data collection, processing, and analysis, the fourth author helped design the equipment deployment at the site and then deployed the equipment, and provided technical support, and the fifth author helped acquire funding and shaped the current research trajectory from the technology side. In the Webcam Case, the third author provided technical support, ingested and annotated the data, and has provided long-term storage support, and the fourth author shaped the current research trajectory. In the Power Case, the third author tested the software stack

In a number of these cases the authors would need to explore alternative author orders or disclosure of author contribution in order to indicate data management contributions. The Hypoxia, Power, and Stream Cases all appear to have received a significant contribution from all authors, in both the research and the data management. In these cases the authors may have been better served by adopting an alphabetical order to the author names to indicate the equality of the author contributions. This would also indicate that the data user will likely need to consult more than the first two authors and the last author with any questions about research data.

The Glider Case is the only one where the first, second, and last authors had the largest contribution to the research, and the other three authors provided contributions that would just push them into the author criteria. In this case, the majority of the data management was performed by the first author. In the Webcam Case the third author contributed significantly to the data management, but not necessarily to the research. The Fungi Case was the only case with just three authors, so it would be assumed that the first, second, and last authors all contributed significantly to the research, and this is the case. They also all contributed significantly to the data management.

A formal disclosure of author contribution, like that required by the ICMJE, might be necessary to support understanding of who handled the data. Authors mentioned that they would be responsible for fielding any questions about the data that might be asked. The PI from the Stream Case indicated that he would be responsible for answering questions about the data, and one of the contributing authors from the Webcam Case indicated that all of the authors would be responsible for answering data questions. But this raises the question of whether someone who was interested in the details of the data handling be benefited by asking the PI or any of the publication authors. The PI is unlikely to have handled the data during research, and on a publication with five authors, like that described in the Webcam Case, there are so many people to choose from.

The disclosure of author contribution statement from the Webcam Case publication match the contributions reported by the three participants interviewed. The statement lacked the nuance of who handled what and when, but would otherwise give the future data user an idea of where to address their questions. Any questions regarding the scientific aspects of the data should go to the first two authors, and any technical questions about data storage and access should go to the third author. Disclosure of author contribution statements, in addition to the other benefits afforded for promotion and tenure committee review and ensuring the responsibility of authors, could support data reuse.

CHAPTER 8: CONCLUSION

The research presented in this dissertation set out to construct a bottom-up answer the question of what roles and data management responsibilities scientific data producers see as applying to themselves. From the perspective of data curators there is a lack of visibility of the science leading to data production and the data management contributions of the data producers. Studying six cases selected from a science and technology research center, I provided descriptions of the research, research data, and data practices of researchers that will eventually contribute to the management of the data. Cases were chosen from long-tail research where with proper data management the data are more likely to be reused and provide a greater impact overall. The six cases represented typical science, technology, and mixed science and technology research performed at the Center for Embedded Networked Sensing, but are atypical of their respective domains, because of the increased data collection capability afforded by the sensors and the collaboration between scientists and technology researchers. According to these researchers, the data volume and collaboration within their respective domains will increase, and it is likely that similar data practices will occur within the communities.

Data Management Core Functions

Research-grade data management is comprised of ad hoc amalgamations of policy, technology, and practices that make data usable for the existing research and for future access. They have developed organically, pulling from collaborator practices, mitigating data disasters, NSF grant review panel experiences, institutional IT support, and a variety of other sources. On the whole the participants interviewed as part of this research had good practices. They assess themselves as being on par or better than their peers in the long-term storage and accessibility of their data.

Their data management practices support current data sharing practices through personal interaction. But it remains to be seen as to whether the practices observed here would support larger-scale, 4th paradigm data sharing.

Setting best practices for presumes that there is some one-size-fits-all data management solution. Even within this small sample the groups performed the tasks in a variety of ways. Focusing on the ends, rather than the means allows the data producers the flexibility to make the processes their own. Setting best practices would at the outset alienate those researchers who already have different practices, hem in the diversity of data management solutions developed within research groups to meet their own needs, and in the future would become out-dated. We want to encourage data management, not quash it through required practices.

Rather than rely on data management best practices, a set of core functions can provide goals that provide flexibility in how they are fulfilled. As seen in these cases the data management tasks performed by the researchers to make their data usable for their own research performed certain core functions, akin to the archival core functions. The core data management functions are: selection for quality, verification for validity, storage for accessibility, and documentation for interpretability. Holding researchers responsible for whatever data management tasks will assure the quality, validity, accessibility, and interpretability of the research data they produce, is more specific than the NSF data management plan requirement language in the current form yet does not set any best practices for how these functions should be fulfilled.

These tasks could also be turned into a point for point questionnaire in lieu of the NSF data management plan requirement. By asking the data producers to answer questions about how they intend to verify outliers or whether they will use protocol sheets to capture documentation, researchers would be forced to reflect on their own processes and whether those processes actually supported the management of their own data.

Invisibility of Data Management

Data management is invisible in four ways. First, the risks of not managing data beyond the contemporary use are invisible to the data producers. Data management is an abstract goal for the data producers, and very few actual success stories are available to motivate large-scale data management efforts. Second, the data management practices of data producers have been invisible from one another until the recent NSF data management plan requirement came into effect. Third, the data management work performed by the data producers that was observed and described by this research appears to be invisible to data curators, who do not acknowledge the data producers' contribution to data management. And fourth, the role of the data curator is invisible to the data producers, who are unaware of this group of individuals who can support their data management needs.

This research speaks to the second and fourth visibility problems. By making the invisible visible, this research is challenging both the data curators and data producers to engage in more conversation. In prior research of the data life cycle, we recommended that data managers and digital curators be involved earlier in the process to assist with data management. From this work we see that there are already researchers are already managing data, to some extent. The data curators do not necessarily need to be participating in all the tasks, but they need to be aware of the data management practices of the data producers to capture all of the products. By presenting an integrated data life cycle model with curation functions with the relationship between these two groups is reinforced rather than buried.

The third visibility problem is clearing up because of newly enacted policy. Data producers need to expose their data management practices to one another. There has been a lack of data management self-esteem within the CENS community that is now starting to shift as researchers are spending more time discussing data management plans with one another. The increase in visibility is shifting practices, and spurring researcher to share their data management tasks with one another. Within this community, the practices already appear to be cutting edge, but this may be attributed to inter-disciplinary collaboration and the emphasis on sharing data that already pervades the center.

Data Management and Author Contribution

Rather than just a "data producer" role, there are a variety of roles that are performed within a research group. Roles and responsibilities were defined by who performed what tasks as well as what researchers perceived to be their responsibilities. The roles with respect to the data were remarkably similar to the roles with respect to the publication: lead, PI, and contributing authors. There is an additional role that does not fit within the author byline, the lab administrator, someone appointed or hired at the lab or departmental level to maintain servers and data structures.

That data management contribution is similar to author contribution can be leveraged, as data users can use author order to inform who if consulted when questions about data arise, assuming that the author order can be interpreted. There are a couple issues with going directly from author position to data management role. The role of the contributing author can vary widely and the lab administrator tends to not be included on the author byline. I would recommend the inclusion of author contribution statements as they are more specific and could be leveraged for data users in addition to making author contribution clear for the purposes of tenure and review.

Future Research Directions

In order to leverage the existing data management practices, I propose at least three additional studies. First, determine how prevalent these practices are across a much larger sample, the typology of data management tasks developed here could inform a survey to capture a much larger pool of participants in a structured manner. With a larger sample, variation between and within disciplines would be more pronounced. Second, determine which data management practices are successful by following the data for a longer time period, until they are reused, interviewing the data reusers as well as the data producers. And finally, determine which data management practices can be supported through policy and infrastructure, by capturing across groups where infrastructure and/or have and have not been adopted to support data management and comparing their relative success over time.

The research reported in this dissertation has not yet exhausted the data collected. The results from this dissertation will be enhanced over time with the addition of impact factors for the journals in which the case publications were published, when they have all been published. Impact factors provide a measure of quality that is absent in this dissertation. A more nuanced analysis of the variation between the science, technology, and mixed science and technology cases will also inform the potential variability to be observed in future research. And finally, the infrastructure adopted by the research groups merits study at more depth, to determine whether infrastructure can be leveraged for use by others.

The fact that there were no issues with distribution merits further research. The roles of individual researchers may be completely entrenched, or power relationships are so imbalanced that there is no option for someone to perform tasks outside their area of expertise. Observing the indoctrination of new research group members into the lab practices would potentially reveal the processes that become buried over time. Finding newly conglomerated research groups with ennui would also show how these roles are negotiated. Unfortunately it is difficult to identify the formation of an entirely new research group, because for many groups at least some subset has worked together before (i.e. people from lab X working together with people from lab Y).

Limitations of the Research

This study was by no means a comprehensive accounting of data management practices on the part of scientific and technology researchers. I have done my best to remove bias from the research and wherever this was impossible, the bias has been clearly identified so that it may be taken into account for future interpretation.

The study was limited to researchers from one multi-disciplinary research center that may not provide an accurate assessment of data management practices across the disciplines sampled. The researchers admit that their own research is atypical of their respective disciplines because of their involvement with the research center. I am gambling on the researchers' assessment that their practices will look more typical over time as the rest of their respective disciplines catch up with the researchers from CENS.

Due to a trade-off between depth and breadth, only six cases were followed, so where this dissertation has strength in depth, it is lacking in breadth. The depth has provided a great richness and texture of the practices of individual researchers and the interplay between research group members. At the same time, the small numbers of both individuals and research groups followed reduces the power of any conclusions from the work. This can be mitigated through the future research proposed in the previous section, to verify the generalizability of the results and the recommendations drawn from them.

The research method relied on the publication as a frame for interviews, document analysis, and field observation. The imposition of this frame both increased the specificity of the data provided from these methods and limited the scope of the view on data practices. Participants provided documents and answers that were a better match for the practices observed in the field than less specific methods utilized in prior data practices research. At the same time, the entire method was oriented to the publication and may have artificially increased the bond between the publication and the research.

APPENDIX I: FOLLOW-THE-DATA INTERVIEW PROTOCOL

1. **Introduction – Type of Research**

    1.3. What type of research do you do?

    1.4. What is the main research project are you working on now?

        1.4.1. When did you begin this project?

        1.4.2. How many people work on this project? [elicit team vs. collaborators]

    1.5. Are there other research projects you are currently a part of?

2. **Research Paper Description**

    (explain why this paper was chosen, and what we hope to use it for)

    2.1. Is this article coming from the main research project you working on now?

        2.1.1. If not, which project is it drawing from? [from those discussed earlier]

        2.1.2. Is this an ongoing project?

    2.2. Where did the particular topic for this paper come from? (concept or idea)

        2.2.1. Under what domain would you classify this research?

        2.2.2. Do you think this made an incremental or radical contribution to the field?

        2.2.3. How does this paper compare to others you have written?

        2.2.4. How does this paper compare to others from the field? (typical, tech heavy, application science heavy, etc)

    2.3. Can you characterize your contribution to this paper? [involved in planning, interpretation, data collection or analysis, etc]

        2.3.1. Who managed the data used in this paper while it was being prepared?

        2.3.2. What were some of the different contributions of the collaborators for this paper with regards to collecting, managing, analyzing the data, and producing tables/ graphics?

        2.3.3. How did everyone know what they needed to do for this paper?

            2.3.3.1. Was there a formal process or was it understood?

            2.3.3.2. Who holds whom accountable?

3. **Data Curation**

    3.1. We have identified this list of variables as used in the research reported in this paper. Can you confirm that this list is correct?

    3.2. Could you walk through some of the steps from designing the data collection plan to analysis? What happens as data are created, cleaned, analyzed, managed, etc.?

        3.2.1. How did you select and verify your data for this paper?

        3.2.2. What tools do you use to interpret and manage your data (Excel, R, MatLab, formulae, etc)?

3.2.3. Did your group keep records or a data log of how you acquired and processed the data? (workflow tools, etc)

3.3. For this paper did you use data from external data source (ie anyone not included in the author list)? (Data)

3.3.1. Can you walk me through the process of acquiring these data? (prompt for tools and resources used, people involved in the data sharing process, etc.)

3.3.2. Were they any particular problems or barriers you encountered in finding, getting, or using the data? (prompt for questions of search/discovery, access, interpretation, trust, metadata, usability, etc.) How did you deal with these?

3.4. Are you having any trouble managing your data collection? In the short term or long term? Why?

3.4.1. Are there data management services you wish others would provide?

3.4.2. Have you written a data management plan yet? If so, can you describe the experience?

3.4.3. Do you feel you have fulfilled your data management responsibility?

3.4.3.1. How would you compare your data management practices to those of your CENS colleagues?

3.4.3.2. How would you compare your data management practices to those of your domain peers?

4. Data Disposition

4.1. Where does the data used for this paper reside? Where is it stored and accessed? (Could you show us?)

4.1.1. Could you locate the data for this [table/graphic/image]?

4.2. What system do you use for naming files, data collections, or versions of your publication or of the tables and graphics used in it?

4.2.1. Can you tell us or show us where you keep your versions?

4.2.2. Do you have additional source files containing data or other resources you used for the publication?

4.3. Do you keep the subsets of data used for the paper? Can you show us where?

4.3.1. You said that you drew from [source outside research group] - How do you work with the subsets of data taken from this source?

5. Data Sharing

5.1. Would the particular data used for this article [table/figure] be relevant to others to re-use?

5.1.1. Would they want to extend these data or compare them to new observations?

5.2. For others to use this dataset, what sorts of additional information might you need to add to the data that is not in the article?

5.2.1. Would you need to include additional information about the instrument, other equipment or software tools used? For example, did you write your own code that might also need to be shared to use the data?

5.2.2. Do you think the publication should be linked to the data used for it?

5.3. Can data be published as an end in itself in your field?

5.4. When would you consider yourself the author of a dataset?

5.4.1. What are the criteria that determine authorship? [Probe Intellectual Property rights and effort expenditure]

5.4.2. Who would you consider to be the owner of the data?

5.4.3. Who is responsible for the dataset? What does responsibility entail? [Try to elicit types of responsibility]

6. Closing

6.1. Who else should we talk to in order to fill in the picture of managing the data that contributed to this paper, and where those data are now?

6.1.1. Any post-docs or grad students?

6.2. Is there anything else about the use of data for publications that you think we are missing, or should be asking about in our interviews?

APPENDIX II: CENS WIND-DOWN INTERVIEW PROTOCOL

The Transformation of Knowledge, Culture, and Practice in Data-Driven Science:

A Knowledge Infrastructures Perspective

Purpose of the Study

The goal of this study is to develop new tools and best practices that provide an integrated framework or infrastructure for the management of the vast quantities of data that are generated in scientific practice. The absence of such tools is a significant barrier not only to the initial discovery and selection of data, but also to the subsequent reuse of the same data by multiple communities of scientists and nonscientists. In this study, we will gather and analyze data on those scientific practices and data management requirements to inform the design process. This research will continue to address the requirements for data-intensive science, focusing on how these needs are articulated and develop during the last year of an NSF-funded Science and Technology Center. Our aim is to enable scientists within CENS to continue to focus on problems that arise within those domains, and to minimize their concern with "the information problem." Our long-term goal is to establish a CI Virtual Observatory for the study of data, data analysis, and visualization.

Demographic Information

1. What research project at CENS are you working on right now, or what was your last research project at CENS?
   - 1.1. What are your research questions for this project?
   - 1.2. What resources/tools/technologies/human assistance do you need to use to perform your research?
   - 1.3. How many people are on this project? Who is from CENS? Who is not?
   - 1.4. Who do you need to talk to on a daily basis? Weekly basis?
   - 1.5. What is the expected duration of this project? Do you expect to continue working on this after you leave CENS?

Project Boundaries

2. Are you working on a project that is outside of CENS?
   2.1. How do you decide what work falls under CENS?

Data Characteristics

3. Within your work, what do you consider to be data? (informing data)
   3.1. What variables are you dealing with?
   3.2. When you look at the data, what are you hoping to find?
   3.3. How do you use different types of data?

Data Management

4. How do you manage your data collection so that you can use it again in the short and long-term future? (Processes)
   4.1. How are data currently shared within your team? (Processes)
   4.2. What are your criteria for selecting and preserving data? For sharing? (Processes)
   4.3. What resources did CENS make available to you for data management? Was this helpful? Was there something they could have done but did not?

Data Sharing

5. How are data shared with people outside your team?
   5.1. Do you make any of your data public available online? Do you use public repositories? (Processes)
   5.2. Who else might be interested in your data?
   5.3. Have you ever been asked by someone outside your project to share data before? If yes, can you elaborate on the process?

Data Legacy

6. What will happen to the data (yours and others)?
   6.1. What would you like to happen with your data?

Establishment of Collaboration

7. How and when did you come to join CENS?
   7.1. Why did you decide to join CENS?
   7.2. What data, tools, students, or staff did you bring with you to CENS?

7.3. Was there anything that you were hoping to get out of CENS?

Information Infrastructure

8. How does CENS compare with other projects that you have worked on?
   8.1. Did CENS provide you with the appropriate resources to carry out your research? How so? If there were perceived weaknesses, how were these addressed?
   8.2. How has your research changed since coming to CENS?
   8.3. If you had not worked at CENS, how would your research or professional life be different?

Collaboration Wind Down

9. What steps are you taking to prepare for the end of CENS?
   9.1. Are you required to follow any end-of-project reporting requirements? Do these pose additional strain on the winding down of CENS?
   9.2. How do think what you learned about organizing or participating in a project like CENS can be passed on to others?

CENS Legacy

10. What will be CENS's most important legacy?
   10.1. What do you expect to take away from CENS - experiences, tools, data?
   10.2. What do expect will happen with the physical components of CENS? What do you think should happen?

Personal Experience

11. What does the end of CENS mean for you personally?
   11.1. What are they going to miss most when the project ends?
   11.2. Have you made any plans for where you will continue after CENS? If so, what are they?

APPENDIX III: CODEBOOK

CENS Data Practices Codebook (CENS v3)

JCW - v3.1 - 2012-02-15

Introduction

The following codebook was designed to construct networks of different classes of thing, such as person, data, research group, publication, etc. In so doing the codes capture fields to construct thick descriptions of each individual item, and collocation (in the same interview or case) allows the interconnections of items to one another. This codebook is based in two prior codebooks, CENS v1 and MMM (CENS v2), which are themselves loosely related. Crosswalks between codebooks are indicated on a code-by-code basis. The interviews to be coded with this codebook have been collected using the Follow the Data/JW Dissertation protocol and the CENS Wind-down protocol, or just the CENS Wind-down protocol.

How to Use the Codebook

Each code has a scope note that explains when the code should be applied, a list of questions from both interview protocols where coding should likely be applied, and the overlap with the CENS v1 and CENS v2 codebooks for future reference.

0 Participant

To be applied to the interview file itself as file attributes; used describe the participant through various demographic dimensions.

0.1 Participant name

Participant name (CENS v1 – 0.1 Researcher; CENS v2 – Header info)

0.2 Participant role

This attribute should describe the employment status of the participant, using one of the following terms to describe them: PI, Faculty, Student, Postdoc, or Staff. (CENS v1 – 0.2 Research role; CENS v2 – Header info)

0.3 Participant's Project

This attribute should be used to record the project with which the participant identifies. (CENS v1 – 0.3 Project; CENS v2 – Header info)

Follow the Data/JW Dissertation

What is the main research project you are working on now?

CENS Wind-down

> What research project at CENS are you working on right now?

0.4 Participant's Domain

This attribute should be used to capture the domain with which the participant identifies. (CENS v1 – combined 0.4 Application area, 0.5 Technology area, and 0.6 Science or technology; CENS v2 – Header info)

Follow the Data/JW Dissertation

> What type of research of research do you do?

0.5 Interview date

Date of interview in YYYYMMDD format. (CENS v1 – 0.7 Interview date; CENS v2 – Header info)

0.6 Interviewer/s

Interviewers present for the interview. (CENS v1 – 0.8 Interviewer/s; CENS v2 – Header info)

1 Project

Various attributes of the project the participant identifies with from 0.3.

1.1 Project description

Use for capturing the general description of the project, this should be a much thicker description of the project than 0.3. This code should be limited to what the participant describes as their core research project, anything beyond that should go in 1.6 Project context.

Follow the Data/JW Dissertation

> What is the main research project you are working on now?

CENS Wind-down

> What research project at CENS are you working on right now?

1.2 Research questions

This code should be used to capture the participant’s RQs for the project being described. This code can also be used to capture other discussion about the theories behind the researcher's current work or what they are trying to learn in their current work, as well as the long term goals. This code should be used to denote the form of the contributions or "findings", or the product of scholarly activities. (CENS v1 – 1.1 Research questions and 1.4 Research contribution)

CENS Wind-down

> What are you research questions for this project?

1.3 Project initiation

Stories about how a project emerged and evolved - who, when, where, why. Should also be used to capture how the participant became involved if the project was already established. For some people there will be a high overlap between what falls under this code and 1.6 Project context. For CENS initiation use 6.1 Joining CENS. (CENS v1 – 1.2 Project duration; CENS v2 – Biography of the project)

Follow the Data/JW Dissertation

> When did you begin this project?

1.4 Project duration

Capture any short-term and long-term projections of how long the project or overall research trajectory will last. For example, the CENS grants will end this year, but the research trajectory will probably take the next 20 years. (CENS v1 – 1.2 Project duration; CENS v2 – Biography of the project)

CENS Wind-down

What is the expected duration of this project?

1.5 Post-CENS

Capture specifically how this project or research trajectory will continue when CENS ends. For more general effects of CENS on research and research trajectory use Effect on research.

CENS Wind-down

Do you expect to continue working on this after you leave CENS?

Have you made any plans for where you will continue after CENS? If so, what are they?

1.6 Project context

Capture how this project is situated with respect to other projects. Most of the CENS researchers are part of a larger network of projects, holding multiple grants and one or two CENS lines. (CENS v2 – Collaborators & interconnections)

Follow the Data/JW Dissertation

Are there other research projects you are currently part of?

CENS Wind-down

Who is from CENS? Who is not?

Are you working on a project that is outside of CENS?

How do you decide what work falls under CENS?

1.7 Project people

Capture how many people, and names if provided of the members of the participant's immediate team. (CENS v1 – 1.1 Group members; CENS v2 – Collaborators & interconnections)

Follow the Data/JW Dissertation

How many people work on this project?

CENS Wind-down

How many people work on this project

Who do you need to talk to on a daily basis? Weekly basis?

1.8 Project support

Capture what is necessary to perform the research.

CENS Wind-down

What resources/tools/technologies/human assistance do you need to use to perform your research?

2 Paper

The following codes should be used to describe attributes of the publication used for the Follow the Data protocol.

2.1 Paper topic

Capture the domain and topic of the paper, which may overlap entirely with the project description. This also includes where the topic comes from - prior research initiative, papers, etc.

Follow the Data/JW Dissertation

Where did the particular topic for this paper?

Under what domain would you classify this research?

2.2 Paper context

Capture how the research being reported in the paper relates to other work being done in this area.

Follow the Data/JW Dissertation

Is this article coming from the main research project you are working on? If not, which project is it drawing from?

Is this on-going research?

2.3 Paper contribution

This code should be used to capture the contributions or "findings" reported in this publication and how they are perceived to relate to comparable publications.

Follow the Data/JW Dissertation

> Do you think this paper made an incremental or radical contribution to the field?
>
> How does this paper compare to others from the field?

3 Data

The following codes should be used to capture the various attributes of the data discussed by the participant.

3.1 Data sources

This code should capture where data are coming from, including simulation, sensors, hand collection, etc. and whether the researcher has used any data captured by others. This should also be used to capture other issues and examples of the utilization of data generated by others. This includes disciplinary attitudes, 1st person narrative, and 2nd person examples, such as the example of post-WWII German researchers performing meta-analysis. There is some expected overlap with “Uses of data.” (CENS v1 – 5.5 Using the data of others; CENS v2 – Data production)

Follow the Data/JW Dissertation

> For this paper did you use data that were produced by anyone not included in the author list?

Can you walk me through the process of acquiring these data?

> Were there any particular problems or barriers you encountered in finding, getting, or using the data?

3.2 Data type, variety, & scope

This code should be used for capturing the various types of information thought of as data, for example images, sensor readings, biological samples, etc. This code should also be used to capture the various data variables being collected, and for what time period. For instance, the NAMOS group will deploy the buoys for 3 days and collect temperature at various depths, ambient light, fluorescence of algae, etc. (CENS v1 – combined 3.2 Data types and 3.3 Data variety, scope; CENS v2 – combined Data characteristics and Data production)

Follow the Data/JW Dissertation

> We have identified this list of variables as used in the research reported in this paper. Can you confirm that this list is correct?

> What would you say are the most essential data for your research?

CENS Wind-down

> Within your work, what do you consider to be data?

> What variables are you dealing with?

3.3 Data use & reuse

This code should be used to capture the possible uses of data by the investigator for the purposes of their own research. This code is also meant to capture the response to probes about the evolution of data use as the data ages. For example the data might become reference data for future deployments, or it might become longitudinal data to describe trends over time. (CENS v1 – combined 3.1 Data use and 3.1.2 Data reuse; CENS v2 – combined Data-code use and Data-code reuse & sharing)

CENS Wind-down

> When you look at the data, what are you hoping to find?

> How do you use different types of data?

3.4 Data authorship

This code is meant to capture whether the participant believes data can be authored, whether they consider themselves and author, and the criteria that determine authorship of data. (CENS v1 – 5.2 Data authorship; CENS v2 – Data authorship (JW personal))

Follow the Data/JW Dissertation

> Would you consider yourself the author of a dataset?

> What are the criteria that determine authorship?

3.5 Data ownership

Same but for ownership (CENS v1 – 5.2 Data authorship; CENS v2 – Data ownership (JW personal))

Follow the Data/JW Dissertation

> Would you consider yourself the owner of a dataset?

4 Data curation

These codes capture broad and specific data curation activities, the tools necessary to perform data curation, and experiences with data management plans.

4.1 Data management

This code is used to capture the general data management and curation practices from planning data collection to preservation, and order-of-operations. For finer-grained detail, use the sub-codes. We expect that the CENS wind-down will use this code for general dm practices, and the dissertation work will use it for order-of-operations (CENS v1 – 2.1 Data management practices; CENS v2 – Work practices)

Follow the Data/JW Dissertation

> Can you walk us through some of the steps from designing the data collection plan to analysis? What happens as the data are created, cleaned, analyzed, managed, etc?

CENS Wind-down

> How do you manage your data collection so that you can use it again in the short and long-term future?

4.1.1 Data selection & appraisal

This code should be used to capture any discussion of data selection or preservation techniques, reasons, goals, etc. (CENS v1 – 2.1.1 Archival issues)

CENS Wind-down

> What are your criteria for selecting and preserving data?

4.1.2 Data QAQC

Use this code to capture any standards and quality control discussion. (CENS v1 – 3.5 Data standards, quality control; CENS v2 – Data systems)

4.1.3 Data metadata & documentation

This code should be used to capture any implicit and explicit discussion of metadata and metadata capture. This includes specific standards, and technologies used for the production and management of metadata. (CENS v1 – 3.5 Data standards, quality control; CENS v2 – Data - metadata and ontology)

Follow the Data/JW Dissertation

> Did your group keep records or a data log of how you acquired and processed the data?
>
> What systems do you use for naming files, data collections, or versions of your publications or of the tables and graphics used in it? Can you tell us where you keep your versions?
>
> For others to use this dataset, what sorts of information might you need to add to the data that is not in the article?

4.1.4 Data storage

How data are stored and shared within the research group. (CENS v1 – 5.1.1 Intra-group sharing; CENS v2 – Data systems)

Follow the Data/JW Dissertation

> Where does the data for this paper reside? Where is it stored and accessed? Could you locate the data for this?
>
> Do you keep subsets of data used for the paper? Can you show us where?

CENS Wind-down

> How are data currently shared within your team?
>
> What will happen to the data (yours and others)? What would you like to happen to your data?

4.2 Data tools & services

This code should be used to capture the various tools and services used in data handling (may end up redundant). (CENS v1 – 2.1.2 Data tools; CENS v2 – Work tools & equipment)

Follow the Data/JW Dissertation

> What tools do you use to interpret and manage your data?
>
> Are there any data management services you wish other would provide?

4.3 Data management plan

This code is meant to capture any experiences with data management plans.

Follow the Data/JW Dissertation

Have you written a data management plan yet? If so, can you describe the experience?

4.4 Data sharing

This code should be used to capture the response to the question about whether data has been shared with others. This also includes general attitudes towards sharing, if it has occurred and the experience. (CENS v1 – 4.2 Repositories, 4.4 Publication of data, 5.1 Sharing data with others; CENS v2 – Data-code reuse & sharing, Knowledge products)

Follow the Data/JW Dissertation

Can data be published as an end in itself in your field?

In your field do they ever link the publication to the data?

Would the particular data used for this article be relevant to others to re-use?

Would they want to extend these data or compare them to new observations?

CENS Wind-down

Do you make any of your data public available online?

Do you use repositories?

How are data shared with people outside of your team?

Who else might be interested in your data?

Have you ever been asked by someone outside your project to share data before? If yes, can you elaborate on the process?

5 Responsibility

The following codes capture aspects of responsibility: what responsibility for data entails, how it is distributed among a group, whether it is fulfilled, and who holds whom accountable.

5.1 Division of labor

Use this code to capture who did what during the paper research, and how it came to be understood who would do what. (CENS v1 – 6.5 Division of labor; CENS v2 – Roles)

Follow the Data/JW Dissertation

> Can you characterize your contribution to this paper?
>
> Who managed the data used in this paper while it was being prepared?
>
> What were some of the different contributions of the co-authors?
>
> How did everyone know what they needed to do for this paper? Was there a formal process or was it understood?

5.2 Accountability

This code captures how accountability works within collaborators. (CENS v1 – 6.4 Consultation; CENS v2 – Coordination work & tools)

Follow the Data/JW Dissertation

> Who holds whom accountable?

5.3 Fulfillment of responsibility

Capturing perceptions of whether or not the participant has fulfilled their responsibility, and comparing their own data practices to those of others.

Follow the Data/JW Dissertation

> Do you feel you have fulfilled your data management responsibilities?

5.4 Data Responsibility

Capturing understanding of who is responsible for data and what responsibility entails. (CENS v1 – 5.2 Data authorship; CENS v2 – Data responsibility (JW personal))

Follow the Data/JW Dissertation

> Who is responsible for the dataset?
>
> What does responsibility include?

6 Center

Capturing attributes of the CENS research center and the effects of the wind-down on research.

6.1 Joining CENS

This code should be used to capture how the participant became involved in CENS or how CENS was initiated if they were part of the initiation. (CENS v1 – 1.2 Project duration; CENS v2 – Biography of the project)

CENS Wind-down

> How and when did you come to join CENS?
>
> Why did you decide to join CENS?
>
> What data, tools, students, or staff did you bring with you to CENS?

6.2 Center support

This code is intended to capture the support the center provided above and beyond what is usually found in academic research, or the lack thereof. (CENS v2 – Coordination work & tools)

CENS Wind-down

> Did CENS provide you with the appropriate resources to carry out your research? If there were perceived weaknesses, how were these addressed?

6.3 Center evolution

Use this code to capture how CENS changed as a center during the participant's tenure.

CENS Wind-down

> Did CENS change as a center during the time you were a part of it? Can you characterize that change?

6.4 Effect on research

The code should be used to capture any mention of the role and effects CENS has had on the participant's research and/or research trajectory. More general effects should be captured under Center Legacy. (CENS v2 – Biography of the project and Governance)

CENS Wind-down

> How does CENS compare with other projects you have worked on?
>
> How has your research changed since coming to CENS?
>
> What does the end of CENS mean for you personally? What are you going to miss most when the project ends?

What do you expect to take away from CENS - experiences, tools, data?

6.5 Center legacy

More general discussion of the effects an STC center can have on research

CENS Wind-down

How do you think what you learned about organizing and participating in a project like CENS can be passed on to the others?

What will be CENS's most important legacy?

6.6 Center dissolution

Use this code to capture center wind-down activities and effects.

CENS Wind-down

What steps are you taking to prepare for the end of CENS?

Are you required to follow any end-of-project reporting requirements? Do these pose additional strain on the winding down of CENS?

What do you expect will happen with the physical components of CENS? What do you think should happen?

7 Cross-grain codes

The following codes can be applied wherever needed.

7.1 Problems

Challenges, controversies, difficulties – a catch-all category. (CENS v2 – Problems)

7.2 Trust

This code should be used to capture any issues of trust, for example not trusting the data coming off a specific type of sensor, etc. (CENS v1 – 4.3 Trust in data/tech/people; CENS v2 – Values & trust)

7.3 Affective issues

This code should be used to capture when participants mention how something makes them feel or more subtle articulation of feeling, negative reaction to terms, really long pauses, etc.

7.4 Successes

This code should be used to capture things that the participant feels very positive about. (CENS v2 – Successes)

APPENDIX IV: SOLICITATION LETTER

Dear ________,

We are writing to request an interview with you in January as part of our NSF-funded research project called Monitoring, Modeling, and Memory (MMM). MMM is a distributed collaboration of 10 investigators from 5 universities studying large-scale cyberinfrastructure projects including CENS, WATERS, and the LTERs.

Our UCLA team is currently studying two aspects of CENS: the data curation tasks performed by CENS researchers, and the impact CENS as a center has had on your research. We want to identify successful collaborative data curation strategies, which will in turn be recommended as best practices to NSF. We also want to capture a record of the harder to measure impacts a large collaboration has on research, especially now as the center is winding down. We have been interviewing researchers at CENS since 2005, and have used what we have learned to build systems and draft cyberinfrastructure policy.

For our current round of interviews, we are examining how CENS researchers curate the particular data sets they include in publications, how the data are generated, selected, and preserved after publication. In order to ground the interview, we would like to discuss the data you used in one of your recent publications. We identified the following article as one that may be relevant for us to discuss because of your use of CENS technologies to collect data:

(Citation)

We assume that this is an article you may have worked on with other colleagues, but we hope you can discuss your role in this research and any contact you have had with the data used for this publication. We are interested in learning about how you typically work with data for publications and collaborate with other authors. We may also follow up with other authors of this paper who had roles in working with the data. If you think another recent article or pre-print would be more relevant for us to discuss with you, please let us know.

The interview would last approximately one to two hours, and could be held at the location of your choice (your office, CENS meeting spaces, etc). With your permission, we would like to record the interview. For some of the questions, you may wish to show us files or databases on your computer that you used for the publication. Before conducting the interview we will ask you to sign an informed consent document similar to those used in other research projects using interviews. We will honor any constraints you require in our research use of your interview.

Please let us know whether you will participate in this round, and we will follow up to schedule an interview. We are eager to answer any questions you might have about our project. We will be grateful for your participation.

Sincerely,

Jillian C. Wallis

Principal Investigator: Christine L. Borgman, Professor & Presidential Chair, Information Studies, UCLA http://is.gseis.ucla.edu/cborgman/

Graduate Student Researchers: Jillian C. Wallis and Lizzy Rolando, Information Studies, UCLA

REFERENCES

Ambrosone, C. B. & Kadlubar, F. F. (1997). Toward and Integrated Approach to Molecular Epidemiology. American Journal of Epidemiology, 146: 912-918.

Anderson, C. (2004). The long tail. Wired Magazine, 12(10). Retrieved from http://wired.com/wired/archive/12.10/tail_pr.html on 17 September 2006.

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, 303(5665): 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, 3: 135-152.

Babbie, E. (2007). The Practice of Social Research (11th ed.). Belmont, CA: Wadsworth.

Baerlocher, M. O., Newton, M., Gautam, T., Tomlinson, G. & Detsky, A. S. (2007). The Meaning of Author Order in Medical Research. Journal of Investigative Medicine, 55(4): 174-180.

Baker, K. (2011). *Personal communication*. Personal Communication to Wallis, J. C.

Baker, K., Millerand, F. & Yarmey, L. (2009). Growing Information Infrastructure: Data Lifecycles and Subcycles. Estes Park, CO. Retrieved from http://oceaninformatics.ucsd.edu/media-gallery/mediaDB/lter/29.pdf on 6 March 2011.

Bennett, D. M. & Taylor, D. M. (2003). Unethical practices in authorship of scientific papers. Emergency Medicine, 15(3): 263-270.

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, 28: 235-242.

Bhopal, R. S., Rankin, J. M., McColl, E., Thomas, L., Kaner, E., Stacy, R., Pearson, P., Vernon, B. & Rodgers, H. (1997). The Vexed Question of Authorship: Views of researchers in a British medical faculty. British Medical Journal, 314: 1009-1012.

Birnholtz, J. P. (2006). What Does it Mean to be an Author? The intersection of credit, contribution, and collaboration in science. Journal of the American Society for Information Science and Technology, 57(13): 1758-1770.

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

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

Borgman, C. L., Wallis, J. C. & Enyedy, N. (2006). Building digital libraries for scientific data: An exploratory study of data practices in habitat ecology. 10th European Conference on Digital Libraries, Alicante, Spain, Berlin: Springer. 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, 7(1-2): 17-3029 September 2007.

Borgman, C. L., Wallis, J. C. & Mayernik, M. S. (forthcoming). Who's got the data? Interdependencies in Science and Technology Collaborations. Journal of Computer Supported Collaborative Work.

Borgman, C. L., Wallis, J. C., Mayernik, M. S. & Pepe, A. (2007). Drowning in data: Digital library architecture to support scientific use of embedded sensor networks. Vancouver, British Columbia, Canada, Association for Computing Machinery: 269-277.

Bourne, P. (2005). Will a biological database be different from a biological journal? PLoS Computational Biology, 1(3): e34. Retrieved from http://dx.doi.org/10.1371/journal.pcbi.0010034 on 28 September 2006.

Bowker, G. C. (2000a). Biodiversity datadiversity. Social Studies of Science, 30(5): 643-683.

Bowker, G. C. (2000b). Mapping biodiversity. International Journal of Geographical Information Science, 14(8): 739-754.

Bowker, G. C. (2000c). Work and information practices in the sciences of biodiversity. VLDB 2000, Proceedings of 26th international conference on very large data bases, Cairo, Egypt, Kaufmann. 693-696.

Bowker, G. C. (2005). Memory Practices in the Sciences. Cambridge, MA: MIT Press.

Brown, C. L., Chan, K. C. & Chen, C. R. (2011). First-author Conditions: Evidence from finance journal coathurship. Applied Economics, 43(25): 3687-3697.

Buehring, G. C., Buehring, J. E. & Gerard, P. D. (2007). Lost in Citation: Vanishing visibility of senior authors. Scientometrics, 72(3): 459-468.

Burman, K. D. (1982). "Hanging from the Masthead": Reflections on authorship. Annals of Internal Medicine, 97(4): 602-605.

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

CIRSS Data Curation Education Program. Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. Retrieved from http://cirss.lis.illinois.edu/CollMeta/dcep.html on 4 March 2011.

CODATA-CENDI Forum on the National Science Board Report on Long-Lived Digital Data Collections. (2005). U.S. National Committee on CODATA, National Research Council. Retrieved from http://www7.nationalacademies.org/usnc-codata/Forum_on_NSB_Report.pdf on 29 September 2006.

Cole, F. T. H. (2008). Taking "Data"(as a Topic): The Working Policies of Indifference, Purification and …. Christchurch, NZ: 240-249.

Collins, H. M. (1998). The Meaning of Data: Open and Closed Evidential Cultures in the Search for Gravitational Waves. American Journal of Sociology, 104(2): 293-338.

Cook, R. B., Olson, R. J., Kanciruk, P. & Hook, L. A. (2001). Best Practices for Preparing Ecological Data Sets to Share and Archive. Bulletin of the Ecological Society of America, 82(2): 138-141.

Costas, R. & Bordons, M. (2011). Do Age and Professional Rank Influence the Order of Authorship in Scientific Publications? Some evidence from a micro-level perspective. Scientometrics, 88: 145-161.

Cragin, M. H., Palmer, C. L., Carlson, J. R. & Witt, M. (2010). Data sharing, small science and institutional repositories. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 368(1926): 4023-4038.

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

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices. Journal of the American Society for Information Science and Technology, 52(7): 558-569.

Cummings, J. N. & Kiesler, S. (2005). Collaborative research across disciplinary and organizational boundaries. Social Studies of Science, 35(5): 703-722.

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

David, P. A. & Spence, M. (2003). Towards Institutional Infrastructures for E-Science: The Scope of the Challenge. Oxford Internet Institute Research Reports: University of Oxford. 92 Retrieved from http://129.3.20.41/eps/le/papers/0502/0502002.pdf on 30 September 2006.

Davies, H. D., Langley, J. M. & Speert, D. P. (1996). Rating authors' contributions to collaborative research: the PICNIC survey of university departments of pediatrics. Pediatric Investigators' Collaborative Network on Infections in Canada. Canadian Medical Association Journal, 155(7): 877-882.

Easterbrook, S. M. & Johns, T. C. (2009). Engineering the software for understanding climate change. Computing in Science & Engineering: 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, 41(5): 667-690.

Engers, M., Gans, J. S., Grant, S. & King, S. P. (1999). First-Author Conditions. Journal of Political Economy, 107(4): 859-883.

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.

Floyd, S. W., Schroeder, D. M. & Finn, D. M. (1994). "Only if I'm First Author": Conflict over credit in management scholarship. The Academy of Management Journal, 37(3): 734-747.

Galison, P. (1997). Image and Logic: A Material Culture of Microphysics. Chicago: University of Chicago Press.

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

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, 24(2): 192-204.

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. Library Trends, 57(2): 280-299.

Hey, A. J. G. & Trefethen, A. (2003). The Data Deluge: An e-Science Perspective. In Berman, F., Fox, G. & Hey, A. J. G. (Eds.). Grid Computing: Making the Global Infrastructure a Reality. Chichester, Wiley. Retrieved from http://www.rcuk.ac.uk/escience/documents/report_datadeluge.pdf on 20 January 2005.

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

Higgins, S. (2008). The DCC Curation Lifecycle Model. International Journal of Digital Curation, 3(1): 134-140.

Hilgartner, S. & Brandt-Rauf, S. I. (1994). Data access, ownership and control: Toward empirical studies of access practices. Knowledge, 15: 355-372.

Hilmer, C. E. & Hilmer, M. J. (2005). How Do Journal Quality, Co-Authorship, and Author Order Affect Agricultural Economists' Salaries? American Journal of Agricultural Economics, 87(2): 509-523.

Hoen, W. P., Walvoort, H. C. & Overbeke, J. P. M. (1998). What are the Factors Determining Authorship and the Order of the Authors' Names? Journal of the American Medical Association, 280(3): 217-218.

Joseph, K., Laband, D. N. & Patil, V. (2005). Author Order and Research Qaulity. Southern Economic Journal, 71(3): 545-555.

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, 2(3-4): 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, 15(4): 321-358.

Laband, D. N. (2002). Contribution, attribution and the allocation of intellectual property rights: economics versus agricultural economics. Labour Economics, 9(1): 125-131.

Laband, D. N. & Tollison, R. D. (2000). Intellectual Collaboration. Journal of Political Economy, 108(3): 632-662.

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

Latour, B. & Woolgar, S. (1979). Laboratory life: The Social Construction of Scientific Facts. Beverly Hills: Sage Publications.

Latour, B. & Woolgar, S. (1986). Laboratory Life: The Construction of Scientific Facts (2nd ed.). Princeton, N.J.: Princeton University Press.

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

Lee, C. A., Tibbo, H. R. & Schaefer, J. C. (2007). Defining What Digital Curators Do and What they Need to Know: The DigCCurr Project. Joint Conference on Digital Libraries. Vancouver, BC, CA. 49-50.

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. 483-492.

Levitan, K. B. (1982). Information as "goods" in the life cycle of information production. Journal of the American Society for Information Science, 33(1): 44-54.

Lord, P. & Macdonald, A. (2003). E-Science Curation Report--Data Curation for E-science in the UK: An Audit to Establish Requirements for Future Curation and Provision. JISC Committee for the Support of Research. 85 pages. Retrieved from http://www.jisc.ac.uk/uploaded_documents/e-scienceReportFinal.pdf on 1 October 2006.

Lynch, C. A. (2008). Big data: How do your data grow? Nature, 455(7209): 28-29.

Lyon, L. (2007). Dealing with data: Roles, rights, responsibilities, and relationships. UKOLN. Retrieved from http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories/project_dealing_with_data.aspx on 23 July 2007.

Martín, E. & Ballard, G. (2010). Data Management Best Practices and Standards for Biodiversity Data Applicable to Bird Monitoring Data, U.S. North American Bird Conservation Initiative Monitoring Subcommittee. Retrieved from http://www.nabci-us.org/aboutnabci/bestdatamanagementpractices.pdf on 22 May 2012.

Mayernik, M. S., Batcheller, A. L. & Borgman, C. L. (2011). How Institutional Factors Influence the Creation of Scientific Metadata. Proceedings of the 2011 iConference, Seattle, WA. 417-425.

Mayernik, M. S., Wallis, J. C. & Borgman, C. L. (2010, in review). Unearthing the infrastructure: Humans and sensors in environmental and ecological field research. Social Studies of Science.

Mayernik, M. S., Wallis, J. C., Pepe, A. & Borgman, C. L. (2008). Whose data do you trust? Integrity issues in the preservation of scientific data. iConference, Los Angeles, CA. Retrieved from http://www.ischools.org/oc/conference08/ on 25 January 2008.

NSF (2011a). Award and Administration Guide: Chapter VI - Other Post Award Requirements and Considerations. National Science Foundation. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4 on 23 May 2012.

NSF (2011b). Grant Proposal Guide: Chapter II - Proposal Preparation Instructions. National Science Foundation. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp on 23 May 2012.

Olson, G. M., Zimmerman, A. & Bos, N. (Eds.). (2008). Scientific Collaboration on the Internet. Cambridge, MA: MIT Press.

Pennock, M. (2007). Digital Curation: A Life-Cycle Approach to Managing and Preserving Usable Digital Information. Library & Archives,(1). Retrieved from http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf on.

Pepe, A., Borgman, C. L., Wallis, J. C. & Mayernik, M. S. (2007). Knitting a fabric of sensor data and literature. Information Processing in Sensor Networks, Cambridge, MA, Association for Computing Machinery/IEEE. http://works.bepress.com/albertopepe/5/.

Pepe, A., Mayernik, M., Borgman, C. L. & Van de Sompel, H. (2009). Technology to Represent Scientific Practice: Data, Life Cycles, and Value Chains. ArXiv.org. Retrieved from http://arxiv.org/abs/0906.2549v1 on 29 June 2009.

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

Pichini, S., Pulido, M. & García-Algar, Ó. (2005). Authorship in manuscripts submitted to biomedical journals: and author's position and its value. Science and Engineering Ethics, 11: 173-175.

Price, D. J. d. S. (1963). Little Science, Big Science. New York: Columbia University Press

Pritchard, S. M., Carver, L. & Anand, S. (2004). Collaboration for knowledge management and campus informatics. University of California, Santa Barbara. 38. Retrieved from http://www.library.ucsb.edu/informatics/informatics/documents/UCSB_Campus_Informatics_Project_Report.pdf on 5 July 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. Retrieved from http://www.nap.edu on 28 September 2006.

Reference Model for an Open Archival Information System (2002). Recommendation for Space Data System Standards: Consultative Committee for Space Data Systems Secretariat, Program Integration Division (Code M-3), National Aeronautics and Space Administration. Retrieved from http://public.ccsds.org/publications/archive/650x0b1.pdf on 4 October 2006.

Rennie, D., Flanagin, A. & Yank, V. (2000). The Contributions of Authors. Journal of the American Medical Association, 284: 89-91.

Rennie, D., Yank, V. & Emanuel, L. (1997). When Authorship Fails: A proposal to make contributors accountable. Journal of the American Medical Association, 278(7): 579-585.

Ribes, D. & Finholt, T. A. (2007). Tensions across the scales: Planning infrastructure for the long-term. Sanibel Island, Florida, Association for Computing Machinery: 229-238.

Riesenberg, D. & Lundberg, G. D. (1990). The Order of Authorship: Who's on first? Journal of the American Medical Association, 264(14): 1857.

SAA Guidelines for College and University Archives. Society of American Archivists. Retrieved from http://www.archivists.org/governance/guidelines/cu_guidelines4.asp on 12 April, 2012.

Savitz, D. A. (1999). What can we infer from the author ordering in epidemiology? American Journal of Epidemiology, 149: 401-403.

Segal, J. (2005). When software engineers met research scientists: A case study. Empirical Software Engineering, 10: 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, 18(5-6), pp. 581–606.

Shapiro, D. W., Wenger, N. S. & Shapiro, M. F. (1994). The Contributions of Authors to Multiauthored Biomedical Research Papers. Journal of the American Medical Association, 271(6): 438-442.

Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility. (2003). Meeting organized by the Wellcome Trust, Fort Lauderdale, Florida, Wellcome Trust. Retrieved from www.wellcome.ac.uk/.../groups/corporatesite/@policy_communications/documents/web_document/wtd003207.pdf on 29 December 2009.

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

Solomon, J. (2009). Programmers, Professors, and Parasites: Credit and co-authorship in computer science. Science and Engineering Ethics, 15: 467-489.

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, 19(3): 387-420.

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

Thibodeau, K. (2007). The electronic records archives program at the National Archives and Records Administration. First Monday, 12(7). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1922/1804 on 17 March 2011.

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

Tscharntke, T., Hochberg, M. E., Rand, T. A., Resh, V. H. & Krauss, J. (2007). Author Sequence and Credit for Contributions in Multiauthored Publications. PLoS Biology, 5(1): 13-14.

Van de Sompel, H., Hammond, T., Neylon, E. & Weibel, S. L. (2006). RFC 4452: The "info" URI Scheme for Information Assets with Identifiers in Public Namespaces. Requests for Comments, Internet Engineering Task Force. Retrieved from http://www.rfc-archive.org/getrfc.php?rfc=4452 on 5 October 2006.

Wallis, J. C. & Borgman, C. L. (2012). Who is responsible for data? An exploratory study of data authorship, ownership, and responsibility. . Proceedings of the American Society for Information Science and Technology, 48(1). http://onlinelibrary.wiley.com/doi/10.1002/meet.2011.14504801188/full.

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

Wallis, J. C., Pepe, A., Mayernik, M. S. & Borgman, C. L. (2008). An exploration of the life cycle of eScience collaboratory data. 2008 iConference, Los Angeles, CA. http://www.ideals.illinois.edu/handle/2142/15122.

Wallis, J. C., Rolando, E. J. & Borgman, C. L. (in progress). If We Share Data, Will Anyone Use Them? Data sharing and reuse in the long tail of science and technology

Weltzin, J. F., Belote, R. T., Williams, L. T., Keller, J. K. & Engel, E. C. (2006). Authorship in ecology: attribution, accountability, and responsibility. Frontiers in Ecology and the Environment, 4: 435-441.

Whitlock, M. C. (2011). Data archiving in ecology and evolution: best practices. Trends in Ecology & Evolution, 26(2): 61-65.

Wren, J. D., Kozak, K. Z., Johnson, K. R., Deakyne, S. J., Schilling, L. M. & Dellavalle, R. P. (2007). The write position. A survey of perceived contributions to papers based on byline position and number of authors. EMBO Reports, 8(11): 988-991.

Wuchty, S., Jones, B. F. & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. Science, 316(5827): 1036-1039.

Yank, V. & Rennie, D. (1999). Disclosure of Researcher Contributions: A study of original research articles in The Lancet. Internal Medicine, 130: 661-670.

Zimmerman, A. S. (2003). Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists Ph.D Dissertation. School of Information. University of Michigan. Ann Arbor, MI. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/61844 on 28 June 2006.

Zimmerman, A. S. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. International Journal of Digital Libraries, 7(1-2): 5-16.

Zimmerman, A. S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. Science, Technology, & Human Values, 33: 631 - 652.

Zimmerman, A. S. & Nardi, B. (2006). Whither or whether HCI: Requirements analysis for multi-sited, multi-user cyberinfrastructures. CHI 2006, Montreal, Association for Computing Machinery. 1601-1606.

Zuckerman, H. A. (1968). Patterns of Name Ordering Among Authors of Scientific Papers: A study of social symbolism and its ambiguity. American Journal of Sociology, 74(3): 276-291.