Title
Data Management in the Long Tail: Science, Software and Service

Journal
The International Journal of Digital Curation, 11(1)

ISSN
1746-8256

Authors
Borgman, Christine L.
Goshen, Milena S.
Sands, Ashley E.
et al.

Publication Date
2016-05-26

DOI
10.2218/ijdc.v11i1.428

IJDC | *Peer-Reviewed Paper*

Data Management in the Long Tail: Science, Software and Service

Christine L. Borgman
University of California, Los Angeles

Milena S. Golshan
University of California, Los Angeles

Ashley E. Sands
University of California, Los Angeles

Jillian C. Wallis
University of Southern California

Rebekah L. Cummings
University of Utah

Peter T. Darch
University of Illinois
at Urbana-Champaign

Bernadette M. Randles
University of California, Los Angeles

Abstract

Scientists in all fields face challenges in managing and sustaining access to their research data. The larger and longer term the research project, the more likely that scientists are to have resources and dedicated staff to manage their technology and data, leaving those scientists whose work is based on smaller and shorter term projects at a disadvantage. The volume and variety of data to be managed varies by many factors, only two of which are the number of collaborators and length of the project. As part of an NSF project to conceptualize the Institute for Empowering Long Tail Research, we explored opportunities offered by Software as a Service (SaaS). These cloud-based services are popular in business because they reduce costs and labor for technology management, and are gaining ground in scientific environments for similar reasons. We studied three settings where scientists conduct research in small and medium-sized laboratories. Two were NSF Science and Technology Centers (CENS and C-DEBI) and the third was a workshop of natural reserve scientists and managers. These laboratories have highly diverse data and practices, make minimal use of standards for data or metadata, and lack resources for data management or sustaining access to their data, despite recognizing the need. We found that SaaS could address technical needs for basic document creation, analysis, and storage, but did not support the diverse and rapidly changing needs for sophisticated domain-specific tools and services. These are much more challenging knowledge infrastructure requirements that require long-term investments by multiple stakeholders.

Received 23 March 2016 ~ *Accepted* 26 May 2016

Correspondence should be addressed to Christine L. Borgman, UCLA, 235 GSE&IS Building, Box 951520, Los Angeles, CA 90095-1520. Email: christine.borgman@ucla.edu

An earlier version of this paper was presented at the 11th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Data management solutions vary widely across research domains, size and duration of projects, scales of data production, characteristics of data, and project goals (Borgman, 2015; Borgman et al., 2015). The study reported here focused on conceptualizing an Institute for Empowering Long Tail Research (IELTR) and was conducted by researchers in information science and in computer science. The project was funded by the National Science Foundation's program on Software Infrastructure for Sustained Innovation (S2I2). Our premise in this project was that larger and longer term scientific projects have far more resources to devote to data management than do scientists working on smaller and shorter term projects. This was an exploratory project to gauge the opportunities for technical solutions to some of the data management needs faced by 'small science.'

The 'long tail' metaphor originated in commerce to describe the popularity distribution of information products, such as books and movies (Anderson, 2004). This power law distribution also is used to describe scientific activity by size of project and scale of data production (Heidorn, 2008; Wallis, Rolando, and Borgman, 2013). At the head of the curve are 'big science' projects in domains such as astronomy and physics that are large and long-term, and that produce vast volumes of data. At the end of the tail are 'small science' projects, such as those in ecology and earth sciences, that are not as large, are shorter term, and produce less data. However, as we have shown elsewhere, research endeavors in domains such as astronomy, microbiology and earth sciences often are composed of many small projects (Darch et al., 2015; Darch and Sands, 2015). Each of these individual projects may generate and consume data of varying volumes and variety. The long tail metaphor is problematic for many reasons, not least of which is that it reduces the rich complexity of science to two dimensions. Rather, we found the most constructive approach to studying scientists in need of better data management tools and infrastructure was to focus on small and medium sized laboratories (SMLs). These laboratories typically piece together their funding from multiple sources. Each grant might be three years or less in duration and support one or two investigators plus a small cadre of students and post-doctoral fellows. These laboratories may function independently or be part of larger local or distributed collaborations.

Even when small teams conduct science projects in a single laboratory, methods and tools may vary widely. In the same field, each scientist may use different tools and techniques to generate datasets similar in form and intent. Role specialization tends to be limited when one investigator is responsible for all activities in a project (Wallis et al., 2007; Wallis, Borgman, Mayernik, and Pepe, 2008). The scientists who produce such data often are solely responsible for the data management. These researchers usually use localized, ad hoc data formats for immediate data use. Consequently, data are often neglected and lost when no longer needed by the research team (Borgman, Wallis, and Mayernik, 2012; Wallis and Borgman, 2011).

While data management is only partly a technical problem, we sought to understand what aspects of the data management in SMLs were amenable to a particular technical solution. Software as a Service (SaaS) is a subscription-based model of licensing and delivering software. These services are hosted, typically using cloud-based technologies, and accessed via a web browser. SaaS is popular for many business applications, such as payroll, content management, accounting, collaboration, and anti-

 doi:10.2218/ijdc.v11i1.428

virus tools. Applications that are suitable for business and scientific work include writing, note-taking, presenting, and simple data analysis. Common software tools include Google Drive, Evernote, and electronic lab notebooks. Domain-specific SaaS tools can provide more extensive data analysis, visualization, and pipeline-processing capabilities.

The IELTR project asked how, when, and whether SaaS can assist small teams with their data management needs. The data practices of SMLs are diverse, but many requirements are common across communities of practice: gathering, describing, cleaning, analysing data, and publishing findings. Similarly, SMLs are subject to data management requirements of their funding agencies and may have few resources available for acquiring, curating, sharing, and sustaining access to their data. These pressures have led to the pursuit of better data management tools, practices, and infrastructures.

As part of a two-year exploratory project to assess the opportunities for applying SaaS to SMLs, we studied the data management practices of three communities. Two of these were National Science Foundation (NSF) Science and Technology Centers: the Center for Embedded Networked Sensing (CENS) and the Center for Dark Energy Biosphere Investigations (C-DEBI). The third SML is a group of data managers for natural reserve sites who were assembled in a workshop in 2014 by the IELTR study (Brooks, Heidorn, Stahlman, and Chong, 2016).

Our research questions are these: What are the typical characteristics of small and medium-sized laboratories and how do these characteristics vary? Who is involved in supporting the data management tasks of these SMLs? What are the data of these SMLs? How do they perceive their own data management needs? What are their concerns for standardization of data, tasks, and tools? Which of these needs might be addressed by SaaS solutions? The UCLA Center for Knowledge Infrastructures has explored scientific data practices in multiple disciplines for over a decade. We drew upon findings from prior studies and collected new data for this project (Borgman, 2007, 2015; Borgman, Wallis, and Enyedy, 2007; Edwards, Mayernik, Batcheller, Bowker, and Borgman, 2011).

Background and Literature Review

Data, data management, and tools are deeply interconnected in the practices of these domains. To set the larger context, we introduce the concept of knowledge infrastructures, identify challenges in data management for small and medium-sized laboratories, and explore opportunities offered by Software as a Service.

Knowledge Infrastructures

We use the term ‘knowledge infrastructures’ in the sense presented by Edwards (2010) as “robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds.” Neither data practices nor information technology can be studied in isolation. Small teams may recognize the benefits of advanced analysis, data visualization, and data sharing, but lack the resources and expertise to exploit available technology. Some of the scientists who work alone or in small teams become experts in technology, whereas others balk at making these investments. Often, small teams rely on graduate students and post-

doctoral fellows to manage their data. Expertise turns over rapidly as these short-term employees graduate or otherwise complete their training (Borgman, Wallis, and Enyedy, 2007; Eddy, 2005; Edwards et al., 2011; Steinhardt and Jackson, 2014). Data and software are deeply intertwined in research practice. Maintaining them, alone or together, and giving credit for contributions to research data and software, are growing concerns in knowledge infrastructures (Howison and Bullard, 2015; Howison and Herbsleb, 2011, 2013; Uhlir, 2012; Velden et al., 2014).

Small teams may work independently or may be part of distributed collaborations. Those in larger collaborations may have access to shared infrastructure, including advanced tools and specialized expertise. Independent teams tend to rely on a smaller array of known tools and infrastructures. Whereas relational databases may offer needed capabilities for data analysis and management, smaller teams often rely on spreadsheets, whether local, such as Microsoft Excel, or SaaS, such as Google Sheets (Borgman, Wallis, Mayernik, and Pepe, 2007; Mayernik, Wallis, and Borgman, 2013). Excel is the preferred tool of many small teams in the earth sciences due to its ubiquity, despite its weaknesses for specialized data analysis, modelling, and visualization (Strasser and Cruse, 2013).

Data Management

In small research teams, data collection and tools are often adapted to scientific conditions as they evolve. While effective for the task at hand, these practices often result in data that are difficult to disseminate or share. A recent survey revealed that, excluding genetic data, only 8% of ecological data could be found online (Hampton et al., 2013). The Dryad data repository, and agreements by ecology journals to require deposit in Dryad, are attempts to increase data access in this community (Greenberg, 2009; Whitlock, 2011).

Small and medium-sized laboratories often encounter ‘science frictions’ (Edwards et al., 2011) that impede managing, describing, and storing data. Such frictions also slow scientific progress and impede the flow of information within and between research teams. Common frictions include lack of standardization, incomplete or conflicting metadata, intellectual property claims, data sharing practices, and human resource constraints (Edwards, 2010; Edwards et al., 2011; Mayernik, Batcheller, and Borgman, 2011).

Metadata friction, in particular, is an example of how failing to document datasets inhibits data use and reuse (Mayernik et al., 2011). Without metadata, datasets may devolve into spreadsheets of unlabelled rows and columns or into indecipherable strings of numbers. Metadata schemas facilitate discovery and sharing data on a global scale. These are sets of elements – e.g. title, creator, publication date – designed to meet the needs of a particular community. Using structured and extensible metadata schemas are a proposed solution to data access, discovery, and sharing data (Edwards et al., 2011; Getty Research Institute, 2008). However, manual creation of metadata is a high barrier to efficient data management (Mayernik, 2011). Automating metadata attachment, where possible, can improve the amount and quality of metadata use in practice.

Software as a Service (SaaS)

SaaS is a method of software distribution based on cloud computing (Riungu, Taipale, and Smolander, 2010). Cloud computing uses remote networked servers to perform tasks that previously would have been conducted on a local server or personal computer.

 doi:10.2218/ijdc.v11i1.428

Many aspects of data storage and management can be accomplished with SaaS, especially those aspects amenable to standardized practices.

SaaS originated as a business solution (Sun, Zhang, Chen, Zhang, and Liang, 2007). One major benefit of SaaS technology in the commercial sector is the ability to share costs through building economies of scale. SaaS is intended to minimize the time spent installing, upgrading, and maintaining tools. Such savings are a major reason for SaaS adoption. However, costs vary considerably between software functions and economic tradeoffs between supporting local data management infrastructure and investing in SaaS are largely unstudied (Benlian and Hess, 2011).

SaaS services subsequently expanded to scientific applications such as such as 'Data as a Service' (DaaS) and the Globus project for grid computing (Allen et al., 2012; Chervenak, Foster, Kesselman, Salisbury, and Tuecke, 2000; Dai, Gao, Guo, Xiao, and Zhang, 2012; Madduri et al., 2013).

SaaS applications are hosted by distributed systems and typically accessed via a web browser. Many SaaS applications, such as Google Docs and DropBox, are based on a 'freemium' (free and premium) model. Basic services are available at no cost and additional storage and services are offered for a fee. While many SaaS applications are developed for personal and business usage, others are intended to serve the research community. Some of these are open source and others offer an array of free and fee-based services. SaaS tools popular in scientific communities include Figshare, Dropbox, Github, Omeka, and Zotero. In a typical SaaS implementation, a company or non-profit organization provides software applications, technical support, and maintains the applications. From the users' standpoint, the essential difference between standalone software and SaaS is that the applications reside in the cloud rather than on users' machines. SaaS providers automatically update software versions, which reduces users' technical workload, while also reducing their autonomy and control (Godse and Mulik, 2009). A particular advantage of SaaS software is that most applications are platform-independent, thus partners on different operating systems (e.g., Apple, Microsoft, Linux) can share the same applications and data resources (Sukow and Grant, 2013). Since researchers in SMLs generally maintain their own technology and data management, they may accrue labor and financial benefits by transferring these tasks to third parties.

Useful SaaS tools for SMLs must accommodate the researchers' relative lack of specialized resources and information technology expertise. One obvious area for SaaS applications is spreadsheets, which are commonly used for data manipulation in SMLs. Web-based relational databases are more sophisticated tools that would accelerate analysis and discovery within labs without the need for a dedicated IT staff (Howe et al., 2011). The ability to transfer data from multiple sources to facilitate analysis also could be helpful in dealing with a variety of instruments and software. SaaS tools could offer advanced analytic, visualization, and information management tools that are too expensive for SMLs to acquire, host, and maintain for exclusive use. An economic challenge is to assess what services SMLs are willing to purchase and those they expect to be available free of charge.

Sites and Methods

To study characteristics and data management needs of small and medium-sized laboratories and to identify which of these needs may be amenable to SaaS solutions,

we reanalysed our previous research data and conducted new research. The findings presented in this article are based on:

- A reanalysis of eleven years of research (2002-2012) conducted at CENS (Borgman, Wallis, and Enyedy, 2007; Wallis et al., 2007, 2013; Wallis and Borgman, 2011),
- Two years (2012-2014) of ethnographic participant observation and interviews on data practices at C-DEBI (Darch et al., 2015; Darch and Borgman, 2014; Darch and Sands, 2015), and
- A 2014 two-day workshop organized by the IELTR team with practicing ecologists.

Methods included ethnographic participant observation, semi-structured interviews, and document analysis. The specific methods used for each of the three sites are noted in Table 1.

Table 1. Interviews and ethnographic participant observation across the three research sites.

Research Sites	Interviews	Ethnographic Participant Observation
CENS	77	11 years
C-DEBI	49	2 years
IELTR Workshop	N/A	2 days

Our existing corpus of CENS interviews, ethnographic field notes, and document analysis was collected from 2006 to 2012. We mined these existing CENS data and findings, including a sample of 30 publications by our UCLA Center for Knowledge Infrastructures from 2006 to 2013. These publications drew on 77 interviews with CENS faculty, students, and research and administrative staff; 11 years of involvement in the CENS community as participant observers during meetings and data collection efforts; and analyses of documents, such as publications, reports, and emails between members.

Our studies of C-DEBI began in August 2012. The first year of C-DEBI ethnography provided a framework to design the IELTR study at one of the small laboratories that received funding from C-DEBI. Specific to the IELTR study, we conducted six interviews in 2013 with faculty, students, and research and administrative staff from that C-DEBI-funded laboratory to examine their data infrastructure needs. The research reported here draws upon those interviews, two years of ethnographic participant observation within the C-DEBI community, and analyses of C-DEBI documents. From 2012 through the end of the IELTR study in 2014, we conducted 49 interviews (43 for our continuing C-DEBI research plus six for the IELTR study).

Following the long-term CENS study and shorter term C-DEBI study, we used the 2014 workshop to test hypotheses from the prior work. Three of the authors participated in the workshop, observing, taking both procedural and ethnographic notes from the various sessions, and leading discussions. Through our interactions at the workshop, we identified similarities between the data challenges discussed at the workshop and those faced by scientists working at the CENS and C-DEBI laboratories.

 doi:10.2218/ijdc.v11i1.428

Research Sites

The three research sites for this study are described below. These are the Center for Embedded Networked Sensing (CENS), the Center for Dark Energy Biosphere Investigations (C-DEBI), and a workshop for natural reserve sites organized by the Institute for Empowering Long Tail Research (IELTR) project team, of which the UCLA Center for Knowledge Infrastructures was a partner.

Center for Embedded Networked Sensing (CENS)

CENS, an NSF Science and Technology Center (STC), spanned multiple institutions and disciplines. Established in 2002, CENS brought together faculty, students, and staff from five Southern California universities. The mission of the Center was to develop embedded networked sensing technologies for use in various scientific application areas. At its peak, CENS had 300 members, with roughly 80% of the researchers coming from computer science and engineering, and the other 20% coming from scientific domain areas that included ecology, environmental engineering, marine biology, seismology, and public health. Other research expanded into social arenas through the use of mobile phone platforms and participatory sensing. About 66 teams submitted annual reports to CENS in the last year of the Center, for which funding ended in 2013 (after a one-year no-cost time extension).

While CENS appeared to be a single entity, a large number of small teams conducted research and managed their own data. CENS collaborations were interdisciplinary, requiring data exchange between scientists and technologists with different backgrounds, methods, and expectations (Borgman et al., 2012). Embedded networked sensing systems developed at CENS generated a deluge of streaming data that overwhelmed traditional methods for data collection, analysis, and storage (Borgman, Wallis, Mayernik, et al., 2007). CENS researchers experienced the data deluge ahead of their scientific peers by experimenting with research-grade automated sensing systems that collected data at a higher density in space and time than the hand-collection methods in common use in most of these sciences.

Center for Dark Energy Biosphere Investigations (C-DEBI)

C-DEBI is an NSF STC that began operations in 2010. Its overall goal is to create a community of researchers to study subseafloor microbial life and the interactions between that life and the physical environment in which it resides. The multidisciplinary Center includes disparate communities of the life and physical scientists, including microbiologists, geochemists, geologists, hydrologists, mineralogists, and sedimentologists. The project is geographically distributed, with project leaders at five US universities and C-DEBI funding activities supporting more than 80 scientists in over 50 universities and organizations in the USA, Europe and Asia (Darch and Borgman, 2014). As of this writing, C-DEBI has supported more than 100 scientists around the world.

Like CENS, C-DEBI brings together multidisciplinary teams of scientists whose work involves producing, processing, analysing, and comparing data about subseafloor microbial life from samples collected mostly through research cruises. Aside from ensuring funding and instruments for research cruises, sample collection, and data production, C-DEBI also provides direct funding to individual scientists and small teams to build a deep subseafloor biosphere research community (Darch et al., 2015).

Workshop for Natural Reserve Sites

This workshop was organized as part of our NSF S2I2-funded project for conceptualizing an Institute for Empowering Long Tail Research (IELTR). The workshop 'Datasphere at the Biosphere II' took place May 5-6, 2014 at the Biosphere II near Tucson, Arizona. Through open-ended questions and neutral facilitation, the workshop helped shape a vision of an IELTR and generated feedback from the earth sciences community on how SaaS might offer solutions to their data management challenges (Brooks, Heidorn, Stahlman, and Chong, 2016; Heidorn, Stahlman, and Chong, 2015). The workshop brought together 19 researchers, among whom were seven faculty members, eight researchers from field stations and laboratories, and three data managers. The participants gathered from 13 U.S. states and included one international partner.

The workshop incorporated plenary and breakout brainstorming sessions. The opening session elicited feedback from the earth science community on what they consider as their data and how they use data in their research. Subsequent questions identified challenges faced in processing, cleaning, appraising, selecting, moving, preserving, and storing data.

Results

We describe the data, data management practices, resources, and expertise needed for each case. The findings examine data and knowledge infrastructure concerns that could be addressed by Software as a Service or other technical solutions.

Center for Embedded Networked Sensing (CENS)

CENS findings address data characteristics, challenges in data management, and necessary resources and expertise.

CENS data

Researchers at CENS collected a wide variety of observational data with methods that ranged from hand measures to physical, chemical, and biological sensors. They captured both phenomena in the biological and geophysical world and observations about how the sensing systems and their robotic platforms functioned.

We discovered that definitions of data, even when referring to the same data artifact, differed from person to person, sometimes within the same research group (Borgman, 2012, 2015; Borgman, Wallis, and Enyedy, 2007). For instance, technologists used hand-collected scientific data only to check the accuracy of their sensor readings, while the domain scientists might consider the hand-collected data as the most important evidence for their research questions (Borgman, 2009; Wallis et al., 2007). The different types of data were managed separately at CENS, with each team controlling the data necessary to conduct their immediate research.

CENS data management challenges

A fundamental challenge in the CENS research was the reliability of the technologies and the availability of connectivity. In the early years of CENS, battery life in the sensors was a major concern. Sensors would fail suddenly or gradually, rebooting themselves at seemingly random times. Some of the sensor deployments had steady

 doi:10.2218/ijdc.v11i1.428

access to the Internet, while many of the field deployments occurred at remote sites without connectivity. In those cases, sensor data were stored offline on laptops or other devices for days or weeks at a time. In other cases, data were collected manually from data loggers, visited periodically by CENS personnel (Hamilton et al., 2007; Mayernik et al., 2013; Rahimi, Hansen, Kaiser, Sukhatme, and Estrin, 2005; Wallis et al., 2007).

Over the course of a decade, multiple tools were developed to manage CENS data including SensorBase (Chang, Yau, Hansen, and Estrin, 2006), the CENS Deployment Center (Mayernik, Wallis, Borgman, and Pepe, 2007), and the UCLA Data Registry (Mandell, 2012a, 2012b). Each of these tools was built in a different stage of the project and to serve different purposes. SensorBase was a 'raw data' repository that incorporated RSS feed technology for automatic streaming of data from sensors to remote databases (Chang et al., 2006; Pepe, Borgman, Wallis, and Mayernik, 2007). SensorBase proved to be an effective tool for the several teams that used it to collect and organize data. These were teams who had full connectivity on their sensor networks that made streaming data possible. SensorBase was easy to adopt because it imposed no constraints on data organization. However, the price of such flexibility was a lack of interoperable metadata schemas or other standards in common use by domain-specific repositories. The data were never open to the CENS community because the system defaulted to password protection.

The CENS Deployment Center (CENSDC) aimed to capture, manage, and reuse relevant information about field activities. It allowed scientists to plan their field deployments – including dates, location, data collectors, equipment, calibration, data to be collected, tasks, and digital notes – within a web-based relational database accessible to all CENS members. This tool was adopted by several of the CENS research groups as a means to stay current with continuing changes in field deployments and personnel turnover. CENSDC had a mix of online and offline features to ensure its usefulness in field conditions.

The CENS Data Registry, which became the UCLA Data Registry, provided a structured way for researchers to make their data visible in the absence of discipline-specific repositories (Mandell, 2012a, 2012b; Wallis et al., 2013). The Data Registry improved data visibility by making the metadata and contact information associated with a dataset publicly available while the data stayed with the researcher. In contrast, SensorBase captured data from sensors and CENSDC captured field practices associated with data. While each of these tools was used by a few CENS teams, none were widely adopted across the Center.

The need to create metadata was a barrier to adoption for each of these tools. The progression of data management approaches from SensorBase to the CENS Data Registry involved decreasing the amount of metadata a researcher would need to provide. However, even the Data Registry apparently asked for more metadata than most CENS researchers were willing to provide. Only eight of the 66 groups registered datasets during the last year of the Center.

Although a single Center, CENS was an amalgam of SMLs with many disciplines, domains, projects, and research goals. Data and data practices varied too widely for CENS-wide data or metadata standards to be adopted (Shankar, 2003; Wallis, Milojevic, Borgman, and Sandoval, 2006).

CENS data management resources and expertise

Most research funding was disbursed to individual laboratories within CENS. As an NSF Science and Technology Center, CENS received about $40 million dollars in base funding during the Center's ten years. In addition to purchasing equipment and

employing graduate students as a part of core CENS research, these funds were used to hire support personnel to cover essential functions, such as accounting, educational programming, personnel, project administration, purchasing, and technical support for equipment and servers. However, no funds were allocated specifically for data management, curation, or a repository. The CENS statistics team developed SensorBase as an experimental project. Similarly, CENSDC and the Data Registry were developed by the data practices team under external funding as part of their studies.

Researchers at CENS managed their data for the use of their small teams. In most cases, these local approaches did not lead to long-term persistence and access. Data management in CENS became more difficult as datasets increased in complexity, size, and granularity (Borgman et al., 2012; Wallis, 2012).

Center for Dark Energy Biosphere Investigations (C-DEBI)

C-DEBI findings are organized parallel to those of CENS, in terms of data characteristics, challenges in data management, and necessary resources and expertise.

C-DEBI data

Researchers at C-DEBI generate data by collecting and analysing physical samples, such as sediments and portions of the basaltic crust, or water. Scientific ocean drilling cruises, or expeditions, conducted by the Integrated Ocean Drilling Program (IODP) in the period 2003-2013 were the major sources of these samples. The organization of these IODP cruises required expedition participants to process cores on the ship with IODP curators. The products from such expeditions become a variety of biological and physical data that can be described differently for the aims of various scientific questions across disciplines, locations, and over time (Darch and Borgman, 2014). The data from initial processing onboard an IODP ship are described and stored under the supervision of professional curators. However, data produced from further analyses of these samples in the onshore laboratories of C-DEBI-affiliated scientists have typical small science characteristics: heterogeneous, documented using local practices, and produced and stored according to localized standards. Because C-DEBI is a cross-disciplinary project, a variety of new methods for data collection and analysis are emerging in response to changing needs.

The notion of data in C-DEBI varies by laboratories and individual scientists. The laboratory we observed for the IELTR study uses a variety of instruments, such as gas chromatographs and microsensors, each of which has different software and is connected to a separate lab computer. Incompatible software forced those working in the laboratory to store data on different computers, despite their preference to keep all the data in one place. The head of this C-DEBI laboratory explained that measurements taken in the laboratory are data to lab members. To ecologists, whose primary data are specimens and measurements taken on board the IODP cruises, these laboratory data might be viewed as noise.

C-DEBI data management needs

C-DEBI-affiliated scientists produce a vast variety of data, each type of which may require different instruments and software. Complex data management needs often result in the use of myriad tools and systems rather than committing to a single technical environment. Members of C-DEBI identified the need for software, more efficient recording and storing of data, data analysis, and better security. As one researcher

 doi:10.2218/ijdc.v11i1.428

stated: "there has to be a better system [than]… trying to keep track of everything yourself without having a really designated, secure, structured place."

Data generated by SMLs may not be comprehensible to people outside the team for reasons such as a lack of common vocabulary, methods, and other data practices. Interviewees at this small lab affiliated with C-DEBI viewed lack of metadata as a barrier to finding, obtaining, or using data. As one member stated: "nobody knows what we're talking about unless you're talking to your lab partner." Some C-DEBI participants expressed a desire for a system that would keep all the data in one place alongside detailed metadata. Their overall goal was to enable scientists to reanalyse and reuse those data.

Large initiatives, such as the Integrated Ocean Drilling Program, address interoperability issues by mandating onboard data description with standardized metadata schemas. The data description process is supervised by curators responsible for keeping records of all the samples (IODP, 2012). While these data are made publicly available after an embargo period, the metadata of all data objects are available to everyone without any restrictions (IODP, 2009). The IODP database is the main data source for this community.

C-DEBI data management resources and expertise

Most of the small laboratories affiliated with C-DEBI-funded scientists lack the resources to employ designated technology staff. When laboratory members cannot manage software problems themselves, they call for external information technology support from their department or university.

Other patterns of software usage in C-DEBI were similar to those of CENS, which did have some technology support in house. Many C-DEBI laboratory members we studied relied on Excel to store data and on Dropbox and email to share their data. Researchers found these tools to be user-friendly, sufficient for their data volume, and widely adopted within their science community. However, interviewees recognized the limits of these current tools for complex data management. Some researchers wished to access their data from multiple computers and operating systems. They expressed a desire for better visualization, control, and long-term storage and access.

These scientists expressed a desire for assistance in describing and analysing data and using appropriate software:

> "If money was no object, I would have specific personnel that could do this stuff so I could do the lab work and the writing … [I] would like to have help with the software and … computational side of things."

Several C-DEBI-affiliated scientists desired a centralized system that would hold their data in one place, and would be compatible with lab instruments to facilitate data transfer and storage. One researcher mentioned the existence of laboratory infrastructure for data management as an advantage in applying for additional funding.

We found limited infrastructure available for collecting and storing data at these SMLs. Researchers keep their data on their hard drive or in some cases they were not sure exactly where their data were located; data might be stored in inaccessible and vulnerable personal computers: "we have one particular computer that has some stuff on it, but I don't know what's there." Those who had experienced working with a data curator onboard IODP cruises expressed how useful it would be to have that kind of expertise in their labs, but viewed the option as prohibitively expensive. These

researchers understood the benefits of a structured system and the difficulties in accommodating such vastly heterogeneous data:

> "The best data storage I have seen is actually in connection with the ocean drilling. … it took them decades to come up with the system, but they are doing a good job … with everything, standard protocols, standard procedures, standard storage, standard everything. It makes it a little bit rigid [and is] complicated to find one system that works for everything."

As with CENS, C-DEBI would have difficulties adopting a single software system because of the multidisciplinary research and variety of data they collect.

Workshop for Natural Reserve Sites

Several common patterns emerged from the workshop discussions. We grouped them by data sources and types, data management needs, data management resources and expertise, and opportunities for SaaS tools.

Workshop data sources and types

Workshop participants used continuous high-frequency data from sensors, data from modeling, permanent plot data, observation and analysis data, and legacy data, such as images, spreadsheets, and handwritten data from field and lab notebooks. Some participants emphasized the importance of distinguishing between characteristics of sensor- and human-collected data. Sensor data are automatic and standardized, whereas human-collected data are richer but harder to combine. Participants desired more standardized data descriptions and methods to transform old data to new formats for reuse. Many participants used external sources of data, such as GenBank.

Workshop data management needs

Data management tasks, such as production and sharing, were deemed challenging. These included data collection, preservation, sharing, and access. The need for data integration was a high priority for most workshop participants. Environmental scientists emphasized the need for data standardization in projects that collect a variety of data types in different media. Due to the heterogeneity of their data and the variety of instruments, many researchers use local formats and local methods to describe data. Merging data while maintaining data integrity is a particular challenge. Researchers need software to analyse data at multiple levels of processing, which vary by type of data and by institution.

Closely related to data integration are concerns for discovery and access. Workshop participants expressed difficulty locating data and databases. Researchers needed more metadata, visualization, taxonomies, and unique identifiers to enhance the ability to locate, access, and retrieve data. The need for visualization, GIS mapping, and modeling were mentioned multiple times. One participant stated: "even when we create a data catalog people don't use it; visualization would help."

Large amounts of previously collected data cannot be exploited due to obsolete formats. One priority was to migrate data from legacy databases that were inherited from previous research projects. These usually were based on older software that is no longer available. Another priority for data access is to digitize and integrate handwritten data from lab and field notebooks and other forms of legacy data.

 doi:10.2218/ijdc.v11i1.428

Workshop data management resources and expertise

Workshop participants identified stakeholders such as science communities, policy makers, businesses, educators, and the general public. They emphasized that long-term ecological studies need long-term support, but they typically get only one to three years of funding for a project. Agencies typically provide support for fieldwork, which is the shortest and the least costly part of these research endeavors. The process of analysis, description, and writing results takes longer and is more costly, but must be done with limited funds.

Some participants would prefer software with more sophisticated capabilities, such as Illumina for genetic data, but were wary of the learning curve for implementing new tools. One participant's organization hired consultants to write code that would upload data seamlessly into a state-owned database. Access to such expertise is expensive.

Small field stations often lack Internet connectivity, which limits the usefulness of SaaS. As one participant explained: "weather data from sensors, field notebooks, community composition, data from loggers (batch data) can't record and upload." Many of these sites are in remote areas of mountains and deserts, far from telecommunications networks. Some may have connectivity in the main building on the natural reserve site, but not a wide area network that reaches all data collection points. Solar power tends to be the primary source of electricity for remote natural reserve sites.

Workshop opportunities for SaaS tools

During the second day of the workshop, the IELTR team invited the participants to envision specific SaaS tools necessary to address data management needs of their individual sites and research projects. Workshop attendees identified a range of tools they hoped SaaS could provide. Desired functionalities included allowing multiple types of user accounts for better data control and assistance generating metadata. They need capabilities such as version control, quality control, annotations, and provenance information. Participants wished for help with hardware, including backup management and automated capture of metadata. They also requested assistance with scalability and workflow synchronization among team members for the medium to long term.

Discussion

This study investigated the data management needs of small and medium-sized laboratories (SMLs) to assess how, when, and whether Software as a Service (SaaS) might address those needs. Our findings identify the complex array of data management problems in SMLs, including lack of resources, expertise, metadata, and standardization. Here we consider which data management problems in these research sites may and may not be amenable to SaaS solutions.

Data management needs in common across the three sites studied – two National Science Foundation Science and Technology Centers and one workshop of natural reserve site managers – spanned the research life cycle. Concerns ranged from the earliest stages of gathering or collecting data, through cleaning, describing, analysing, and storing data, to publishing their findings. The two STCs were most concerned for managing data for current use, whereas the natural reserve managers were most concerned with long-term data management for integration and reuse. This comparison of different kinds of scientific needs is important for understanding the full scope of data management for small- and medium-sized laboratories. In the short term, all three

sites were concerned with metadata, standards, and common tools that could be used to integrate data from multiple sources, whether their own data or those of their collaborators.

While both STCs received ten years of funding support (two five-year awards), most individual projects within these centers were conducted with grants of one to three years in length. The natural reserve sites, in contrast, may exist for decades at a time. However, individual research studies at these sites also tend to be based on one- to three-year grants. CENS, whose funding preceded NSF data management plan requirements, had relatively little concern for maintaining data beyond the life of the Center. Management defaulted to the individual teams that collected the data.

C-DEBI, founded eight years after CENS, is more concerned with developing centralized and standardized data management methods. The most standardized data production for C-DEBI occurs in the IODP expeditions. Once specimens reach the C-DEBI laboratories, most data practices and management default to local tools and storage mechanisms. C-DEBI-affiliated scientists are concerned with long-term sustainability of their data as they endeavor to build a new field of research. However, the means to establish common approaches to data management are a source of tension within this community (Borgman et al., 2015; Darch et al., 2015; Darch and Borgman, 2014; Darch and Sands, 2015).

Natural reserve managers most consistently expressed their concerns for long term data sustainability. They are faced with legacy data inherited from prior investigators, students, post-doctoral fellows, and staff. These legacy data exist in numerous formats of audio, video, images, numbers, and text, and reside on a similarly disparate array of hardware and software. Data migration and integration were their greatest challenges.

Natural reserve site managers were more explicit about the balance of labor between data collection, analysis, and management than were members of CENS and C-DEBI. Their grants fund the labor and equipment to collect scientific data. They find that the labor and expertise to exploit those data for findings and publication is a much longer and more expensive process. Thus, they continually assemble funding to analyse extant data or to integrate existing data with new data. We found similar accumulation and transfer of data in our observations of CENS and C-DEBI. The competing time frames of grant-funded projects, staff time, and student work complicate their ability to sustain the usefulness of data over long periods of time (Edwards et al., 2011; Jackson, Ribes, Buyuktur, and Bowker, 2011).

All of the small and medium-sized laboratories studied were constrained by lack of technical expertise on their teams, technical support from their institutions, and overall availability of human resources. Their data were scattered across many computers, platforms, and tools acquired over long periods of time. Often the tools and platforms were incompatible, both within their laboratories and between their laboratories and their collaborators. Similarly, their data and their practices were highly diverse due to the broad range of disciplinary expertise and expansive array of research questions being studied. A common concern was the desire for more sophisticated tools than those to which they had access or could afford.

Some of these data management problems are amenable to technical solutions such as those afforded by SaaS technologies. Obvious tools are those that support basic document production, data analysis, and storage. SaaS document services that are free or freemium include Google Drive (formerly Google Docs), Office Online (Microsoft), Open Office, and newer services that are designed for academic publishing, such as Authorea and Overleaf. Spreadsheets included in the generic document services offer basic data analysis capabilities, and web-based databases are more sophisticated

 doi:10.2218/ijdc.v11i1.428

options. Storage options are growing rapidly, not only with Dropbox and Box, but cloud services for larger volumes of data offered by Amazon, Microsoft, and others. These services alleviate some of the needs for local IT support by being platform-independent and including continuous upgrades and security patches. Offline use of SaaS technologies is a decreasing problem to the extent that these applications can operate on local devices and merge data when connectivity becomes available.

Metadata and standards concerns can be harder to address with SaaS services unless well tailored to the domain. To the extent that basic metadata can be generated automatically from data collection devices, these can be captured by software technologies. However, automated metadata may be largely syntactic, such as time stamps, measurements, instrument calibrations, and so on. Semantic metadata that adds meaning to data remains a largely human endeavor. The larger needs for sophisticated tools that are tailored to domain-specific environments, especially those that can support heterogeneous data from disparate sources, are not readily amenable to technical solutions, SaaS or otherwise.

Conclusion

Research teams in small- and medium-sized laboratories struggle with a data deluge and with a lack of sufficient human expertise and knowledge infrastructures to manage their data. SaaS has helped small- and medium-sized businesses compete with larger organizations that can afford dedicated infrastructures. We investigated the potential of SaaS applications to address data management needs in small- and medium-sized scientific laboratories.

The basic features of SaaS technologies are attractive for these kinds of scientific environments because they can reduce costs, labor, and the need for local technical support. Basic SaaS tools designed for business applications, such as document creation and spreadsheets, already are in wide scientific use. To the extent that SaaS scientific tools can support large enough communities to be economically viable, they are finding users in areas such as bioinformatics and genomics.

The harder problems in data sustainability faced by small and medium sized laboratories tend not to be amenable to technical solutions. These are the larger knowledge infrastructure needs for highly specialized tools and expertise in rapidly changing environments. Expertise in both scientific domains and in specific tools tends to turn over very quickly due to rapidly evolving research fronts, the cyclical nature of student projects and post-doctoral fellow appointments, and funding cycles that are much shorter than the time frame in which the data remain scientifically useful. Investments in scientific practice, by funding agencies, educators, and researchers, must take a much longer term view of sustaining access to scientific data (Borgman, 2015). To quote Jim Gray (National Research Council, 2004): “may all your problems be technical.”

Acknowledgements

This research was funded by the National Science Foundation Award #1216754: Collaborative Research: Conceptualizing an Institute for Empowering Long Tail Research. Ian Foster, PI; co-PIs Christine L. Borgman, Bryan Heidorn, Bill Howe, and

Carl Kesselman. We thank our collaborators on the IELTR project and other partners in the UCLA Center for Knowledge Infrastructures for comments on earlier drafts of this paper: Ian Foster, Bill Howe, Irene Pasquetto, and Sharon Traweek. All authors of this paper were affiliated with the UCLA Center for Knowledge Infrastructures during the time they worked on this project. Jillian C. Wallis is now at the University of Southern California, Rebekah L. Cummings is at the University of Utah, and Peter T. Darch is at the University of Illinois at Urbana-Champaign.

References

Allen, B., Pickett, K., Tuecke, S., Bresnahan, J., Childers, L., Foster, I., … Martin, S. (2012). Software as a service for data scientists. *Communications of the ACM, 55*(2), 81. doi:10.1145/2076450.2076468

Anderson, C. (2004). The long tail. *Wired Magazine, 12*(10). Retrieved from http://www.wired.com/wired/archive/12.10/tail_pr.html

Benlian, A., & Hess, T. (2011). Opportunities and risks of software-as-a-service: Findings from a survey of IT executives. *Decision Support Systems, 52*(1), 232–246. doi:10.1016/j.dss.2011.07.007

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

Borgman, C.L. (2009). The digital future is now: A call to action for the humanities. *Digital Humanities Quarterly, 3*. Retrieved from http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html

Borgman, C.L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059–1078. doi:10.1002/asi.22634

Borgman, C.L. (2015). *Big data, little data, no data: Scholarship in the networked world.* Cambridge, MA: The MIT Press.

Borgman, C.L., Darch, P.T., Sands, A.E., Pasquetto, I.V., Golshan, M.S., Wallis, J.C., & Traweek, S. (2015). Knowledge infrastructures in science: Data, diversity, and digital libraries. *International Journal on Digital Libraries, 16*(3-4), 207–227. doi:10.1007/s00799-015-0157-z

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), 17–30. doi:10.1007/s00799-007-0022-9

Borgman, C.L., Wallis, J.C., & Mayernik, M.S. (2012). Who's got the data? Interdependencies in science and technology collaborations. *Computer Supported Cooperative Work, 21*(6), 485–523. doi:10.1007/s10606-012-9169-z

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. In Joint Conference on Digital Libraries. Vancouver, British Columbia, Canada: Association for Computing Machinery. doi:10.1145/1255175.1255228

Brooks, C.F., Heidorn, P.B., Stahlman, G.R., & Chong, S.S. (2016). Working beyond the confines of academic discipline to resolve a real-world problem: A community of scientists discussing long-tail data in the cloud. *First Monday, 21*(2). Retrieved from http://ojs-prod-lib.cc.uic.edu/ojs/index.php/fm/article/view/6103

Chang, K., Yau, N., Hansen, M., & Estrin, D. (2006). SensorBase.org – A centralized repository to Slog sensor network data. In Proceedings of the International Conference on Distributed Networks (DCOSS)/EAWMS. Retrieved from http://escholarship.org/uc/item/4dt82690

Chervenak, A., Foster, I., Kesselman, C., Salisbury, C., & Tuecke, S. (2000). The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets. *Journal of Network and Computer Applications, 23*(3), 187–200. doi:10.1006/jnca.2000.0110

National Research Council (2004). *Computer Science: Reflections on the Field, Reflections from the Field.* Washington, D.C.: National Academy Press. Retrieved from http://www.nap.edu/catalog/11106.html

Dai, L., Gao, X., Guo, Y., Xiao, J., & Zhang, Z. (2012). Bioinformatics clouds for big data manipulation. *Biology Direct, 7*, 43. doi:10.1186/1745-6150-7-43

Darch, P.T., & Borgman, C.L. (2014). Ship space to database: Motivations to manage research data for the deep subseafloor biosphere. In Proceedings of the 77th Annual Meeting of the Association for Information Science and Technology. Seattle, WA. Retrieved from http://www.asis.org/asist2014/proceedings/submissions/papers/156paper.pdf

Darch, P.T., Borgman, C.L., Traweek, S., Cummings, R.L., Wallis, J.C., & Sands, A.E. (2015). What lies beneath?: Knowledge infrastructures in the subseafloor biosphere and beyond. *International Journal on Digital Libraries, 16*(1), 61–77. doi:10.1007/s00799-015-0137-3

Darch, P.T., & Sands, A.E. (2015). Beyond big or little science: Understanding data lifecycles in astronomy and the deep subseafloor biosphere. In iConference 2015 Proceedings. Newport Beach, CA: iSchools. Retrieved from https://www.ideals.illinois.edu/handle/2142/73655

Eddy, S.R. (2005). "Antedisciplinary" science. *PLoS Computer Biology, 1*(1), e6. doi:10.1371/journal.pcbi.0010006

Edwards, P.N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming.* Cambridge, MA: The MIT Press.

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. doi:10.1177/0306312711413314

Getty Research Institute. (2008). *Introduction to metadata. (M. Baca, Ed.) (2nd ed)*. Los Angeles, CA: Getty Research Institute. Retrieved from http://www.getty.edu/research/publications/electronic_publications/intrometadata/index.html

Godse, M., & Mulik, S. (2009). An approach for selecting Software-as-a-Service (SaaS) product. In 2013 IEEE Sixth International Conference on Cloud Computing (pp. 155–158). Los Alamitos, CA, USA: IEEE Computer Society. doi:10.1109/CLOUD.2009.74

Greenberg, J. (2009). Theoretical considerations of lifecycle modeling: An analysis of the Dryad repository demonstrating automatic metadata propagation, inheritance, and value system adoption. *Cataloging and Classification Quarterly, 47*(3-4), 380–402. doi:10.1080/01639370902737547

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. doi:10.1089/ees.2006.0045

Hampton, S.E., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller, A.L., … Porter, J.H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment, 11*(3), 156–162. doi:10.1890/120103

Heidorn, P.B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends, 57*(2), 280–299. doi:10.1353/lib.0.0036

Heidorn, P.B., Stahlman, G.R., & Chong, S. (2015). Datasphere at the Biosphere II: Computation and data in the wild. Retrieved from https://www.ideals.illinois.edu/handle/2142/73759

Howe, B., Cole, G., Souroush, E., Koutris, P., Key, A., Khoussainova, N., & Battle, L. (2011). Database-as-a-service for long-tail science. In J.B. Cushing, J. French, & S. Bowers (Eds.), *Scientific and Statistical Database Management* (pp. 480–489). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-22351-8_31

Howison, J., & Bullard, J. (2015). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology.* doi:10.1002/asi.23538

Howison, J., & Herbsleb, J.D. (2011). Scientific software production: Incentives and collaboration. In Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, pp. 513–522. New York, NY, USA: ACM. doi:10.1145/1958824.1958904

Howison, J., & Herbsleb, J.D. (2013). Incentives and integration in scientific software production. In Proceedings of the 2013 Conference on Computer Supported Cooperative Work, pp. 459–470. New York, NY, USA: ACM. doi:10.1145/2441776.2441828

IODP. (2009). IODP site survey data confidentiality policy. Retrieved from https://www.iodp.org/iodp-site-survey-confidentiality-policy/file

IODP. (2012). IODP sample, data, and obligations policy. Retrieved from https://www.iodp.org/iodp-sample-data-and-obligations-policy/file

Jackson, S.J., Ribes, D., Buyuktur, A., & Bowker, G.C. (2011). Collaborative rhythm: Temporal dissonance and alignment in collaborative scientific Work. In Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, pp. 245–254. New York, NY, USA: ACM. doi:10.1145/1958824.1958861

Madduri, R.K., Dave, P., Sulakhe, D., Lacinski, L., Liu, B., & Foster, I.T. (2013). Experiences in building a next-generation sequencing analysis service using Galaxy, Globus Online and Amazon Web Service. In Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery, pp. 34:1–34:3. New York, NY, USA: ACM. doi:10.1145/2484762.2484827

Mandell, R.A. (2012a). Researchers' attitudes towards data discovery: Implications for a UCLA data registry (SSRN Scholarly Paper No. ID 2129539). Rochester, NY: Social Science Research Network. doi:10.2139/ssrn.2129539

Mandell, R.A. (2012b). Researchers' attitudes towards data discovery: Implications for a UCLA data registry. In Libraries in the Digital Age (LIDA) Proceedings, 12. Zadar, Croatia. Retrieved from http://ozk.unizd.hr/proceedings/index.php/lida2012/article/view/59/43

Mayernik, M.S. (2011). Metadata realities for cyberinfrastructure: Data authors as metadata creators (PhD Dissertation). UCLA, Los Angeles, CA. doi:10.2139/ssrn.2042653

Mayernik, M.S., Batcheller, A.L., & Borgman, C.L. (2011). How institutional factors influence the creation of scientific metadata. In Proceedings of the 2011 iConference, pp. 417–425. Seattle, WA: Association for Computing Machinery. doi:10.1145/1940761.1940818

Mayernik, M.S., Wallis, J.C., & Borgman, C.L. (2013). Unearthing the infrastructure: Humans and sensors in field-based research. *Computer Supported Cooperative Work, 22*(1), 65–101. doi:10.1007/s10606-012-9178-y

Mayernik, M.S., Wallis, J.C., Borgman, C.L., & Pepe, A. (2007). Adding context to content: The CENS deployment center. *Annual Meeting of the American Society for Information Science and Technology, 44*, pp. 1–7. Milwaukee, WI: Information Today, Inc. doi:10.1002/meet.1450440388

Pepe, A., Borgman, C.L., Wallis, J.C., & Mayernik, M.S. (2007). Knitting a fabric of sensor data resources. In Proceedings of the 2007 ACM IEEE International Conference on Information Processing in Sensor Networks. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.596.2608

Rahimi, M., Hansen, M., Kaiser, W.J., Sukhatme, G.S., & Estrin, D. (2005). Adaptive sampling for environmental field estimation using robotic sensors. In 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005), pp. 3692–3698. doi:10.1109/IROS.2005.1545070

Riungu, L.M., Taipale, O., & Smolander, K. (2010). Research issues for software testing in the cloud. In 2010 IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom), pp. 557–564. Indianapolis, IN: IEEE. doi:10.1109/CloudCom.2010.58

Shankar, K. (2003). Scientific data archiving: The state of the art in information, data, and metadata management (White Paper). Retrieved from http://works.bepress.com/borgman/234

Steinhardt, S.B., & Jackson, S.J. (2014). Reconciling rhythms: Plans and temporal alignment in collaborative scientific work. In Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 134–145. ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2531736

Strasser, C., & Cruse, P. (2013). The DMPTool and DataUp: Helping researchers manage, archive, and share their data. University of California Curation Center, California Digital Library. Retrieved from https://rdmi.uchicago.edu/sites/rdmi.uchicago.edu/files/uploads/Strasser%2C%20C%20and%20Cruse%2C%20P_The%20DMPTool%20and%20DataUP-Helping%20Researchers%20Manage%2C%20Archive%2C%20and%20Share%20their%20Data.pdf

Sukow, A.E.R., & Grant, R. (2013). Forecasting and the role of churn in software-as-a-service business models. *iBusiness, 05*(01), 49. doi:10.4236/ib.2012.51A006

Sun, W., Zhang, K., Chen, S.-K., Zhang, X., & Liang, H. (2007). Software as a service: An integration perspective. In Proceedings of the 5th International Conference on Service-Oriented Computing, pp. 558–569. Berlin, Heidelberg: Springer-Verlag. doi:10.1007/978-3-540-74974-5_52

Uhlir, P.F. (2012). For attribution – developing data attribution and citation practices and standards: Summary of an international workshop. Washington, D.C.: The National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=13564

Velden, T., Bietz, M.J., Diamant, E.I., Herbsleb, J.D., Howison, J., Ribes, D., & Steinhardt, S.B. (2014). Sharing, re-use and circulation of resources in cooperative scientific work. In Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work; Social Computing, pp. 347–350. New York, NY, USA: ACM. doi:10.1145/2556420.2558853

Wallis, J.C. (2012). The distribution of data management responsibility within scientific research groups (Ph.D.). University of California, Los Angeles, United States. Retrieved from http://escholarship.org/uc/item/46d896fm

Wallis, J.C., & Borgman, C.L. (2011). Who is responsible for data? An exploratory study of data authorship, ownership, and responsibility. In Annual Meeting of the American Society for Information Science and Technology, 48, pp. 1–10. New Orleans, LA: Information Today. doi:10.1002/meet.2011.14504801188

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. doi:10.2218/ijdc.v3i1.46

Wallis, J.C., Borgman, C.L., Mayernik, M.S., Pepe, A., Ramanathan, N., & Hansen, M. A. (2007). Know thy sensor: Trust, data quality, and data integrity in scientific digital libraries. In Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries, LINCS 4675, pp. 380–391. Berlin: Springer. doi:10.1007/978-3-540-74851-9_32

Wallis, J.C., Milojevic, S., Borgman, C.L., & Sandoval, W.A. (2006). The special case of scientific data sharing with education. *Annual Meeting of the American Society for Information Science and Technology, 43*, pp. 1–13. Austin, TX: Information Today, Inc. doi:10.1002/meet.14504301169

Wallis, J.C., Rolando, E., & Borgman, C.L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE, 8*(7), e67332. doi:10.1371/journal.pone.0067332

Whitlock, M.C. (2011). Data archiving in ecology and evolution: Best practices. *Trends in Ecology and Evolution, 26*(2), 61–65. doi:10.1016/j.tree.2010.11.006