January 1, 2014

Ship Space to Database: Scientific and Social Motivations for a Database to Support Deep Subseafloor Biosphere Research

Peter Darch, *University of California - Los Angeles*
Christine L Borgman, *University of California - Los Angeles*

Ship Space to Database: Motivations to Manage Research Data for the Deep Subseafloor Biosphere

Peter T. Darch
UCLA Department of Information Studies
GSEIS Bldg, Rm. 210, Box 951520, Los Angeles
California 90095-1520
petertdarch@ucla.edu

Christine L. Borgman
UCLA Department of Information Studies
GSEIS Bldg, Rm. 235, Box 951520, Los Angeles
California 90095-1520
borgman@gseis.ucla.edu

ABSTRACT
What motivates data management strategies of scientific collaborations? In this paper, we argue that infrastructures for data management are designed to support scientific work *per se* and to perform a variety of social functions. We present findings from a longitudinal ethnographic case study of a large, multidisciplinary, distributed scientific project studying subseafloor microbial life. A critical element of this project's plan for data management is to construct an online portal that includes a data registry and data repository. We found that a range of factors motivate the construction of these systems and their features. In addition to scientific concerns for curation and accessibility of diverse and scarce data, we argue that the building of the registry and repository is also motivated by social factors. One factor is the attempt to build a community of domain researchers to endure beyond the end of this project in 2020. Another is the potential use of the registry and database as tools to demonstrate the productivity of the project in negotiations about the allocation of ocean drilling cruise resources. Considering the social and scientific factors together enriches accounts of how knowledge infrastructures are constructed.

Keywords
Big Data, data management, data curation, microbiology, scientific databases, knowledge infrastructures, little science, small science, multidisciplinary science

INTRODUCTION
The deluge of scientific data offers new opportunities to compare and integrate data across disciplines. It also poses an immense challenge, as large volumes of data are difficult to manage and exploit effectively. Data often are not captured or managed for reuse. Requirements of funding agencies and journals to release research data highlight the complexity of modern science: not only the contested notion of data, but competing views of research, innovation, and scholarship, disparate incentives for collecting and releasing data, the economics and intellectual property of research products, and public policy. The promise of technology-enabled, data-intensive digital scholarship in science is predicated upon available systems, services, tools, content, policies, practices, and human resources to discover, mine, and use research products. Collections of scientific data are essential components of *knowledge infrastructures*, defined as "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds" (P. N. Edwards, 2010, p. 17). Not only is this infrastructure not yet in place, it is not yet clear what should be built or how to build it (Bell, Hey, & Szalay, 2009; Borgman, Forthcoming, 2007; P. N. Edwards et al., 2013).

Knowledge infrastructures are expensive to construct and maintain. The value proposition and burden of costs are much debated (Berman & Cerf, 2013). Their design rests on the ability to explicate the socio-technical structures that are embodied in the data, in the practices, in technical arrangements, and in policies. These interdependencies are known to present significant risks to adoption and implementation (Altman, 2009; Larsen, 2008). Infrastructures for research data serve many more purposes than disseminating resources. They also must support data collection, analysis, use, and reuse. Ideally, they also support new scientific methods and improve access to information. Technologies may include metadata schemas, computational infrastructure, data repositories, and other tools and services.

We present findings about the creation of an online scientific data registry and a data repository from a longitudinal case study of infrastructure development in a large, collaborative, distributed, multidisciplinary scientific project studying subseafloor microbial life, namely the *Center for Dark Energy Biosphere Investigations* (*C-DEBI*). The registry and repository are intended to advance of the scientific work of C-DEBI, but they also serve important social roles in the broader institutional contexts in

77th ASIS&T Annual Meeting, October 31- November 5, 2014, Seattle, WA, USA.

which the project is embedded. Knowledge infrastructures have the potential to reconfigure social arrangements (P. N. Edwards et al., 2013). Most information science research on the construction of infrastructure for data management addresses motivations for supporting scientific work. Here we consider the impact of scientific and social factors on the development of such infrastructure.

MOTIVATIONS FOR BUILDING DIGITAL SCIENTIFIC DATA INFRASTRUCTURE

In this section, we motivate our research questions regarding what factors drive and influence the development of infrastructure for scientific data. We first examine how elements of infrastructure may be motivated by scientific needs. Then we consider what social functions the products of scholarly work can perform, raising the possibility that infrastructure for scientific data may also be constructed and used with a view to serving both social scientific functions.

How Data Management Infrastructure Supports Scientific Work

Most information science research on the development of scientific data management infrastructure concerns how to support the scientific work of the communities they are intended to serve (Palmer, Cragin, Heidorn, & Smith, 2007; Ribes & Finholt, 2009). Some of this research addresses supply-side issues, such as building infrastructures that ensure the successful management and curation of data. These studies tend to concern the support of long-term multidisciplinary collaborations through the building of databases. One set of challenges is the need to integrate multiple types of data produced by scientists from different communities of practice, with disparate knowledge practices and standards (Baker & Bowker, 2001; Bowker, 2000; Leonelli, 2012, 2013). A related challenge is the difficulty of getting scientists to contribute their data to the databases. Scientists' concerns include control of data, authorship rights, and incentives to manage and share their data (Borgman, Wallis, & Enyedy, 2007; Borgman, 2012; Wallis, Mayernik, Borgman, & Pepe, 2010).

Another body of research addresses the demand for digital data management infrastructure, for instance, recognizing the potential use of databases as tools in day-to-day scientific work (Hine, 2006). Some studies consider the contexts in which databases are used and the needs of the scientists in these contexts, such as accessing data, integrating data from multiple sources, and analyzing and preserving data (Borgman, Wallis, Mayernik, & Pepe, 2007). Multidisciplinary databases serve scientists working in disparate contexts with disparate requirements and standards (Wallis et al., 2010). Use of these data may occur long after data were collected and deposited, and long after the scientific collaboration has ended (Karasti, Baker, & Halkola, 2006; Karasti, Baker, & Millerand, 2010). Other studies have considered how potential users of a database might be encouraged to use data contained therein, for example by asking how to improve these users' trust in these data (Faniel & Jacobsen, 2010; Van House, Butler, & Schiff, 1998; Wallis et al., 2007).

How Databases and Other Scholarly Products Support Collaboration and Infrastructure

Relatively little research on scholarly communication has addressed work products such as datasets or databases. Data can support a variety of types of work at the community-of-practice level (Birnholtz & Bietz, 2003). One is helping to define boundaries between different communities of practice; the differing roles that data play in the work of these communities can emphasize their separateness. Another is that decisions regarding whether an individual scientist is granted access to particular types of data can be a proxy for whether this scientist is admitted as a member of a particular community of practice. Whether or not one produces one's own data is often a marker of status within a particular community (Zimmerman, 2008).

Databases also perform work at the communities-of-practice level. For example, they can be used to define and negotiate boundaries between scientific communities (Bietz & Lee, 2009). Furthermore, as "boundary negotiating objects" (Lee, 2007, p. 307), databases can be used to coordinate scientific work spanning multiple disciplines.

Other products of scholarly work can also serve social functions. For example, the production of journal articles and conference papers is part of the competition for scarce resources such as funding and the use of physical infrastructures. Scholarly credit accrues to the producers, which can then be transformed into funding for more scientific work (Latour & Woolgar, 1986). Products of scholarly work also are used to build communities of scholars. Scientific newsletters in biology may build community not only in terms of defining membership and providing opportunities for information exchange, but also for spreading common norms of practice (Kelty, 2012). A wide range of scholarly products has therefore been demonstrated to perform social functions, in addition to scientific or technical functions. This raises the possibility that infrastructure for scientific data management, too, might perform social functions.

CASE STUDY

A number of scientific factors drive the building of infrastructure for scientific data management. Further, scholarly work products may be motivated by their potential social functions. The discussion in the previous section motivates our three research questions for this paper:

1. What scientific, technical, and social roles does infrastructure for scientific data management play in multidisciplinary collaborations?
2. How do these collaborations negotiate and create the infrastructure they require?

3. What sociotechnical factors drive the development of this infrastructure?

We address these research questions via a case study of a large, multidisciplinary, distributed scientific collaborative project. The project and methods are introduced here and developed more fully in the findings.

The Center for Dark Energy Biosphere Investigations

The Center for Dark Energy Biosphere Investigations (C-DEBI) is an NSF *Science and Technology Center* (*STC*) that launched in September 2010. It was initially funded for five years, with the possibility of renewal for another five years. The project's fundamental aim is to build a community of researchers to study subseafloor microbial life and to understand interactions between this life and the physical environment inhabited. It is a highly interdisciplinary project, bringing microbiologists together with a wide variety of physical scientists, including geologists, mineralogists, sedimentologists, geochemists, and hydrologists. These researchers are geographically distributed, with the Principal Investigator (PI) and four co-PIs based at five US universities distributed coast-to-coast. C-DEBI funding covers projects conducted by over 80 scientists in more than 50 universities and research institutions across the USA, Europe, and Asia.

Scientists involved with C-DEBI work towards the project's scientific goals through the collection and analysis of physical samples, such as sediments and portions of the basaltic crust, or water. Fundamental to these scientists' work is the production, analysis, and correlation of data about the microbial communities in these samples and the physical properties, such as geochemical or hydrological, of these samples. The most significant source of data for C-DEBI researchers during the period of our case study were scientific ocean drilling cruises conducted by the *Integrated Ocean Drilling Program* (*IODP*) which ran from 2003-2013 (it should be noted that the IODP was replaced with a new drilling program in 2013, namely the *International Ocean Discovery Program*, also known as IODP. For the purposes of this paper, the acronym "IODP" will be used to refer to the Integrated Ocean Drilling Program throughout). The IODP generally referred to drilling cruises as *"Expeditions"*.

C-DEBI is a good case study for addressing research questions about the role of scientific databases in research collaborations because its highly interdisciplinary and distributed nature exemplifies the challenges involved in designing and implementing knowledge infrastructures for contemporary scientific research. The launch of C-DEBI has afforded opportunities to observe how the work of negotiating, building, and maintaining data management practices and infrastructures unfolds in a new collaborative setting.

Data Management in C-DEBI

The portion of our research addressed in this paper is the process of constructing infrastructure for scientific data management in C-DEBI. This infrastructure is part of an online portal currently being developed, and will be hosted by one of the participating US research universities. The portal will contain both a data registry and a data repository, along with information about C-DEBI publications, technological products developed during the course of C-DEBI-related activities that may be used in technology transfer, and materials for C-DEBI's extensive Education and Outreach activities. Given the scope of this paper, we focus below on the element of the portal that relates to scientific data, namely the registry and the repository.

C-DEBI was established prior to the National Science Foundation's requirements for Data Management Plans, which began with proposals submitted in 2011 (National Science Foundation, 2010). However, C-DEBI has been required to develop a plan for its renewal application in 2014 (Center for Dark Energy Biopshere Investigations, 2012). The registry and repository are critical elements for the fulfilment of this plan. C-DEBI will mandate that all datasets produced during the course of C-DEBI-funded scientific work that support results reported in scientific publications will be registered on the C-DEBI data registry, and will be deposited in a relevant, online, publicly-accessible database. The C-DEBI data registry and repository are intended to organize C-DEBI data and to make them publicly available. The data registry will report multiple metadata elements for each dataset (to be explained below), along with a link to where the dataset is hosted. The C-DEBI repository is being designed as a repository for datasets produced during C-DEBI-funded work for which no other relevant repository currently exists.

Work began on the portal in summer 2013. C-DEBI allocated substantial funding to portal development: $95,000 in 2013 and an additional $287,000 for 2014 (Center for Dark Energy Biosphere Investigations, 2014). The portal is being developed in stages. The first phase, which includes prototyping and site architecture, was completed in time for the NSF Site Visit in January 2014. Initial design concerned decisions about what datasets to be stored or linked, and the categories and granularity of metadata necessary for each dataset.

METHODS

We present selected findings from an eighteen-month ethnographic case study of C-DEBI that includes participant observation, semi-structured interviews, and document analysis. Scientists have been observed at work in a range of contexts. We have been embedded for eight months in a laboratory headed by a leading figure in C-DEBI at a large US research university, and conducted week-long observational work in two other participating laboratories. We have attended scientific meetings of both

		Career stage	Interviewees	Involved with IODP
C-DEBI	USA-based	Undergraduate	**5**	0
		Graduate student	**9**	0
		Postgraduate	**7**	1
		Faculty	**13**	2
		Non-scientists	**4**	0
	Non-USA-based	Faculty	**3**	3
		TOTAL C-DEBI	**41**	6
IODP	Cruise operations	Curator	**2**	
		Staff Scientist	**2**	
		Technical support	**1**	
	Ocean Leadership	Policy	**2**	
		Data management	**1**	
		TOTAL IODP	**8**	

Table 1: The composition of our interview sample.

C-DEBI and the broader scientific communities in which the Center is embedded (conferences, workshops, seminars, colloquia) and participated in domain-specific scientific conferences (Darch & Cummings, 2013). One member of our project team has observed construction of the C-DEBI portal since work commenced during summer 2013. This has involved meeting with the member of C-DEBI tasked with developing the portal and attending teleconference meetings with the three C-DEBI project members involved in building the portal.

Our current interview sample comprises 49 people, including C-DEBI-affiliated scientists and scientists, curators, and managerial staff involved in related activities such as the scientific ocean research cruises. Interviewees were selected based on the relevance of their experiences and roles within C-DEBI and the ocean drilling program to our research questions. Our sample is detailed in Table 1, which distinguishes between respondents involved in C-DEBI and those working for the Integrated Ocean Drilling Program (IODP). The C-DEBI sample is broken down further by geographic location (USA or not), and career stage. The column "Involved with IODP" indicates which interviewees are involved in policy- or decision-making in the IODP. The IODP interviewees are further split into two groups: those in cruise operations, and those with the Consortium for Ocean Leadership, which was responsible for administering US involvement in the IODP.

Interviews ranged in length from 35 minutes to two hours and 30 minutes, with the majority being between one and two hours long. Scientists were interviewed about their data practices, scientific work, and professional backgrounds, and involvement with the IODP. Non-scientists were asked about their work within the C-DEBI project or the IODP and how they are involved with the implementation and maintenance of infrastructure and policies within C-DEBI

We have also assembled a corpus of documents including official C-DEBI documents, contextual materials, and information from the NSF and the IODP.

Data were analyzed using a grounded theory approach (Glaser & Strauss, 1967). Interview transcripts and other documents were read closely, and a number of themes emerged. These documents were then coded according to these themes using NVivo.

FINDINGS

Our findings about the C-DEBI data registry and data repository address the scientific and social roles of infrastructure for data management in this collaboration. The primary sources of data for C-DEBI are scientific ocean drilling cruises. Samples and observations from these cruises become many kinds of biological and physical data for C-DEBI scientists, represented in many ways, to address many different scientific questions, across disciplines, locations, and over time. These data sources, in turn, interact with the science requirements of C-DEBI's primary funding source, the National Science Foundation. However, considering the diversity of data management infrastructure built by comparable STCs in response to NSF requirements

shows that the form taken by such infrastructure is not completely determined by NSF requirements. Other social and scientific factors also influence design. In this section, we present these other factors. Then, we draw them together to explain how they have motivated the construction of the C-DEBI data registry and data repository. First, however, we set the scene by outlining the scientific data that require management within C-DEBI.

Scientific Data in C-DEBI

C-DEBI-funded scientists produce a wide range of datasets about a variety of physical and biological phenomena. In the domain of microbiology, scientists used physical rock and water samples taken from expeditions to generate a range of data characterizing the communities of microbes found in these samples, including genomic data used to perform analyses of community composition and proteomic data to understand microbial function.

In the domains of physical sciences, C-DEBI-affiliated scientists produce multiple types of data based on analysis of samples from expeditions. Such datasets result from a variety of analyses, such as the mineralogical or geochemical composition. These analyses may be conducted in the scientists' own laboratories or at large infrastructural facilities where scientists must apply for time to use the equipment.

Scientists combine data about microbiological and physical phenomena to produce hybrid datasets that span both biological and physical science domains. In some cases, this involves integrating laboratory-produced biological data with laboratory-produced physical data. In other cases, this involves combining laboratory-produced biological data and physical data available in the IODP databases. A common example involves the correlation of the existence and abundance of different microbial species in a single location with the physical properties of the environment in that location.

We also found a significant diversity in methods used to produce datasets across C-DEBI-affiliated scientists. Individual scientists employed different methods and techniques to produce datasets that were similar in form and intent. Even within a single laboratory, we found significant variation from one laboratory bench to the next, and for individual stages involved in the production of a single dataset. This heterogeneity of methods was driven to a large extent by the highly interdisciplinary and multi-institutional nature of C-DEBI, with scientists bringing multiple techniques, perspectives and expertise from their respective disciplinary backgrounds and from the array of laboratories in which they had trained and worked.

Collaboration and Infrastructure Drivers for Development of Digital Data Management Infrastructure

Characteristics of the science, of the state of collaboration, and the state of infrastructure for the scientific fields participating in C-DEBI all appear to be drivers for the particular design of the registry and repository.

National Science Foundation Requirements

C-DEBI was launched in late 2010 with five years of funding and the possibility of renewal for an additional five years. Although the C-DEBI proposal was submitted prior to the implementation of data management plan requirements for all National Science Foundation (NSF) grant proposals, Science and Technology Centers (STCs) such as C-DEBI, subsequently were required to develop and implement a management plan for data resulting from their scientific work. This plan must include requirements for long-term management and curation of data, and strategies for making them publicly available. Ahead of the NSF renewal review, scheduled to take place in summer 2014, C-DEBI developed a Data Management Plan (Center for Dark Energy Biopshere Investigations, 2012). Scientists who receive C-DEBI funding must make data they produce publicly available. The data registry and data repository introduced above are being developed and implemented so that scientists can comply with this requirement.

Curation and Accessibility of Scarce Data

The scientific work products of C-DEBI are particularly valuable because data about the deep subsurface biosphere is scarce. Serious data collection began only in the late 1990s – largely since the launch of the IODP in 2003 – in part because of the expense and logistics of participating in expeditions. Currently, data exist only for a few sites in the ocean and, compared to data in other domains of microbiology, is about relatively basic phenomena. C-DEBI scientists frequently refer to their own work as "discovery-driven" rather than "hypothesis-driven" to emphasize the lack of maturity of deep subseafloor biosphere research relative to studies of microbes in other environments. Furthermore, a dataset about a particular phenomenon in a particular location is highly likely to be the only dataset in existence on this topic, and thus the consequences of data loss are significant.

Another factor motivating the C-DEBI data registry and data repository is unevenness of data availability across the disciplines involved in C-DEBI. Policies for creating and contributing to community databases vary across the participating disciplines. Scientists who publish in most microbiological journals are required to deposit genomic sequence data supporting their conclusions to publicly accessible databases such as *GenBank* (http://www.ncbi.nlm.nih.gov/genbank). No similar requirement exists for physical science data in the realm of C-DEBI. Some appropriate databases do exist (such as *Pangaea* in the field of the earth sciences, http://www.pangaea.de) but contributions of data are at the discretion of the scientist.

Despite the scientific aims of C-DEBI to integrate physical and biological data, the sparse availability of data across

participating scientific disciplines makes this goal difficult to accomplish. The problem is particularly acute when findings involve both biological and physical data. For example, a journal article may present an analysis correlating biological sequence data about a microbial community with data about the physical environment this community inhabits. Sequence data are contributed to a community database where they are accessible. The physical science data, in contrast, may be kept solely in the possession of the scientist who produced them and may get lost over time. As a result, other scientists who may wish to reproduce or extend this analysis with the combination of data are unable to do so.

Building and Sustaining Community

One of the major aims of C-DEBI is to build an enduring community of researchers studying subseafloor microbial life. C-DEBI provides funding support for these scientists to enroll them into the field of deep subseafloor biosphere research. However, it is NSF policy regarding Science and Technology Centers (STCs) such as C-DEBI that there be a 10-year limit on funding, after which the STC must either find other sources of funding to continue or disband. The C-DEBI data registry and data repository are intended to help build a community of researchers and to foster an open, collaborative spirit that will sustain this community beyond the ten-year NSF funding period.

Ocean Drilling for Scientific Research

Scientists studying subseafloor microbial life use samples of rock and water collected from the seafloor to characterize subseafloor microbial communities and the environments they inhabit. These samples are collected on ocean research cruises that can be organized by a number of agencies. The most significant for C-DEBI researchers during the period of our case study were scientific ocean drilling cruises conducted by the Integrated Ocean Drilling Program (IODP) that ran from 2003-2013.

The IODP was an international organization that brought together scientists from 23 countries and covered a wide range of scientific disciplines. Along with microbiology, the IODP also comprised 11 physical science disciplines (including geochemistry, hydrology, paleomagnetology, and sedimentology). The products of each IODP expedition included cores, which are lengths of rock taken from the holes. These cores were processed on the ship by the IODP curators and by expedition participants. Some cores were first subjected to a wide range of analyses of physical properties on board the ships: these analyses were standard across all IODP expeditions and results are stored in the IODP database. Data were subject to a one-year moratorium for expedition participants, after which they were made publicly accessible. Other cores were allocated amongst expedition participants, who either analyzed the cores on board the ship or transported them to their onshore laboratories after the cruise. Other cores were (and continue to be) stored at the IODP core repositories. These samples also were subject to the one-year moratorium for expedition participants and available thereafter to all scientists.

The forerunner organizations to the IODP, namely the *Deep Sea Drilling Program* (*DSDP*) which ran from 1968-1983 and the *Ocean Drilling Program* (*ODP*), focused almost exclusively on the physical sciences. The DSDP did not include microbiology at all. Microbiologists became involved with the ODP from the late 1990s, though this involvement was generally ad hoc and sporadic: microbiologists were only systematically included in scientific ocean drilling activities from the launch of the IODP. The 1990s saw the emergence of subseafloor life as a domain of serious scientific study, leading to the expansion of the IODP's scientific agenda to include microbiology. This domain of study is now a major component of the successor to the IODP, namely the International Ocean Discovery Program that launched in late 2013.

Decision-making processes within the IODP about the shape of the Program's major research areas usually involved discussion and negotiation between representatives of the multiple disciplines involved with the IODP. Sometimes, representatives of microbiology have encountered resistance from other disciplines during negotiations about the organization of the IODP's scientific work. The IODP's scarce resources (ship space, physical samples) needed to be distributed among participating disciplines. The forerunners to the IODP did not involve microbiology. Thus, the inclusion of microbiology in the IODP meant other disciplines had to give up some of their share of the IODP resources, in a number of different ways.

One way was to focus selected IODP expeditions on microbiology by including more microbiologists – at the expense of scientists from other disciplines – and devoting more time to performing microbiological experiments and analyses in the boreholes and on board the ship. A second way was to allocate more of the very limited number of places available to scientists on other IODP expeditions to microbiologists. Third was to retrofit one ship, the IODP *Joides Resolution* (or, *JR*), which had been used in the forerunner to the IODP, by converting physical science laboratory to a dedicated microbiological laboratory. The inclusion of microbiology in the IODP has also reduced the number of cores available to physical scientists. Cores intended for microbiological analysis must be handled differently to those for physical science analysis. For instance, samples for physical analysis are typically stored at -4 ^{o}C, while samples for microbiological analysis are typically stored at -80 ^{o}C.

Given that physical science disciplines have been required to give up valuable infrastructural resources to enable the inclusion of microbiologists in IODP expeditions, microbiologists need to demonstrate that this inclusion has resulted in scientifically valuable output. Although academic journal articles can be used to demonstrate this

productivity, articles often involve the integration of data from multiple sources, including cruises conducted under the auspices of organizations other than the IODP: thus, the precise contribution of the IODP data to these articles is often unclear. The challenge facing the C-DEBI community, therefore, is to make the relationship between its scientific output and IODP cruises more explicit.

Building a Registry and Database for C-DEBI Data

Multiple challenges facing C-DEBI have been identified above, namely meeting NSF requirements for data management, addressing data scarcity and unevenness, building and sustaining a nascent community, and securing participation of microbiologists on ocean drilling cruises. These challenges are motivating both the construction of the C-DEBI data registry and data repository and the inclusion of particular features in this infrastructure.

NSF Requirements for Data Management

One of the motivations for building the C-DEBI registry and repository is to fulfill NSF requirements by providing infrastructure to help ensure all data produced under C-DEBI funding is publicly accessible. However, the existence of the registry and repository, and the forms they take, are not completely determined by these requirements.

As a point of comparison, we examined how current NSF Science and Technology Centers in related areas of science address data management requirements. C-DEBI is one of 11 current STCs; these were launched between 2005 and 2010. All ten of the other STCs are subject to NSF requirements that their affiliated researchers make publicly available the data collected during their STC-funded scientific work. However, only three of these ten STCs operate an online, publicly-accessible registry or repository that makes these data accessible by downloading datasets, by providing links to other sources, or by contact information of scientists who have produced datasets. These STCs are the *Center for Coastal Margin Observation and Prediction* (*CMOP*, www.stccmop.org); the *Center for Microbial Oceanography Research and Education* (*C-MORE*, cmore.soest.hawaii.edu); and the *Center for Remote Sensing of Ice Sheets* (*CReSIS*, www.cresis.ku.edu).

The four STC data registries or repositories (CMOP, C-MORE, CReSIS, C-DEBI) vary in types of datasets they contain and in the scope of metadata about each dataset. C-MORE is the most comparable to C-DEBI, in that they study microbial life in the ocean (a neighboring domain to that studied by C-DEBI) and scientists use samples collected from ocean research cruises (in the case of C-MORE, these cruises are operated directly by the STC).

C-MORE operates an online registry (cmore.soest.hawaii.edu/datasearch/data.php). For each dataset, the registry lists the laboratory where the dataset is located, the name of the particular cruise from which the dataset was generated, a few words briefly describing the dataset, and email contact details for the dataset's owner. In some cases, there is a link to download the dataset.

Compared to the planned registry for C-DEBI, the C-MORE registry and repository contains many fewer categories of metadata. For example, both the C-DEBI and C-MORE registry have a category to name the research cruise from which each dataset was derived. However, C-DEBI has a separate category for the precise geographic location (i.e., the specific drilling hole) from where the sample was drawn, whereas C-MORE does not even though each C-MORE cruise typically involves taking samples from multiple geographic locations. Another example is that the C-DEBI registry is planned to have metadata categories detailing the publication(s) in which a particular dataset has been used; by contrast, the C-MORE registry does not.

The NSF requirements consist of general principles that can be implemented in many ways. The current STCs each have taken different approaches, despite their scientific and organizational similarities. Thus it is important to consider the impact of other factors that have influenced decisions to implement the C-DEBI data management infrastructure and with what form and functions.

Infrastructure to Build and Sustain Community

The C-DEBI data registry and data repository are intended to help bring together and build a community of researchers, to sustain this community beyond the ten-year NSF funding period, and to foster norms of openness.

Since C-DEBI's original conception, the C-DEBI leadership have recognized the value of cyberinfrastructure in enabling and fostering links between a very distributed group of scientists (K. Edwards, 2009). In particular, infrastructure for data sharing, of which the C-DEBI data registry and data repository will be the fundamental component, is being developed to help "support the connection among scientists and others in the C-DEBI project research community" (Center for Dark Energy Biopshere Investigations, 2013, p. 9).

Additionally, our interviews have revealed that the registry and repository are regarded as an important element of C-DEBI's legacy beyond the project's anticipated end in 2020, in order to continue establishing and renewing links and collaboration across disciplines and institutions. The data management infrastructure is intended not only to bring a community together, but also to sustain and to continue to build this community in the longer term.

Finally, the data management infrastructure is intended not only to bring a community together, but also to foster particular norms of behavior amongst this community, namely those of openness and data sharing. Open access to data produced during C-DEBI-funded research is at the core of the philosophy underpinning C-DEBI's Data Management Plan (Center for Dark Energy Biopshere Investigations, 2012). The provision of the registry and

repository has the aim of enabling and encouraging behavior amongst C-DEBI-funded scientists that is consistent with norms of openness and data sharing.

Infrastructure to Support Microbiology in Scientific Ocean-Drilling Cruises

Categories of metadata about datasets will include information about the origin of the physical samples from which the dataset has been derived. In the case of samples collected from IODP expeditions, metadata includes the name of the research site, the expedition number, and the specific drill hole(s) where the samples originated.

By tying each dataset to the specific IODP expedition that the sample that produced the dataset was collected, the registry is intended to make more explicit the link between expeditions and the scientific work in the study of the deep subsurface biosphere these expeditions have enabled. In turn, the ability of microbiologists to advocate for continued and deepened participation in the IODP should be strengthened. Thus, tying datasets to specific IODP expeditions has motivated not only the development of the C-DEBI data registry and data repository themselves, but also the inclusion of particular categories of metadata that tie datasets the particular IODP expedition on which the sample was collected, but the specific borehole as well.

DISCUSSION

Existing studies of building of knowledge infrastructures to support data management and collaborative scientific work tend to focus only on how these infrastructures address scientific challenges (Palmer et al., 2007; Ribes & Finholt, 2009). Our case study reveals that, instead, both the very construction of the C-DEBI registry and repository and the inclusion of particular features are motivated by both social and scientific factors. By highlighting differences between how C-DEBI and other comparable STCs have approached the fulfilment of NSF data management requirements, we can see how the factors that are particular to C-DEBI have influenced the development of the registry and repository.

Scientific Motivations for Data Management Infrastructure

Some motivations for the data registry and repository relate to facilitating more and better scientific work. C-DEBI faces particular challenges in this respect. One is that the data are scarce: thus, long-term preservation is critical. Scarcity is a particularly critical issue for observational data (as opposed to experimental or simulation data) because they cannot be replicated and are often expensive to collect (National Science Board (U.S.), 2005).

Another challenge is that the data are highly diverse in a number of ways, including the phenomena that the data represent, the methods involved in producing these datasets (even for data that may be of a similar type), and the standards (e.g. metadata, requirements for depositing in disciplinary databases) to which data are managed and curated (Baker & Bowker, 2001; Bowker, 2000; Leonelli, 2012, 2013). Some of the factors motivating the registry and repository relate to meeting these challenges so that the database can enhance users' scientific work (Hine, 2006), ensuring a more even provision of, and access to, data across different disciplines to enable scientists to answer interdisciplinary research questions at the heart of C-DEBI. These scientists frequently need to use data produced in different contexts and to integrate data from multiple contexts (Borgman, Wallis, Mayernik, et al., 2007; Wallis et al., 2010).

Social Motivations for Data Management Infrastructure

Both the leadership of, and the scientists funded by, C-DEBI can be understood as social actors embedded in multiple and intersecting social contexts. These contexts include ocean drilling cruises and scientific disciplines such as microbiology. We found that C-DEBI personnel are well aware of the challenges and opportunities of these contexts. The C-DEBI data registry and data repository are intended to be mobilized both as tools in promoting deep subseafloor biosphere scientists' interests in these contexts.

Within the context of microbiology, the deep subseafloor biosphere is a recently established domain of study, lacking much of the infrastructure and community that characterizes more mature domains. Some knowledge products, such as scientific newsletters, have been shown to help build communities and establish scientific norms of practice (Kelty, 2012). Our study extends this scope to infrastructures for scientific data management.

The C-DEBI data registry and data repository are also intended to help sustain and continue to build a community of deep subseafloor biosphere researchers in the longer term. Temporality – in the sense of building infrastructures that endure for the long-term – has been found to be a scientific motivation to build infrastructure for data management (Karasti et al., 2006, 2010); we have found that temporal concerns also relate to the social work that the registry and repository are being designed to perform.

In the context of scientific ocean drilling cruises, we can see another form of social work that the C-DEBI data registry and data repository is intended to perform, namely the negotiation, advancement and protection of the position of microbiology within this context. The data registry and data repository are intended to make the relationship between IODP cruises and the scientific work of C-DEBI more explicit, motivating both the registry's and repository's existences and the inclusion of some categories of metadata. C-DEBI leaders hope that the registry and repository will provide a resource to be mobilized in the negotiations for allocation of scarce resources (ship space, physical samples) within the International Ocean Discovery Program, just as other scholarly products are used by scientists to accrue resources (Latour & Woolgar, 1986).

CONCLUSIONS

We have seen how both social and scientific factors come together to motivate the very existence, and specific features, of a data registry and a data repository supporting a large, distributed, multidisciplinary scientific collaboration. Considering both social and scientific factors enriches accounts of how infrastructures supporting scientific work are built and maintained. The decision-making processes of key stakeholders in the construction of such infrastructures are better understood by considering the social contexts in which they and the scientific community served by the infrastructure are embedded.

The C-DEBI data registry and data repository are still undergoing development. By continuing to study their construction, we can see how different motivations emerge over time or change in importance to C-DEBI. Furthermore, because the registry and repository are only two components among many of the C-DEBI infrastructure, we will be able to study how they interact with other infrastructural components in meeting C-DEBI's core aims. This will help us to characterize how, from the point of view of C-DEBI, its registry and repository are intended to be more than just tools to facilitate scientific work.

ACKNOWLEDGEMENTS

The work in this paper has been supported by the Sloan Foundation Award #20113194, *The Transformation of Knowledge, Culture and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective*. We also acknowledge the contributions of Milena Golshan, Ashley Sands, Sharon Traweek and Jillian Wallis for commenting on earlier drafts of this paper, and to Rebekah Cummings for assistance with conducting the case study. We are deeply grateful to those C-DEBI and IODP personnel who we interviewed and observed at work.

REFERENCES

Altman, M. (2009). Transformative Effects of NDIIPP, the Case of the Henry A. Murray Archive. *Library Trends*, *57*(3), 338–351. doi:10.1353/lib.0.0040

Baker, K. S., & Bowker, G. C. (2001). Designing an Infrastructure for Heterogeneity of Ecosystem Data, Collaborators and Organizations. *LTER Network Newsletter*.

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the Data Deluge (Computer Science). *Science*, *323*(5919), 1297–1298. doi:10.1126/science.1170411

Berman, F., & Cerf, V. G. (2013). Who Will Pay for Public Access to Research Data? *Science*, *341*(6146), 616–617. doi:10.1126/science.1241625

Bietz, M. J., & Lee, C. (2009). Collaboration in Metagenomics: Sequence Databases and the Organization of Scientific Work. In *Proc. Europ. Conf. on Computer-Supported Cooperative Work* (pp. 243–262). doi:10.1007/978-1-84882-854-4_15

Birnholtz, J. P., & Bietz, M. J. (2003). Data at Work: Supporting Sharing in Science and Engineering. In *Proceedings of the 2003 International {ACM} {SIGGROUP} Conference on Supporting Group Work* (pp. 339–348). New York, {NY}, {USA}: Association for Computing Machinery. doi:10.1145/958160.958215

Borgman, C. L. (Forthcoming). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge MA: MIT Press.

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

Borgman, C. L. (2012). The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. doi:10.1002/asi.22634

Borgman, C. L., Wallis, J. C., & Enyedy, N. D. (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., & Pepe, A. (2007). Drowning in Data: Digital library architecture to support scientific use of embedded sensor networks. In *Joint Conference on Digital Libraries* (pp. 269–277). Vancouver, British Columbia, Canada: Association for Computing Machinery. doi:10.1145/1255175.1255228

Bowker, G. C. (2000). Biodiversity Datadiversity. *Social Studies of Science*, *30*(5), 643–683. doi:10.1177/030631200030005001

Center for Dark Energy Biopshere Investigations. (2012). *C-DEBI Data Management Philosophy and Policy*. Retrieved from http://www.darkenergybiosphere.org/internal/docs/C-DEBIDataManagementPlan_2012draft.pdf

Center for Dark Energy Biopshere Investigations. (2013). *C-DEBI Strategic Implementation Plan 2013-2014*. Retrieved from http://www.darkenergybiosphere.org/internal/docs/C-DEBI_SIP_2013-2014.pdf

Center for Dark Energy Biosphere Investigations. (2014). *Center for Dark Energy Biosphere Investigations STC Annual Report 2013*. Retrieved from http://www.darkenergybiosphere.org/internal/docs/C-DEBI-Annual-Report-2013.pdf

Darch, P. T., & Cummings, R. L. (2013, December 9). *Buried deep: How data about subseafloor life becomes dark and why*. Presented at the American Geophysical Union 46th Annual Fall Meeting, San Francisco, CA.

Edwards, K. (2009). *Center for Dark Energy Biosphere Investigations (C-DEBI): A center for resolving the extent, function, dynamics and implications of the Subseafloor*

Biosphere. Retrieved from http://www.darkenergybiosphere.org/internal/docs/2009C-DEBI_FullProposal.pdf

Edwards, P. N. (2010). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., … Calvert, S. (2013). *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges* (p. 40). Ann Arbor, MI: University of Michigan. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/97552

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. *Journal of Computer Supported Cooperative Work*, *19*(3-4), 355–375. doi:10.1007/s10606-010-9117-8

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory* (Vol. 5). Retrieved from http://www.ualberta.ca/iiqm/backissues/5_1/pdf/mills.pdf

Hine, C. (2006). Databases as Scientific Instruments and Their Role in the Ordering of Scientific Work. *Social Studies of Science*, *36*(2), 269–298. doi:10.1177/0306312706054047

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. doi:10.1007/s10606-006-9023-2

Karasti, H., Baker, K. S., & Millerand, F. (2010). Infrastructure Time: Long-term Matters in Collaborative Development. *Computer Supported Cooperative Work (CSCW)*, *19*(3-4), 377–415. doi:10.1007/s10606-010-9113-z

Kelty, C. M. (2012). This is not an article: Model organism newsletters and the question of "open science." *BioSocieties*, *7*(2), 140–168. doi:10.1057/biosoc.2012.8

Larsen, R. L. (2008). On the Threshold of Cyberscholarship. *Journal of Electronic Publishing*, *11*(1).

Latour, B., & Woolgar, S. (1986). *Laboratory life: the construction of scientific facts*. Princeton, {N.J.}: Princeton University Press.

Lee, C. P. (2007). Boundary Negotiating Artifacts: Unbinding the Routine of Boundary Objects and Embracing Chaos in Collaborative Work. *Computer Supported Cooperative Work (CSCW)*, *16*(3), 307–339. doi:10.1007/s10606-007-9044-5

Leonelli, S. (2012). When humans are the exception: Cross-species databases at the interface of biological and clinical research. *Social Studies of Science*, 0306312711436265.

Leonelli, S. (2013). Global data for local science: Assessing the scale of data infrastructures in biological and biomedical research. *BioSocieties*, *8*(4), 449–465.

National Science Board (U.S.). (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. Arlington, Virginia: National Science Foundation. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/

National Science Foundation. (2010). *NSF Data Management Plans*. Washington, D.C. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp

Palmer, C. L., Cragin, M. H., Heidorn, P. B., & Smith, L. C. (2007). Data curation for the long tail of science: The case of environmental studies. Presented at the 3rd International Digital Curation Conference, Washington, DC. Retrieved from https://apps.lis.uiuc.edu/wiki/download/attachments/32666/Palmer_DCC2007.rtf?version=1

Ribes, D., & Finholt, T. (2009). The Long Now of Technology Infrastructure: Articulating Tensions in Development. *Journal of the Association for Information Systems*, *10*(5). Retrieved from http://aisel.aisnet.org/jais/vol10/iss5/5

Van House, N. A., Butler, M. H., & Schiff, L. R. (1998). Cooperative work and practices of trust: Sharing environmental planning data sets. In *CSCW '98: ACM conference on computer supported cooperative work, Nov14-18* (pp. 335–343). Seattle, WA: New York: ACM.

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* (Vol. LINCS 4675, pp. 380–391). Budapest, Hungary: Berlin: Springer. doi:10.1007/978-3-540-74851-9_32

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. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries* (pp. 333–340). Gold Coast, Queensland, Australia: ACM. doi:10.1145/1816123.1816173

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*(5), 631–652. doi:10.1177/0162243907306704