PeerJ Computer Science

Ship space to database: emerging infrastructures for studies of the deep subseafloor biosphere

Peter T. Darch[1] and Christine L. Borgman[2]

[1] School of Information Sciences, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, United States

[2] Information Studies, University of California, Los Angeles, CA, United States

ABSTRACT

Background. An increasing array of scientific fields face a "data deluge." However, in many fields data are scarce, with implications for their epistemic status and ability to command funding. Consequently, they often attempt to develop infrastructure for data production, management, curation, and circulation. A component of a knowledge infrastructure may serve one or more scientific domains. Further, a single domain may rely upon multiple infrastructures simultaneously. Studying how domains negotiate building and accessing scarce infrastructural resources that they share with other domains will shed light on how knowledge infrastructures shape science.

Methods. We conducted an eighteen-month, qualitative study of scientists studying the deep subseafloor biosphere, focusing on the Center for Dark Energy Biosphere Investigations (C-DEBI) and the Integrated Ocean Drilling Program (IODP) and its successor, the International Ocean Discovery Program (IODP2). Our methods comprised ethnographic observation, including eight months embedded in a laboratory, interviews ($n = 49$), and document analysis.

Results. Deep subseafloor biosphere research is an emergent domain. We identified two reasons for the domain's concern with data scarcity: limited ability to pursue their research objectives, and the epistemic status of their research. Domain researchers adopted complementary strategies to acquire more data. One was to establish C-DEBI as an infrastructure solely for their domain. The second was to use C-DEBI as a means to gain greater access to, and reconfigure, IODP/IODP2 to their advantage. IODP/IODP2 functions as infrastructure for multiple scientific domains, which creates competition for resources. C-DEBI is building its own data management infrastructure, both to acquire more data from IODP and to make better use of data, once acquired.

Discussion. Two themes emerge. One is data scarcity, which can be understood only in relation to a domain's objectives. To justify support for public funding, domains must demonstrate their utility to questions of societal concern or existential questions about humanity. The deep subseafloor biosphere domain aspires to address these questions in a more statistically intensive manner than is afforded by the data to which it currently has access. The second theme is the politics of knowledge infrastructures. A single scientific domain may build infrastructure for itself and negotiate access to multi-domain infrastructure simultaneously. C-DEBI infrastructure was designed both as a response to scarce IODP/IODP2 resources, and to configure the data allocation processes of IODP/IODP2 in their favor.

Submitted 4 March 2016
Accepted 10 October 2016
Published 14 November 2016

Corresponding author
Peter T. Darch, ptdarch@illinois.edu

Academic editor
Sally Jo Cunningham

Additional Information and Declarations can be found on page 25

DOI **10.7717/peerj-cs.97**

Subjects Human-Computer Interaction, Digital Libraries
Keywords Knowledge infrastructures, Scientific data, Microbiology, Long tail, Big science, Little science, Big data, Little data, Cyberinfrastructure, Data infrastructure

INTRODUCTION

The array of scientific fields facing a "data deluge" has grown rapidly in the years since the term was coined (*Hey & Trefethen, 2003*, p. 809). New instruments are capable of collecting data at greater volume, variety, and velocity than ever before. Consequences of these developments include the emergence of new infrastructures, changes in epistemologies, and new forms of collaborative work (*Kitchin, 2014*; *Leonelli, 2014*).

However, in many fields data are scarce and hard won (*Borgman, 2015*; *Kitchin & Lauriault, 2014*); their problem is the opposite of "big data." Domains suffering data scarcity may have lower epistemic status than those enjoying "data wealth" (*Sawyer, 2008*, p. 355). Consequently, they may attempt to increase their resources by developing better infrastructure to produce, manage, curate, and circulate the data they do have.

We examine how a community of researchers studying the deep subseafloor biosphere experiences data scarcity, and how they develop knowledge infrastructures to address this scarcity. This scientific domain experiences acute data scarcity because it aspires to address questions of societal concern and existential questions about humanity in a more statistically intensive manner than is presently possible.

Two infrastructures have been critical to this community's development (*Darch & Sands, 2015*). One was established long before the emergence of the deep subseafloor biosphere as a topic of scientific study and is shared with other domains. This is the *Integrated Ocean Drilling Program* (*IODP*, 2003–2013) and its successor, the *International Ocean Discovery Program* (*IODP2*, from 2013), an international organization that conducts scientific ocean drilling cruises on behalf of scientists studying physical and biological phenomena related to the seafloor (*International Ocean Discovery Program, 2014*). The second infrastructure is specific to the deep subseafloor biosphere, namely the *Center for Dark Energy Biosphere Investigations*, or *C-DEBI* (*Edwards, 2009*).

We explore the relationships between IODP/IODP2 infrastructure and C-DEBI infrastructure. Multiple domains compete for the scarce infrastructural resources of IODP/IODP2, such as space on drilling cruises for people and equipment, cores that are acquired on those cruises, and the work of data analysts and curators employed by IODP/IODP2. The features of C-DEBI, including its structure, modes of providing financial support for researchers, and its data infrastructure, are designed to enable deep subseafloor biosphere researchers to do the following:

1. exploit existing IODP/IODP2 resources allocated to deep subseafloor biosphere research; and
2. make more effective interventions in the inter-domain politics of IODP/IODP2 infrastructure, thereby securing a greater share of IODP/IODP2 resources for their community.

Many scientific fields rely on both single- and multi-domain infrastructure, thus our findings apply to infrastructure far beyond the subseafloor biosphere domain.

BACKGROUND AND RESEARCH QUESTIONS

In this section we introduce our research questions with a review of relevant literature. First we frame the concept of knowledge infrastructures, exploring how they develop, support, and constrain scientific practice. We consider the interaction between infrastructure specific to a single domain and infrastructure shared between domains. Second, we discuss how and why the availability of data is a critical concern in science, with attention to how domains address data scarcity. Third, we present more background on deep subseafloor biosphere research to explain why this scientific domain is an ideal site to study the interaction of knowledge infrastructures.

Knowledge infrastructures

The term *infrastructure* is often used in reference to large-scale systems with multiple social and technical components that provide services, resources, and facilities (*Edwards et al., 2009*). Infrastructures, in the senses used herein, are "best understood as ecologies or complex adaptive systems" (*Borgman, 2015*, p. 33). Infrastructures are complex in the sense that they comprise multiple systems that may have been devised, built, and configured by different actors with varying objectives, but which function together. They are continuously adaptive in the sense that components change or are introduced (*Edwards et al., 2013*). A particular type of infrastructure is "knowledge infrastructures," which *Edwards* (*2010*, p. 17) defined as "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds."

Infrastructure is a "fundamentally relational concept" in the sense that a sociotechnical configuration "becomes infrastructure in relation to organized practices" (*Star & Ruhleder, 1996*, p. 113). From the point of view of a particular domain, what is considered to be a knowledge infrastructure is precisely those configurations of social and technical components that provide resources to support a community's shared objectives and practices. Indeed, as a domain's interests and practices change, the knowledge infrastructure must adapt if the domain is to remain relevant (*Ribes & Polk, 2015*).

Single vs. multi-domain infrastructures

In some instances, a component of a knowledge infrastructure may serve one scientific domain alone. In other cases, a component serving one scientific domain may also serve as part of a knowledge infrastructure for others. Over time, these relationships tend to shift, requiring both infrastructures and domains to adapt, sometimes willingly, sometimes less so.

The *Large Synoptic Survey Telescope* (*LSST*), a telescope under construction with a projected budget of $1.1 billion, is an important example of a multi-domain infrastructure. LSST promises to provide data unprecedented in scale and scope for multiple domains within astronomy (*LSST Science Collaboration et al., 2009*). The process of building LSST requires negotiation between domains on decisions such as how to allocate scarce observing time during the planned ten years of data collection. This multiple-domain infrastructure,

in turn, interacts with infrastructures serving single domains within astronomy. These include the *Dark Energy Survey*, and the *Gaia Telescope*, which focuses on the Milky Way (*Dark Energy Survey, 2014*; *European Space Agency et al., 2016*). These single-domain infrastructures were planned and began operations after significant funds were invested in LSST. The sociotechnical configurations of each project are shaped by LSST and vice versa (*European Space Agency et al., 2016*; *LSST Science Collaboration et al., 2009*). The authors of this paper also are involved in a case study of LSST (*Borgman et al., 2015*).

Examples abound of interactions between single- and multiple-domain knowledge infrastructures. One example is between infrastructure that serves researchers studying earthquakes in Southern California only (*Southern California Earthquake Center , 2016*) and infrastructure that serves multiple domains related to the formation of volcanoes in the continent of North America (*Earthscope, 2016*). A second example is between infrastructure intended to support a single domain studying coastal dynamics and multi-domain infrastructure that collects and distributes data about oceans, coastal regions, and the US Great Lakes (*Center for Coastal Margin Observation & Prediction, 2015*; *NOAA, 2016*).

Domains sharing infrastructure

While infrastructures that serve individual domains have received the most study, other important infrastructures span multiple, and often competing domains. Domain-spanning infrastructures often represent significant investment of public research funding and are critical sources of data for the research communities they serve. Among the few shared infrastructures receiving scholarly attention is *Argos*, whose study by *Benson (2012)* revealed how marine biologists were able to negotiate a share of its infrastructural resources. Argos is a satellite-based environmental surveillance system that provides data for oceanography, meteorology, and marine biology (*Ortega, 2003*). Marine biologists persuaded Argos leaders to collect data elements important to them by appealing to the interests of their commercial partners. In return, the biologists compromised other elements of their data collection methods to satisfy Argos partners in oceanography and meteorology.

Another perspective on negotiations in shared infrastructure is offered by *Ribes & Finholt (2008)*, in their study of building infrastructure for studying the water environment. The forerunner to this infrastructure served a single community of researchers in environmental engineering, but the new infrastructure was intended to serve hydrological scientists as well. Ribes and Finholt show that the spokespeople of these domains negotiate infrastructure building more effectively when they represent a strong, clearly defined, and cohesive community. Various mechanisms for defining, canvassing the opinions of, and representing a community—such as organizing forums and conducting surveys—play an important role in facilitating these spokespeople's effectiveness. Unfortunately, the effort to build this infrastructure fell apart before it began scientific operations (*Jackson & Buyuktur, 2014*).

These studies suggest the importance of studying how domains negotiate processes of building and accessing infrastructural resources shared with other domains, and how the configurations of these infrastructures shape the work and organization of individual communities.

Data scarcity and abundance

As science became more professionalized in the early twentieth century, more emphasis was placed on mathematical methods. These methods, in turn, require data generation and statistical confirmation (*Bowler & Morus, 2010*). That shift in focus endowed science with greater cultural authority as the primary knowledge-generating institution within society (*Porter, 1996*). The tight coupling of mathematization, data, and cultural authority helps to explain why domains that experience data scarcity are often so concerned with increasing their volume of data (*Sawyer, 2008*). For example, in molecular biology, increasing quantification and statistical inference were driven by "an ever-present *methodological anxiety* manifested in the constant search for an increased *objectivity* –or in its converse: the avoidance of *subjectivity*" (italics in original). These methodological changes both require and accommodate increasing quantities of data (*Suárez-Díaz, & Anaya-Munoz, 2008*, p. 452). By increasing quantification and data-intensive practices, communities can increase their scientific credibility (*Hagen, 2003*; *Kay, 2000*). Indeed, this increasing quantification frequently shapes, and is shaped by, hypothesis testing with laboratory-generated data (*Lenoir, 1999*; *Paul, 2009*).

Least studied are situations where "no data" exist, whether because no data were collected about a particular phenomenon, data may have existed but were not curated for longer term use, or data still exist but cannot be discovered or retrieved for a variety of reasons (*Borgman, 2015*).

Increasing data production

Social mechanisms can encourage greater production of data and use of statistical methods. As domains develop norms that promote data-intensive research, those who eschew such approaches may be marginalized (*Keller, 1984*). A domain's quest for enhanced status may drive changes at the institutional level, even leading to the development of new domains. Natural historians in the early 20th century, in the face of profound challenges to their domain's status, made alliances with the more data-intensive and mathematically based discipline of population genetics, forming the new discipline of evolutionary biology (*Ceccarelli, 2001*).

Domains sometimes address scarcity by producing large volumes of data. A notable example is cosmology, traditionally regarded as the poor relation to other domains of study in astronomy (such as lunar and solar astronomy) because they lacked sufficient data to support fundamental conjectures (*Kragh, 1996*). Cosmologists' concern with this lowly status motivated the emergence of large telescope projects, known as sky surveys, that collect vast quantities of data about astronomical phenomena (*Sloan Digital Sky Survey, 2016*). Cosmology is now characterized by the use of data-intensive statistical methods (*Strauss, 2014*). Similarly, molecular biology addressed data scarcity by developing the *Human Genome Project* (*Lenoir & Hays, 2000*).

Gaining access to extant data

One way for a domain to address its data scarcity is to negotiate access to extant infrastructures, which often occurs in parallel with adopting computational- and data-intensive epistemologies (*Chow-White & García-Sancho, 2012*). Scientific databases are a

means to organize and classify information, providing "a contemporary key to both state and scientific power" (*Bowker, 2005*, p. 108).

Infrastructure projects, in turn, may increase data production and circulation by shaping the behavior of scientists. Interoperability strategies include imposing or embedding common standards, such as what counts as reliable evidence, acceptable research methods, and data management practices (*Bowker, 2005*; *Leonelli & Ankeny, 2012*). Infrastructure can also foster norms of behavior that encourage greater openness within a scientific community, including openness around data (*Leonelli, 2010*). Contributions to a database can, in turn, encourage greater willingness to contribute to a database, resulting in a self-reinforcing cycle of increased data availability and normative shifts towards greater data openness. Databases can help encourage the circulation of types of data that the database does not support by creating expectations that scientists will share data when asked (*Leonelli & Ankeny, 2012*). *Kelty (2012)* discusses how scientific newsletters in biology not only constituted an infrastructure to build communities around particular organisms, but also promoted sharing of research objects by requiring openness of researchers as a condition of receiving the newsletter (and thus continued membership of the community).

Lastly, and perhaps most importantly for knowledge infrastructures, is to address scarcity by building infrastructure that aggregates extant data or databases (*Meyer, 2009*), or that integrates existing sites of data production into a single infrastructure (*Aronova, Baker & Oreskes, 2010*). These databases play critical roles in supporting research and in fostering data-intensive methods (*Bowker, 2000*; *Leonelli, 2012*). One example is the *Long Term Ecological Research Network*, whose infrastructure attempts to improve the management and accessibility of data produced by distributed ecological research stations. This network originated in the efforts of biologists to leverage "Big Science" funding opportunities by collecting and organizing large-scale datasets (*Aronova, Baker & Oreskes, 2010*).

However, a major barrier to data circulation and integration is the use of disparate standards by individual scientists, a problem particularly acute when scientists come from different disciplines or communities of practice (*Baker, Jackson & Wanetick, 2005*; *Bietz & Lee, 2009*; *Borgman et al., 2015*; *Borgman et al., 2007*). Scientists' concerns about control of data, authorship rights, and incentives to undertake the work necessary to make their data shareable also limit the adoption of these infrastructures (*Borgman, 2015*; *Borgman, Wallis & Enyedy 2007*).

Research questions

Data scarcity is a critical problem for many scientific communities. Much richer accounts are needed of how domains employ infrastructural approaches to data scarcity. One approach is to emphasize scarce resources when negotiating access to infrastructure. As discussed above, competition arises when an infrastructure attempts to accommodate all types and forms of data present in the participating communities, particularly in the face of conflicting data standards and formats. A second approach is to consider relationships between an infrastructure and others with which it overlaps. A single domain may rely upon multiple infrastructures simultaneously. A domain that builds an infrastructure to increase its flow of data may be expressing dissatisfaction with the existing infrastructure.

New infrastructure is likely to address the shortcomings and exploit affordances of existing infrastructures. These considerations motivate our three research questions:

(1) How do scientific domains experience data scarcity?
(2) How does a scientific domain address data scarcity through developing knowledge infrastructures?
(3) What interactions occur between infrastructure specific to a domain and infrastructure shared with other domains?

DEEP SUBSEAFLOOR BIOSPHERE RESEARCH

Studies of the deep subseafloor biosphere draw together scientists from multiple physical and life science backgrounds who bring a wide variety of perspectives, tools, and methods (*Darch et al., 2015*). Researchers pursue their scientific goals by collecting and analyzing rocks from the seafloor, known as *cores*. Their work involves data about the microbial communities in these cores and the core's physical properties, such as geochemical or hydrological.

Scientific ocean drilling cruises are the primary source of cores for this community. From 2003 to 2013, the Integrated Ocean Drilling Program (IODP) operated these cruises. Scientific studies of the oceans receive extensive financial support from governments. As *Mukerji (2014)* argues, this support depends on the ability of oceanographic research to address questions of major social and political concern. Such concerns can include national defense, environmental issues, fisheries, and more recently, telecommunications. Thus, in 2013, US government support for scientific ocean drilling led to the International Ocean Discovery Program (IODP2) as the replacement for IODP. IODP2 uses the same ships, drilling technologies, and other resources as IODP, and provides infrastructure for many scientific domains, including plate tectonics.

The Center for Dark Energy Biosphere Investigations (C-DEBI) is a *Science and Technology Center* (*STC*) funded by the US National Science Foundation (NSF), and launched in September 2010. C-DEBI was initially funded for five years, and successfully renewed for an additional five years (2015–2020). C-DEBI, which provides infrastructure for deep subseafloor biosphere researchers, has two main aims. One is to foster a community of researchers to study deep subseafloor microbial life, and the second is to promote scientific work on microbial ecology of the seafloor. This scientific work explores the role of subseafloor microbes in global environmental processes to improve knowledge of the origin, evolution, and extent of life on earth.

These researchers are geographically distributed, with the Principal Investigator (PI) and four co-PIs based at five universities across the US. C-DEBI funding covers projects conducted by over 90 scientists in more than 40 universities and research institutions across the USA, Europe, and Asia (*Center for Dark Energy Biosphere Investigations, 2016*).

C-DEBI was established before the NSF's requirements for Data Management Plans, which began with proposals submitted in 2011 (*National Science Foundation, 2010*). However, C-DEBI was required to develop a plan for its renewal application in 2015 (*Center for Dark Energy Biosphere Investigations, 2012a*). By this time, senior personnel in

C-DEBI had become more aware of the inter-domain politics in IODP/IODP2, having participated in expeditions during 2011 and 2012. In response to this awareness, and to requirements for a data management plan, the process of constructing infrastructure for scientific data management in C-DEBI began. Thus, C-DEBI is itself a knowledge infrastructure for the domain of deep subseafloor biosphere research and C-DEBI has responsibility for developing further infrastructure components.

Our focus in this article is the relationships between the IODP/IODP2 infrastructure and the C-DEBI infrastructure, and the influence of data scarcity or abundance. Our account pays attention to these relationships in the period up to C-DEBI's renewal in summer 2015, although work on infrastructure continues. We highlight the critical role of negotiations between scientific domains as they contest scarce infrastructural resources. Single-domain infrastructures are both a response to, and an intervention in, these negotiations. Our study advances research on interactions between representatives of different domains (such as in formal meetings, and informal encounters), and as domains build infrastructures for themselves. We also advance research on the motivations for domains to build infrastructure for themselves by examining how such infrastructure is configured to leverage and intervene in the control of shared infrastructure.

Analyzing this process, therefore, provides a valuable opportunity to understand the relationships between single- and multi-domain infrastructure, and is part of a larger study of knowledge infrastructures at multiple scientific sites (*Borgman et al., 2015*).

RESEARCH METHODS

We present findings from an eighteen-month, qualitative case study of scientists studying the deep subseafloor biosphere by focusing on C-DEBI and the ocean drilling programs on which it depends, the Integrated Ocean Drilling Program (IODP) and its successor, the International Ocean Discovery Program (IODP2). This community affords rich opportunities for answering our research questions, enabling us to explore relationships between C-DEBI and IODP/IODP2; to ask how and why scientists engage in building, configuring, and negotiating these infrastructures to access data; and how the scarce resources of IODP/IODP2 are contested between multiple domains. Our human subjects research was approved by the UCLA North General Institutional Review Board (#10-000909-CR-00002).

Data collection

A key feature of this case study is long-term ethnographic observation of C-DEBI (*Hammersley & Atkinson, 2007*). We were embedded for eight months in a laboratory headed by a leading figure in C-DEBI at a large US research university, which involved one of the authors visiting the laboratory for two or three days per week to observe bench work and laboratory meetings. The first author also conducted weeklong observational work in two other participating laboratories in the US and joined researchers on a three-day field research expedition. We made extensive notes about what we observed, including the physical layout of the laboratories and laboratory benches, tools and methods used, and

patterns of collaboration. Our informants told us about their backgrounds, aspirations, and experiences in the laboratories and on scientific cruises.

These organizations are distributed across multiple institutions and countries, which posed issues of scalability for the researcher (*Star, 1999*). The work of C-DEBI and IODP/IODP2 spans more sites than a small team of researchers can visit, much less meet face-to-face with all personnel. We focused on the techniques and technologies—the "*scalar devices*"—employed by our research subjects to understand C-DEBI, IODP/IODP2, and their domain (*Ribes, 2014*, p. 158).

One scalar device that we observed was gatherings such as the C-DEBI All-Hands' Meeting and research workshops. A larger gathering was the American Geophysical Union Fall Meeting 2013 in San Francisco, a major conference for the deep subseafloor biosphere community and IODP/IODP2; we also presented our early findings at this conference. These events enabled our research subjects to take stock of the scale of the communities and infrastructures in which they are embedded, in terms of the people involved, organizational hierarchies, and the range of scientific work conducted. Another form of scalar devices is reports and workplans, which we studied as part of document analysis (see below).

The distributed nature of C-DEBI and IODP/IODP2 also means that work in these organizations often takes place through communications media. By using multiple forms of media, we could establish "co-presence" when "co-location" was not possible (*Beaulieu, 2010*). Co-presence involves the researcher witnessing how the work of scientific collaborations is conducted even when they are not physically (or necessarily temporally) collocated with the subjects of research. Lacking the opportunity to observe practices on board an IODP cruise, given the expense and limited places available, we studied IODP work conducted elsewhere. We attended online meetings and seminars where participation and data collection were planned, and watched a feature-length documentary that used footage from a deep subseafloor biosphere-focused cruise. Other online observations included workshops, meetings where key C-DEBI personnel planned infrastructure to coordinate data management across the project, and websites of organizations and people.

We assembled a corpus of documents for analysis. Documents such as instruction manuals for laboratory equipment help to explain the work conducted by C-DEBI-affiliated scientists in their laboratories. Other documents help us to interpret contexts in which C-DEBI scientists operate. These include official C-DEBI documents such as the funding proposal, Annual Reports to the NSF, and operating documents for C-DEBI and IODP/IODP2. These documents function as scalar devices by giving details and metrics about activities, plans, and available infrastructural resources.

Our research also draws heavily on semi-structured interviews; the sample for this article consists of 49 people, which includes C-DEBI-affiliated scientists, curators and managerial staff involved in IODP/IODP2, as detailed in Table 1. The column *Involved in IODP* indicates which C-DEBI interviewees are involved in decision- or policy-making in IODP/IODP2. These interviewees are further split into two groups: those in cruise operations, and those with the Consortium for Ocean Leadership, responsible for administering US involvement in IODP (but not IODP2). Interviews ranged in length

Table 1 The composition of our interview sample.

		Career stage	Interviewees	Involved in IODP
C-DEBI	USA-based	Undergraduate	**5**	0
		Graduate student	**9**	0
		Postdoctoral researcher	**7**	1
		Faculty	**13**	2
		Management	**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**	

from 35 min to two and one-half hours, with the majority being between one and two hours long. With the consent of the interviewees, interviews were recorded and sent to an external company for transcription.

C-DEBI interviewees were initially recruited from those scientists being observed in the laboratory, and were typically interviewed after an extended period of observation. Other C-DEBI interviewees were recruited from those who had been awarded C-DEBI-funded grants, with these interviews typically taking place over Skype. We have interviewed undergraduate and graduate students, postdoctoral researchers, faculty members, and non-scientists with senior level roles in administering and operating C-DEBI. IODP interviewees were identified and approached through a range of methods, including personal introductions from C-DEBI-affiliated scientists and other IODP personnel, and from public websites.

Our interviews covered a range of topics, including interviewees' backgrounds and career trajectories. We asked scientists and technical staff detailed questions about the scientific work they are undertaking and the importance and role of data in their work. We also asked IODP/IODP2 technical staff and C-DEBI scientists who have participated in IODP cruises about their work on board expeditions, how they negotiate for access to cruise resources, and how they transfer data, methods, techniques, and collaborative networks between cruises and their onshore laboratories. Non-scientists were asked about their roles in building and administering C-DEBI and IODP/IODP2 infrastructure.

Where interview quotations are used in this paper, we add a code in parentheses after each quote indicating whether they are IODP or C-DEBI, their career stage, and a number unique to each interviewee: (IODP curator, #1) or (C-DEBI faculty, #3) etc.

Data analysis

Our initial data analysis involved close reading of our ethnographic notes, interview transcripts, and documents. Based on these readings, we identified emerging themes about the relational, complex, and dynamic nature of knowledge infrastructures, and coded our

data accordingly. In particular, we focused on themes relating to how those we interviewed described their own work (scientific, organizational, building infrastructure); how they identified and defined communities of which they considered themselves members; what resources, both current and anticipated, they identified as necessary to their own work and to deep subseafloor biosphere research as a whole; what they considered to be infrastructure; and how they and their community engaged with other scientific communities to negotiate, access, and build infrastructure. We refined our coding scheme iteratively, going back and forth between our scheme and the data. Using a range of sources enables us to triangulate, cross-checking our data to validate our findings (*O'Donoghue & Punch, 2004*).

We began our data analysis after completing approximately six months of laboratory observation and fifteen interviews. We have thus been able to strike a balance between ensuring that our observations have not been biased by preconceived ideas and being able to assess our emerging findings and tentative hypotheses against further observations. We presented our emerging findings to the deep subseafloor biosphere research community at major scientific meetings (see above) for feedback and clarification.

RESULTS

Deep subseafloor biosphere research is a relatively new domain of study. Significant momentum for its emergence came in the late 1990s upon the publication of an article that extrapolated data about the size of microbiological communities in coastal sediments to the deep subseafloor (*Whitman, Coleman & Wiebe, 1998*). That article concluded that deep subseafloor microbes might constitute up to one-third of all of Earth's biomass. This claim had major implications for important questions of scientific and human concern, such as the global nitrogen cycle. As a new domain of scientific study, and one for which little data existed prior to its emergence, the strategy of founding scientists rested on acquiring more data for research. This pursuit of more and better quality data continues to be a critical force to this day.

Results are grouped into three sections. First, we frame the data scarcity problem in the terms of the deep subseafloor biosphere research community. Second, we discuss how this new research community established relationships with the international drilling programs, which is their major data source, and built some of their own complementary infrastructure. In this section, we also compare C-DEBI's choices of infrastructure for data management to those of other NSF Science and Technology Centers that were founded in the same time period. Third, we explore how the C-DEBI and ocean drilling programs worked together, in ways more and less successful, to develop a robust research community for deep subseafloor biosphere research.

Data scarcity

Complaints by deep subseafloor biosphere scientists about the "dearth of data" for their core research questions led to the founding of C-DEBI (*Edwards, 2009*: 5). Relevant data still exist only for a few sites in the ocean and, compared to studies of microbial ecology in other environments, is about relatively basic phenomena.

Two reasons emerged for the community's continuing concern with data scarcity. The first is constraints on scientists' ability to pursue their field's immediate research objectives, which are to characterize deep subseafloor microbial communities in terms of the quantity and types of microbes that exist, how these microbes interact with the physical environment they inhabit, and how microbial communities vary between geographic sites on the seafloor (*Orcutt et al., 2013*). As a C-DEBI report stated, "Evidence for microbial alteration [of the physical environment] exists, yet scientists lack robust molecular, biochemical, or physiological data so needed" (*Center for Dark Energy Biosphere Investigations, 2011*, p. 11).

The second concern about data scarcity relates to the status of deep subseafloor biosphere research relative to studies of microbial ecology in other environments, and thus for their ability to attract future resources and funding. In the words of many scientists that we studied, research on the deep subseafloor biosphere is largely exploratory or *discovery-driven*, while research on microbial ecology in other environments is largely *hypothesis-driven*:

> *"Our work in the lab in general tends to be classified as rather exploratory as opposed to hypothesis-driven. This is something…that I met researchers who take issue with because they insist that to be a true science, proper science, you need to have a question and then you need to have a methodology that will either answer 'yes' or 'no', or some number. Whereas, our kind of science is oftentimes, it's more like, 'I wonder if...' And then you try something and the results are occasionally interesting, and then you go, 'Look what I found.' You didn't know what you were looking for, you just cast a big net out." (C-DEBI graduate student, #1)*

As some C-DEBI scientists stated the problem, if the approach of deep subseafloor biosphere researchers has a lower scientific status than studies of microbial ecology in other environments, they will receive fewer or smaller grants because "funding agencies will rarely fund basic discovery science" (*Teske, Biddle & Schrenk, 2011*, p. 9).

As the deep subseafloor biosphere emerged as a domain of study, researchers adopted a strategy of building and leveraging infrastructure to acquire more data. One of their strategies was to build infrastructure specifically for deep subseafloor biosphere researchers, first by establishing C-DEBI. The second strategy was to use C-DEBI as a means to gain greater access to, and to reconfigure, IODP2 for their advantage. Notably for our research questions, IODP/IODP2 is an infrastructure that deep subseafloor biosphere research shares with other scientific domains.

IODP/IODP2 provides the requisite infrastructure for the C-DEBI community to access the geographic sites and to acquire the physical samples necessary to produce data about the deep subseafloor biosphere. C-DEBI was initially established as a way to consolidate the position of the deep subseafloor biosphere within IODP and to recruit new researchers into the domain. Over time, C-DEBI also began to build its own infrastructure to respond to the limitations of IODP in providing data, and to reconfigure the IODP2 infrastructure so that a greater share of IODP2 resources will be allocated to deep subseafloor biosphere researchers, thus increasing their supply of data.

Ocean drilling meets deep subseafloor biosphere research

Interest in deep subseafloor microbial life that emerged in the late 1990s coincided with institutional challenges facing scientific ocean drilling programs. The predecessors to IODP that ran from 1968 to 2003 focused almost exclusively on physical science research. These programs facilitated major scientific successes. Best known is the evidence for the theory of plate tectonics and continental drift (*Committee on the Review of the Scientific Accomplishments and Assessment of the Potential for Future Transformative Discoveries with US-Supported Scientific Ocean Drilling, 2012*). Many scientists were concerned that funding for ocean drilling would cease with the anticipated end of the *Ocean Drilling Program*, IODP's immediate predecessor, in 2003. Expanding the drilling mission to include the deep subseafloor biosphere provided the momentum necessary to secure funding for IODP to launch in 2003. One of our interviewees, a senior administrator within IODP, quoted Admiral Watkins, who then headed the Joint Oceanographic Institutions, "I can remember …him saying, 'Give me bugs and I can give you a new program" (IODP policy, #1).

C-DEBI as a single-domain infrastructure

To reconstruct the origins of C-DEBI as an infrastructure for the emergent community of deep subseafloor biosphere researchers, we drew heavily upon documentary sources to complement our interviews and ethnography. The round of proposals for the 2003 launch of IODP was a critical inflection point. Deep subseafloor biosphere research was one of three major scientific foci for IODP (*Integrated Ocean Drilling Program (IODP) Planning Sub Committee (IPSC) Scientific Planning Working Group, 2001*), which planned four to five expeditions per year. Proposals were required to state the scientific objectives of the cruise and to identify the sites a cruise would visit. Three separate teams of scientists independently, and successfully, submitted proposals (in 2003, 2005, and 2007 respectively) for expeditions focused primarily on the deep subseafloor biosphere (although an IODP/IODP2 cruise will typically have a focus on one particular scientific domain, it will nevertheless involve scientists from all domains of study represented within IODP/IODP2).

Rather than work independently, the successful teams joined forces to coordinate the three biosphere-focused IODP cruises, which were planned for 2010 and 2011. To consolidate the position of deep subseafloor biosphere research within IODP, in 2008 they established the *Dark Energy Biosphere Investigations Research Coordination Network*. This network had four specific goals (*Edwards & Amend, 2008*, p. 3):

- "Develop an interactive *community* of deep-biosphere researchers
- Facilitate *coordination* of science *between* deep-biosphere drilling projects
- Stimulate *interaction* and *education* among disparate disciplines
- Enable *synthesis* and *integration* of data and technology advances"

This network brought together scientists in regular face-to-face meetings, with the proposal to the NSF for C-DEBI in 2009 arising from one such meeting. C-DEBI, as introduced above, is a Science and Technology Center (STC) funded by the US National Science Foundation (NSF), initially for five years from 2010–2015, and subsequently renewed for another five years from 2015–2020. The proposal to establish C-DEBI set

out two major goals. One is "to coordinate, integrate, support, and extend the science" of deep subseafloor biosphere research, and the second to help "foster and educate an interdisciplinary community of deep subseafloor biosphere researchers, with a focus on students and junior researchers" (*Edwards, 2009*: 1). The focus on students and junior researchers highlights the aspiration to establish an enduring community. The significance of this long-term view was highlighted, sadly, by the deaths of two of the five founding co-investigators during the first five years of C-DEBI. It is a tribute to their vision that the collaboration was sufficiently robust to reorganize with new investigators and to win the second five-year award.

C-DEBI has developed social, technical, and policy means to pursue its aims, including funding support for scientists, a website and meetings to circulate knowledge and to bring together globally distributed researchers, and an infrastructure for data management that continues to evolve. In its early years, C-DEBI focused more on recruiting scientists to study the deep subseafloor biosphere than on building technical or policy infrastructure. Recruiting occurs by distributing grants to support small-team (usually two or three persons) laboratory-based research projects (typically one to three years in length), or graduate and postdoctoral fellowships. These grants typically support projects in which scientists use cores collected from scientific ocean drilling cruises to produce new data in their onshore laboratories, and to support projects that develop new techniques to study the deep subseafloor biosphere. To date, nearly 90 such grants have been distributed across more than 40 institutions in the USA (*Center for Dark Energy Biosphere Investigations, 2016*). However, in the early years of C-DEBI, these grants were not accompanied by a strategy for managing the data produced by the funded projects.

C-DEBI data portal

Although the C-DEBI proposal aspired to build data management infrastructure, stating that "C-DEBI will develop and maintain a website for public access and data sharing among the C-DEBI research community" (*Edwards, 2009*: 24), little was accomplished toward this goal in the first two years of operation. The first five-year Strategic Implementation Plan 2010–2015 claimed "technical difficulty" as the barrier to establishing the data management infrastructure (*Center for Dark Energy Biosphere Investigations, 2010*; 11).

However, by 2012, data management infrastructure became part of the work of C-DEBI. This plan stresses that the "C-DEBI STC is committed to open access for all information and data gathered during scientific research that is conducted as part of C-DEBI" (*Center for Dark Energy Biosphere Investigations, 2012a*: 1). Starting in 2012, C-DEBI responded to new NSF requirements by mandating that participating scientists must make data publicly available after a moratorium, typically two years. C-DEBI developed their data management strategy as part of their application for renewal of NSF funding for the second five-year period, 2015–2020. The most critical strategic challenge for C-DEBI during its first years of operation was to navigate this NSF renewal process successfully.

The C-DEBI Data Portal, comprising a registry and repository, is a central element of C-DEBI's strategic goals for their second five-year project (*Center for Dark Energy Biosphere Investigations, 2014a*). C-DEBI requires participating teams to register all associated datasets

that support published results on the portal, and to deposit them in a relevant, online, publicly accessible database. The portal contains metadata about each dataset, including details about the provenance of the cores used (including the name of the research site, the cruise number, and the specific drill hole(s) where the samples originated), and a link to the dataset if it is hosted externally.

In mid-2013, C-DEBI assembled a team to build the data portal, which included one co-PI, a scientific database expert from another C-DEBI institution, and the C-DEBI Administrative Assistant. Since the Center's inception, the Administrative Assistant had responsibility for building and maintaining the C-DEBI project website. His job was expanded to include leading the development of the data portal. 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, 2014b*). The first phase, which included prototyping and site architecture, was completed for the NSF visit to C-DEBI in January 2014. Subsequently, the job title of the Administrative Assistant was changed to Data Manager, reflecting the shift in the importance to C-DEBI of data management infrastructure from a marginalized aspiration to a central goal.

Alternative approaches to single-domain infrastructure

The existence of the data portal, and the form it takes, is only partially determined by NSF requirements. To determine the degree to which the form of C-DEBI data management infrastructure is driven by NSF data management requirements, we examined how other NSF Science and Technology Centers have addressed these requirements. C-DEBI is one of 11 current STCs, all of which are subject to NSF data release requirements. However, the STCs interpret these requirements in a variety of ways. Hence, only three of the other ten STCs operate an online, publicly-accessible registry or repository to make these data accessible, whether by downloading datasets, by providing links to other sources, or by providing contact information for people associated with datasets (*Center for Coastal Margin Observation & Prediction, 2015*; *Center for Microbial Oceanography Research and Education, 2015a*; *Center for Remote Sensing of Ice Sheets, 2015*).

The data registries or repositories of these STCs vary in the types of datasets they contain and in the scope of metadata they capture about each dataset. The *Center for Microbial Oceanography Research and Education*, or *C-MORE*, is the STC most scientifically comparable to C-DEBI, in that they study microbial ecology in the ocean and scientists use samples collected from ocean research cruises operated directly by C-MORE.

C-DEBI is distinguished from other STCs with online data infrastructure by requiring the most comprehensive set of metadata for datasets in its registry and repository. For example, while both the C-DEBI and C-MORE registries have a category to name the research cruise from which each dataset was derived (*Center for Microbial Oceanography Research and Education, 2015b*), C-DEBI also has a category for the precise geographic location from where the sample was drawn. Another example is that the C-DEBI registry has metadata categories detailing the publication(s) in which a particular dataset has been used; by contrast, the C-MORE registry does not.

The current STCs have implemented the NSF requirements in different ways, despite their scientific and organizational similarities. In other words, the existence of the NSF requirements do not completely account for what data management policies or infrastructure are implemented by C-DEBI, nor do they account for the specific form and function of these policies and infrastructure.

Shared infrastructure and community building

IODP/IODP2 expeditions are the primary source of physical samples and data used in the onshore laboratories of the deep subseafloor biosphere researchers we studied. Some scientists also participate in cruises operated by other organizations (*University-National Oceanographic Laboratory Services, 2015*). However, these non-IODP/IODP2 cruises usually revisit sites drilled during previous IODP/IODP2 expeditions.

International ocean drilling programs

IODP/IODP2 is infrastructure, consisting of ships, scientific equipment and personnel, and an organizational structure shared between the deep subseafloor biosphere community and several other domains. Consequently, these domains compete for scarce resources. Their representatives are involved in decision-making processes within IODP/IODP2 about the scope of their programmatic research and the organization of specific cruises. The expansion of IODP scientific activities to include deep subseafloor biosphere research required other domains to concede some of their share of IODP resources. While our C-DEBI participants generally reported a collegial atmosphere among scientists on IODP/IODP2 cruises, many deep subseafloor biosphere researchers nevertheless encounter resistance, such as objections to the number of cores allocated to deep subseafloor biosphere researchers:

> *"We encounter resistance [from other domains] when we apply to sail. We encounter it when we apply for sample requests. We encounter it when we set up for post-cruise fundings, and also for regular grant writing." (C-DEBI faculty, #1)*

Increased competition for places on cruises is but one of the ways in which other domains concede resources to accommodate research about the deep subseafloor biosphere. A wider variety of domains also compete for allocation of cores, the precious sources of physical and microbiological data from cruises. Initial decisions made on board the ships determine the core's value to these competing communities. The most critical initial decision for the microbiology community is the temperature at which cores are stored. While samples for physical analysis are typically stored at -4 °C, samples for microbiological analysis are typically stored at -80 °C to avoid contamination and to stabilize biological material. Other handling decisions include the ways in which the cores are cut and distributed to participating scientific teams. The number of cores per cruise is finite, therefore producing more cores suitable for microbiological analysis results in fewer cores suitable for physical science analysis:

> *"What you do with this core you just split these one meter sections lengthwise, open it up so you have two halves of it.... For microbiology and geochemistry you do it somewhat differently. You take the core, you're not cutting it up lengthwise but you cut out short*

sections...So you lose the entire stratigraphic information from that core." (C-DEBI faculty, #2)

The allocation of cores to research teams requires intensive negotiation. Pre-cruise meetings to negotiate allocations result in a Sampling Plan for the voyage. During a cruise, Sampling Plans frequently are adapted to changing conditions and to the relative success of core sampling, leading to further negotiation of resources:

"Sometimes . . . we don't receive that much [biological] core material. Sometimes you may have a small piece and 15 people want something from the small piece so then that has to go through iterations of compromise." (IODP curator, #1)

As a new area of science, a particular challenge for deep subseafloor biosphere researchers is their own methodological diversity. The lack of agreement within the deep subseafloor biosphere research community on standard practices for data handling works against them in negotiating for more cores. As discussed in more depth elsewhere (*Darch & Sands, 2015*), workflows to characterize microbiological communities in cores vary significantly, even between scientists working on adjacent benches in the same laboratory. Scientists from other domains sometimes use this variation to argue against allocation of cores to deep subseafloor biosphere researchers, as one of our interviewees encountered:

"Those that are competing with us for sediment material, the hard rocks guys, the sedimentologists, those guys that are then lobbying for the same sediment samples that we're going after, they can turn to us and say, 'Well, you know what, if we handed you half of this and you half of this, you guys would come back with two different datasets, so what's the point of handing it to any of you because you guys can't describe it in the same way anyway?' " (C-DEBI faculty, #1)

As IODP approached the end of its funding period in 2013, concerns arose among stakeholders about continued government funding for scientific ocean drilling cruises, and indeed, funding was reduced for IODP's successor, IODP2 (*Committee on Guidance for NSF on National Ocean Science Research Priorities: Decadal Survey for Ocean Sciences, 2015*). However, the physical infrastructure (cruise ships, core repositories, data management systems) of IODP2 is largely that of IODP.

In addition to maintaining their position within IODP2, deep subseafloor biosphere researchers also aspire to secure more resources for deep subseafloor biosphere research in the future. One aspiration is to produce more standardized data on board the ocean drilling cruises, akin to the standardized set of analyses of physical properties that are routinely conducted and made publicly available through an IODP2 database. As no comparable set of standards exists for the microbiological properties, cruises neither conduct nor report basic microbiological descriptions of cores. Many of the C-DEBI-affiliated scientists interviewed mentioned this lack of agreement around standardized methods as a serious constraint on their scientific progress (*Orcutt et al., 2013*). Instead, individual scientists devote much effort to basic microbiological analyses in their home laboratories. The time and expense of basic description limits the resources available to conduct more advanced analyses, as explained by one of our interviewees:

> *"Post-expeditions Awards provide $15,000 worth of support for up to two years, for you to do the research that you proposed to do while you were at sea … The difference between what we can use that money for, and say what a sedimentologist can use that money for, is grossly different. Because a sedimentologist, the geochemist, the petrologist, the paleo-mag guys, all of them pretty much have all the data. And so, they're looking at the $15,000 as seed money to maybe do some analysis that they maybe pay for a grad student, maybe pay for a technician, maybe pay for somebody's time, to analyze it, to maybe take it another direction…For the biologist, we have $15,000 to now process all of our samples, do all the sequence analysis, do the bulk labor of all of our work on the equipment that we already have to have in our lab versus what everybody else is using on the ship." (C-DEBI faculty, #1)*

However, to include microbiological description of cores on board all cruises would require major reconfiguration of existing cruise practices. These practices were established over many years before microbiology's inclusion in cruises:

> *"It took them decades to come up with the system... standard protocols, standard procedures, standard storage. It makes it a little bit rigid like I said, when you do something new and novel, like the living things don't really have a place yet." (C-DEBI faculty, #4)*

Consequently, attempts to change IODP2 practices in support of deep subseafloor biosphere research would require a significant amount of effort and would diminish the resources available for more physical science-oriented domains in IODP2. To date, these attempts have faced considerable resistance:

> *"Geochemical and geophysical research objectives represented on IODP expeditions are routinely provided by dedicated shipboard scientists and technicians assigned to completing standard procedures on all core material. The call for an additional biological workload on these individuals is typically met with an argument claiming a lack of time and resources on board." (Orcutt et al., 2013, p. 8)*

One of our interviewees told us that much of the resistance from these physical scientists is that, "as a community [of deep subseafloor researchers], we can't agree on anything" regarding methodology. While this resistance on the part of physical science disciplines appears to be motivated by a fear of losing IODP resources, their arguments are that the lack of standardization of microbiological workflows means that microbiological analyses cannot be included as part of the workflow on IODP cruises.

Consequently, many C-DEBI scientists recognize that the deep subseafloor biosphere community must undertake the work necessary to standardize microbiological workflows.

Infrastructure convergence

The development of the C-DEBI data management infrastructure can be understood in the light of the aspirations of deep subseafloor biosphere researchers to improve exploitation of currently-available resources from IODP/IODP2 cruises, and to reconfigure IODP2 infrastructure to secure a greater share of drilling cruise resources. Those developing C-DEBI infrastructure are working towards three goals: better curation and circulation of

extant data; building a community with norms of open data; and explicating the roles of IODP/IODP2 data in C-DEBI research. Here we examine how C-DEBI is addressing these three goals, and how the goals both shape, and are shaped by, the relationships between C-DEBI and IODP/IODP2 infrastructure.

Improving curation of extant data

The first goal of the C-DEBI data management infrastructure is to improve the curation, circulation, and accessibility of data handled by the deep subseafloor biosphere researchers. At a project workshop held in 2012 that brought together many leading members of C-DEBI, "encouragement of data sharing …was identified as an important priority" (*Center for Dark Energy Biosphere Investigations, 2012b*, p. 3). Despite the scarcity of cores and of resources to analyze them, deep subseafloor biosphere researchers often do not take the steps necessary to preserve these data beyond the short-term or to make them easily accessible to fellow members of their domain.

This situation is compounded by disparate policies and uneven provision of community databases across the scientific disciplines involved in deep subseafloor biosphere research. Scientists who publish in most microbiological journals are required to deposit genomic sequence data supporting their conclusions to publicly accessible databases such as *GenBank* (*Benson et al., 2013*), although these databases do not cover the full range of microbiological data produced by deep subseafloor biosphere scientists. No similar requirement to deposit exists for complementary physical science data. Some appropriate databases do exist but contributions of data are at the discretion of the scientist, and occur rarely, as illustrated by this quotation:

> *"Nowadays they won't publish your work if it has molecular [biology] data and it's not in the database somewhere… There are now databases where you could, I guess, submit this type of data like the geological data. But I haven't started doing that yet." (C-DEBI postdoctoral researcher, #1)*

The C-DEBI Data Portal is intended to provide a home for these microbiological data. At one level, the goal of improved curation, accessibility, and circulation of data can be understood as a desire to increase the amount of scientific work accomplished from the limited amount of cores and data currently available. However, this goal also can be understood in terms of pursuing longer-term strategic objectives of the deep subseafloor biosphere community. Standardized approaches will support data integration and meta-analyses, for example. Better data curation will enable "our community to develop and recommend broad standards" (*Center for Dark Energy Biosphere Investigations, 2013b*, p. 4), in turn helping to promote some of the methodological standardization necessary to address the criticisms of physical science researchers.

Deep subseafloor biosphere researchers have promoted standardized methods as a means to advance hypothesis-driven science and replicability (*Teske et al., 2011*). They expect better replicability to address concerns about the status of their field.

The belief that methods for deep subseafloor biosphere research should be standardized was far from unanimous, however. As explored in more depth in *Darch (2016)*, some

members view methodological heterogeneity as a key strength of the domain, with researchers from diverse disciplinary backgrounds bringing new methods to bear on research questions. A particular concern is that standardizing methods of this emergent domain is premature, and will foreclose possibilities for more efficient or reliable methods in the future. Although these debates are ongoing, proponents of standardization hold key decision-making positions within C-DEBI. Consequently, the design of C-DEBI data infrastructure promotes methods standardization.

Community building

Since C-DEBI's conception, project leadership recognized the potential of data infrastructure to link distributed scientists (*Edwards, 2009*). In particular, the C-DEBI Data Portal is intended to "support the connection among scientists and others in …C-DEBI" (*Center for Dark Energy Biosphere Investigations, 2013a*, p. 9). The portal is an important element to sustain and expand the community beyond the project's anticipated end in 2020, given that 10 years is the maximum NSF STC funding. As one of our interviewees explains, "The web-based database for the entire subseafloor biosphere community will be an important legacy of C-DEBI's contribution" (C-DEBI management, #1). C-DEBI hopes that its Data Portal will emulate other successful scientific databases that began as project-specific endeavors and became institutionalized with subsequent funding.

However, C-DEBI does not merely seek to build a community through its data management infrastructure; it also seeks to foster particular norms among community members, notably a collaborative ethos predicated on openness and sharing of knowledge and knowledge products. As a consequence of this design and policy, scientists who do not conform to the norms of openness and data sharing are likely to be excluded. In turn, by fostering openness norms, data will be more widely available and exploited more fully. The norms are not ends in themselves, but intended to address scientific and strategic goals of studies of the deep subseafloor biosphere to obtain the needed volume and variety of data. Furthermore, C-DEBI also hopes that by helping to foster and strengthen an enduring community of researchers, their project's legacy will ensure continued strength in advocating for deep subseafloor biosphere's role in IODP2.

Explicating IODP/IODP2 contributions to C-DEBI

Given the resistance from physical science disciplines, and in light of future uncertainties around funding, deep subseafloor biosphere researchers must demonstrate that their inclusion in scientific ocean drilling cruises has resulted in scientifically valuable output. By demonstrating scientific value, they hope to secure continued IODP2 resources and to reconfigure IODP2 infrastructure in their favor. Articles in scholarly journals demonstrate scientific productivity, but do not necessarily highlight the essential or precise contributions of IODP data to their findings. Subseafloor biosphere research reports tend to integrate data from multiple sources, including cruises conducted under the auspices of organizations other than IODP/IODP2. Journal articles also report analyses of data derived in laboratory conditions that are based on cores or on instruments in drill holes left by those cores.

The challenge facing the C-DEBI community, therefore, is to make more explicit the relationship between their own scientific output, IODP/IODP2 cruises, and other kinds of

data. One way in which C-DEBI is addressing this challenge is by assigning appropriate categories of metadata in its Data Portal. These categories include information about the origin of the cores from which the data have been derived, such the name of the research site, the cruise number, and the specific drill hole(s) where the samples originated.

Such metadata serves several purposes. One is to improve the provenance of the data, which enhances replicability. Another is to improve integration of data from multiple sources. A longer-term benefit is to provide evidence that deep subseafloor biosphere researchers can use in negotiations with IODP2, both to gain more access to cruises and to reconfigure IODP2 practices in describing biological characteristics of cores. The need to demonstrate the value of IODP/IOPD2 to C-DEBI research thus motivates the development of the C-DEBI Data Portal and the choices of metadata categories.

DISCUSSION

Deep subseafloor biosphere researchers faced data scarcity, which they addressed by attempting to increase the volume and variety of data available to them. The entire research area emerged in the late 1990s and early 2000s when scientific data about the subseafloor became a viable goal. As more data became available, strategies for growing their research base evolved. This community continues to seek more data as a means to accomplish science at faster rates. As they matured as a community, they became increasingly concerned about their scientific status in relation to studies of microbial ecology in other environments. They altered their strategy for advancement accordingly (*Kragh, 1996*; *Lenoir, 1999*; *Paul, 2009*).

Here, we explicate two broad themes that emerge from our case study. The first is data scarcity—what it means for a scientific domain to experience data scarcity, what the implications are for its status, and how the domain addresses this scarcity. The second is the politics of knowledge infrastructures. A scientific domain may build and configure infrastructure specific to itself and also infrastructure shared with other domains.

Data scarcity

Terms like "big data" and "little data" are commonly employed to denote the scale of data to which scientific domains have access (*Borgman, 2015*). "Data scarcity" is a more poignant term as it suggests a state that is unsatisfactory for a domain's practitioners (*Sawyer, 2008*). The degree of data scarcity can be understood only in relation to that domain's scientific objectives.

As emergent scientific domains such as the deep subseafloor biosphere struggle to survive and thrive, they must attract resources to support researchers and infrastructure. To justify support for public funding in highly competitive environments, scientific domains must demonstrate their utility by contributing to one or both of the following:

1. Issues of major political and social concern (*Mukerji, 2014*), such as those relating to the environment, agriculture, and national defense;
2. Existential questions about humanity, such as the origins and evolution of life, and the origins and extent of the universe (*Bowler & Morus, 2010*).

C-DEBI makes both claims: study of the deep subseafloor biosphere will contribute significantly to understanding the effects of global environmental change, and to the

origins and evolution of life. The deep subseafloor biosphere domain faces data scarcity because it aspires to pursue its scientific objectives in a more statistically intensive manner than is afforded by the data to which it currently has access. For instance, domain scientists would like to answer questions about the global distribution of microbes, which is essential to understanding the role of the deep subseafloor biosphere in important environmental processes. To answer these questions, scientists must be able to perform meta-analyses that involve aggregating datasets about the size and composition of microbial communities in many different sites of the ocean. At present, insufficient data currently exists to make accurate estimates about the global distribution of microbes.

Leaders of the deep subseafloor biosphere community wish to shift the research emphasis from discovery to hypothesis-driven science, bringing their domain into line with other domains of microbiology. Hypothesis-driven science, involving statistical methods to test hypotheses, is generally regarded as more scientifically credible than discovery-driven science (*Lenoir, 1999*; *Paul, 2009*). These scientists wish to test the effects of environmental changes on the ability of different types of microbes to survive and thrive. Others wish to study microbes' abilities to survive and adapt to extreme conditions, such as those that may have been present when life on earth began.

Knowledge infrastructures

Our account of C-DEBI infrastructure is enriched by understanding relationships between this single-domain infrastructure and the complexities of IODP/IODP2, and by examining how infrastructure negotiations influence access to resources. We observe a mutual shaping of the single-domain and shared infrastructure, driven by the deep subseafloor biosphere researchers' desire to address their data scarcity.

The case presented in this paper contributes to studies of knowledge infrastructures (*Edwards, 2010*; *Edwards et al., 2013*) in two respects. First, although infrastructure components often are shared and negotiated between multiple domains, surprisingly few studies have paid close attention to the difficulties of sharing infrastructural resources (*Benson, 2012*; *Ribes & Bowker, 2008*; *Ribes & Finholt, 2008*). Our study pays close attention to how scientists from the deep subseafloor biosphere have negotiated a share of scarce resources of IODP/IODP2, thus extending research on negotiating shared infrastructure.

Second, very few studies offer accounts of how an infrastructure emerges in relation to other infrastructures upon which a domain may depend. The case presented in this article is an exemplar of configurations of knowledge infrastructure common to many domains of science. Here, the knowledge infrastructure of the deep subseafloor biosphere includes a major component shared with other domains, and another major component that it wholly controls. Further, this case demonstrates how building a single-domain infrastructure unfolds both in response to, and as a significant intervention in, the configuration of shared infrastructure. Although building the single-domain C-DEBI infrastructure may appear intended to provide immediate resources to scientists, it is also a means to access a greater portion of the shared infrastructure of IODP/IODP2.

How domains share or build their own infrastructures

Infrastructural approaches are commonly used to gain access to data resources, and to facilitate shifts toward more data-intensive epistemologies (*Bowker, 2000*; *Bowker, 2005*; *Chow-White & García-Sancho, 2012*). Scientific domains can engage in infrastructure-building activities to increase opportunities for producing, analyzing, aggregating, accessing, circulating, and long-term curating of data. Existing studies suggest these activities follow one of two strategies. The strategy most widely studied involves a domain that builds infrastructure intended exclusively for itself. The second strategy, documented in a few studies, involves a domain that constructs infrastructure to share with other domains, either by building new multi-domain infrastructure or by negotiating access to extant infrastructure that already serves other domains (*Benson, 2012*; *Ribes & Bowker, 2008*; *Ribes & Finholt, 2008*).

One contribution of our study is to demonstrate that a single scientific domain may pursue both strategies simultaneously. The deep subseafloor biosphere scientists constructed infrastructure specific to the domain (C-DEBI and its Data Portal) and negotiated greater access to infrastructure shared with other domains (ocean drilling programs such as IODP/IODP2).

The single-domain C-DEBI infrastructure was built in response to the constraints and opportunities of the multi-domain IODP/IODP2 infrastructure. When multiple domains share infrastructure, they compete for elements of design and operation that serve their needs, such as choices of what data are to be collected and what standards are to be applied. Infrastructure is built on an installed base, so that adaptations are both afforded and constrained by the configurations of extant infrastructure (*Star & Ruhleder, 1996*). Hence, when a scientific domain first seeks access to an infrastructure controlled by other domains, they may need to adapt their scientific practices accordingly (*Benson, 2012*). This was the situation faced by deep subseafloor biosphere scientists in gaining access to the IODP/IODP2 infrastructure, which had established procedures for collecting, curating, and accessing samples and data; for ship-based facilities; and had favored geographic locations for scientific ocean-drilling cruises. These pre-existing practices helped to shape and constrain the scientific research opportunities of deep subseafloor biosphere scientists.

One way to acquire more data is to distribute funding accordingly. Other studies have observed cases where infrastructure is designed to increase data production (*Lenoir & Hays, 2000*; *Strauss, 2014*) or to coordinate distributed sites of data collection (*Aronova, Baker & Oreskes, 2010*). The motivations of C-DEBI are similar, with top priority assigned to exploiting cores whose allocation between domains is hotly contested.

A second way to maximize scientific work is to aggregate and integrate existing data more effectively (*Leonelli & Ankeny, 2012*; *Meyer, 2009*), for instance by embedding common standards within or across domains (*Bowker, 2005*; *Leonelli & Ankeny, 2012*) or by building links between members of the domain and fostering norms of data sharing and openness across this domain (*Kelty, 2012*; *Leonelli, 2010*). The C-DEBI data infrastructure exists to foster collaboration among other deep subseafloor biosphere researchers, and to assemble and circulate data produced by distributed domain scientists.

Building and converging infrastructures

The single-domain C-DEBI infrastructure was designed in response to scarce IODP/IODP2 resources for deep subseafloor biosphere science. However, C-DEBI also can be understood as a means to negotiate a greater share of the IODP/IODP2 infrastructure.

Negotiations between domain representatives are a typical feature of projects to build shared infrastructure (*Ribes & Bowker, 2008*; *Ribes & Finholt, 2008*). *Ribes & Finholt (2008)* focus on the importance of defining and representing the interests of the communities involved. C-DEBI, and its associated data infrastructure, exemplifies such a strategy. It was formed to build a strong, enduring community that speaks with one voice, creating a larger presence of deep subseafloor biosphere scientists in IODP2 negotiations.

One way that C-DEBI increases the domain's status is by making explicit the contribution of IODP/IODP2 resources to their research. The C-DEBI Data Portal, and the choices of metadata categories, provide evidence that deep subseafloor biosphere researchers can use to negotiate further involvement in IODP2. A second way is to enable meta-analyses and cross-comparisons of methods that promote methodological standardization across the domain, with the goal of making microbiological workflows standard practice on future cruises. This standardization, in turn, is deemed necessary by many C-DEBI scientists to reconfigure how IODP2 cruises operate in the future, and thus to secure more data for their community.

Our case also demonstrates how relationships between single-domain and multi-domain infrastructure change over time. The infrastructure studied by (*Ribes & Finholt, 2008*), namely WATERS, fell apart even before it began its scientific operations. Our study, on the other hand, exposes how the configuration of C-DEBI, and the priorities of those building C-DEBI infrastructure, has shifted. In the early years of C-DEBI, distributing grants to enable data production and to recruit more scientists into the community was very much the priority. Over time, C-DEBI's priorities changed, following from experiences in negotiating with other domains for resources during the three biosphere-focused IODP expeditions in 2010 and 2011. Biosphere scientists realized both that standardizing microbiological workflows and making explicit how they are using cores to produce data are critical challenges for acquiring more data from future drilling cruises. In turn, this awareness promoted a greater focus of C-DEBI on building and configuring online data management infrastructure. Single-domain infrastructure is subject to change and reconfiguration as domain scientists gain more experience of, and become increasingly sophisticated operators in, using shared infrastructure. In turn, these shifts will change the configurations of resources available to scientists as they seek to go about their day-to-day work.

CONCLUSIONS

Scientists in many fields are concerned with increasing access to data as a means to advance their scientific work, to increase access to resources, and to enhance their status in the larger scientific community. Many scientific domains have addressed these concerns through infrastructural strategies. Very often, these strategies involve sharing infrastructure with other domains, whether this infrastructure is built anew, such as in the case of the Large

Synoptic Survey Telescope, or is gained by membership in other infrastructures, as in the case of the deep subseafloor biosphere and IODP/IODP2. Even when funding and data seem abundant, as in LSST, resources may be scarce and must be contested between domains.

In many cases, scientific domains will participate in shared infrastructure and in domain-specific infrastructure. In this article, we explored how the deep subseafloor biosphere community pursued an infrastructural approach to addressing data scarcity. Their data scarcity can be understood as a response to the challenges they face as an emergent domain in demonstrating their ability to contribute credibly to issues of critical social importance and interest. Both independent and shared infrastructures proved essential to this community's creation and maturation. Further, we identified the mutual shaping of these shared and independent infrastructures. The independent infrastructure was built both in response to, and as an intervention in, the configuration of the shared infrastructure.

We continue to study infrastructure, data management, and standards processes in the deep subseafloor biosphere research community—in particular as infrastructure continues to evolve in light of C-DEBI's successful renewal in 2015—and in astronomy to advance our understanding of relationships between infrastructure and epistemological practices (*Borgman et al., 2015*). In this article, we focus on the relationships between shared and domain-specific infrastructures and the difficulties of sharing infrastructures. The deep subseafloor biosphere community continues to evolve. Current pressures to standardize methods reveal the significant challenges in ensuring that this community can act as a single, strong, and united entity in negotiating access to shared infrastructure (*Darch, 2016*). Relationships between shared and domain-specific infrastructures should be studied across a wider range of scientific endeavors, as points of friction often reveal deeper truths about scientific practice (*Borgman et al., 2015*; *Edwards et al., 2011*).

ACKNOWLEDGEMENTS

This article is based in part on a paper presented at Association for Information Science and Technology (ASIS&T) Annual Meeting 2014, *Ship Space to Database: Motivations to Manage Research Data for the Deep Subseafloor Biosphere* (*Darch & Borgman, 2014*). We acknowledge the contributions of Milena Golshan, Irene Pasquetto, Ashley E. 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/IODP2 personnel who we interviewed and observed at work.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Research reported here has been supported by two grants from the Sloan Foundation Award: #20113194 (The Transformation of Knowledge, Culture and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective, Christine L. Borgman, PI, Sharon Traweek, Co-PI, UCLA); and #201514001 (If Data Sharing Is the Answer, What Is the Question?, Christine L. Borgman, UCLA, Subcontract to Peter T. Darch, University of

IL at Urbana-Champaign). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Sloan Foundation Award: #20113194, #201514001.

Competing Interests

Christine L. Borgman is on the Academic Advisory Board for PeerJ Computer Science.

Author Contributions

- Peter T. Darch conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Christine L. Borgman conceived and designed the experiments, wrote the paper, reviewed drafts of the paper.

Ethics

The following information was supplied relating to ethical approvals (i.e., approving body and any reference numbers):

UCLA North General Institutional Review Board
IRB#10-000909-CR-00002.

Data Availability

The following information was supplied regarding data availability:

The data in this article is human subjects data, and cannot be released given their sensitive nature and the infeasibility of de-identification.

REFERENCES

Aronova E, Baker KS, Oreskes N. 2010. Big science and big data in biology: from the international geophysical year through the international biological program to the long term ecological research (LTER) network, 1957–present. *Historical Studies in the Natural Science* **40(2)**:183–224 DOI 10.1525/hsns.2010.40.2.183.

Baker KS, Jackson SJ, Wanetick JR. 2005. Strategies supporting heterogeneous data and interdisciplinary collaboration: towards an ocean informatics environment. In: *Proceedings of the 38th annual Hawaii international conference on system sciences, 2005. HICSS '05*. 219b–219b.

Beaulieu A. 2010. Research note: from co-location to co-presence: shifts in the use of ethnography for the study of knowledge. *Social Studies of Science* **40(3)**:453–470 DOI 10.1177/0306312709359219.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. *Nucleic Acids Research* **41(Database issue)**:D36–D42 DOI 10.1093/nar/gks119.

Benson E. 2012. One infrastructure, many global visions: the commercialization and diversification of Argos, a satellite-based environmental surveillance system. *Social Studies of Science* **42(6)**:843–868 DOI 10.1177/0306312712457851.

Bietz MJ, Lee CP. 2009. Collaboration in metagenomics: sequence databases and the organization of scientific work. In: Wagner I, Tellioğlu H, Balka E, Simon C, Ciolfi L, eds. *ECSCW 2009*. London: Springer, 243–262.

Borgman CL. 2015. *Big data, little data, no data: scholarship in the networked world*. Cambridge: The MIT Press.

Borgman CL, Darch PT, Sands AE, Pasquetto IV, Golshan MS, Wallis JC, Traweek S. 2015. Knowledge infrastructures in science: data, diversity, and digital libraries. *International Journal on Digital Librariex* **16(3–4)**:207–227 DOI 10.1007/s00799-015-0157-z.

Borgman CL, Wallis JC, 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 CL, Wallis JC, Mayernik MS, Pepe A. 2007. Drowning in data: digital library architecture to support scientific use of embedded sensor networks. In: *Joint conference on digital libraries*. Vancouver: Association for Computing Machinery, 269–277.

Bowker GC. 2000. Biodiversity datadiversity. *Social Studies of Science* **30(5)**:643–683 DOI 10.1177/030631200030005001.

Bowker GC. 2005. *Memory practices in the sciences*. Cambridge: MIT Press.

Bowler PJ, Morus IR. 2010. *Making modern science: a historical survey*. Chicago: University of Chicago Press.

C-MORE. 2015a. Center for Microbial Oceanography Research and Education. *Available at http://cmore.soest.hawaii.edu* (accessed on 16 July 2015).

C-MORE. 2015b. C-MORE / Data. *Available at http://cmore.soest.hawaii.edu/datasearch/data.php* (accessed on 16 July 2015).

Ceccarelli L. 2001. *Shaping science with rhetoric: the cases of Dobzhansky, Schrodinger, and Wilson*. Chicago: University of Chicago Press.

Center for Dark Energy Biosphere Investigations. 2010. C-DEBI strategic implementation plan, 2010–2015. *Available at http://198.16.6.27/internal/docs/C-DEBI_SIP_2010Sep.pdf*.

Center for Dark Energy Biosphere Investigations. 2011. Center for dark energy biosphere investigations STC annual report 2011. *Available at http://www.darkenergybiosphere.org/wp-content/uploads/docs/2011C-DEBI_AnnualReport_noApp.pdf*.

Center for Dark Energy Biosphere Investigations. 2012a. C-DEBI data management philosophy and policy. *Available at http://www.darkenergybiosphere.org/internal/docs/C-DEBIDataManagementPlan_2012draft.pdf*.

Center for Dark Energy Biosphere Investigations. 2012b. C-DEBI "Activity" theme team 2012 workshop—report. *Available at http://www.darkenergybiosphere.org/wp-content/uploads/docs/2012C-DEBIactivity_WorkshopReport.pdf*.

Center for Dark Energy Biosphere Investigations. 2013a. C-DEBI strategic implementation plan 2013–2014. *Available at http://198.16.6.27/internal/docs/C-DEBI_SIP_2013-2014.pdf*.

Center for Dark Energy Biosphere Investigations. 2013b. C-DEBI "Activity" theme team 2013 workshop—report. Bigelow Laboratory for Ocean Sciences & Ocean Point Inn, East Boothbay. ME: C-DEBI. *Available at http://www.darkenergybiosphere.org/wp-content/uploads/docs/2013_C-DEBI_Activity_Workshop_Report.pdf*.

Center for Dark Energy Biosphere Investigations. 2014a. CDP. *Available at http://cdp.darkenergybiosphere.org*.

Center for Dark Energy Biosphere Investigations. 2014b. Center for dark energy biosphere investigations STC annual report 2013. *Available at http://www.darkenergybiosphere.org/wp-content/uploads/docs/C-DEBI-Annual-Report-2013.pdf*.

Center for Dark Energy Biosphere Investigations. 2016. C-DEBI-funded Projects. *Available at http://www.darkenergybiosphere.org/research-activities/funded-projects/* (accessed on 23 July 2015).

Chow-White PA, García-Sancho M. 2012. Bidirectional shaping and spaces of convergence interactions between biology and computing from the first DNA sequencers to global genome databases. *Science, Technology & Human Values* **37(1)**:124–164 DOI 10.1177/0162243910397969.

CMOP. 2015. Center for Coastal Margin Observation & Prediction. *Available at http://www.stccmop.org* (accessed on 16 July 2015).

Committee on the Review of the Scientific Accomplishments and Assessment of the Potential for Future Transformative Discoveries with US-Supported Scientific Ocean Drilling. 2012. Scientific ocean drilling: accomplishments and challenges. Washington, D.C.: National Academies Press.

CReSIS. 2015. Center for Remote Sensing of Ice Sheets. *Available at http://www.cresis.ku.edu* (accessed on 16 July 2015).

Darch PT. 2016. Many methods, many microbes: methodological diversity and standardization in the deep subseafloor biosphere. In: *IConference 2016 proceedings*. *Available at https://www.ideals.illinois.edu/handle/2142/89330*.

Darch PT, Borgman CL. 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. *Available at http://www.asis.org/asist2014/proceedings/submissions/papers/156paper.pdf* .

Darch PT, Borgman CL, Traweek S, Cummings RL, Wallis JC, Sands AE. 2015. What lies beneath? Knowledge infrastructures in the subseafloor biosphere and beyond. *International Journal on Digital Librarie* **16(1)**:61–77 DOI 10.1007/s00799-015-0137-3.

Darch PT, Sands AE. 2015. Beyond big or little science: understanding data lifecycles in astronomy and the deep subseafloor biosphere. In: *iConference 2015 proceedings*. Newport Beach, CA: iSchools, *Available at https://www.ideals.illinois.edu/handle/2142/73655*.

Dark Energy Survey. 2016. Dark energy survey. *Available at https://www.darkenergysurvey.org/* (accessed on 23 July 2016).

Earthscope. 2016. Earthscope. *Available at http://www.earthscope.org/* (accessed on 23 July 2016).

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. *Available at http://www.darkenergybiosphere.org/internal/docs/2009C-DEBI _FullProposal.pdf*.

Edwards PN. 2010. *A vast machine: computer models, climate data, and the politics of global warming*. Cambridge: MIT Press.

Edwards K, Amend J. 2008. Towards coordination and integration of deep marine biosphere research: the Dark Energy Biosphere Institute (DEBI). *Available at https://www.marum.de/Binaries/Binary18149/Edwards_DeepBiosphereDEBI.pdf*.

Edwards PN, Bowker GC, Jackson SJ, Williams R. 2009. Introduction: an agenda for infrastructure studies. *Journal of the Association for Information Systems* **10(5)**:364–374.

Edwards PN, Jackson SJ, Chalmers MK, Bowker GC, Borgman CL, Ribes D, Burton M, Calvert S. 2013. *Knowledge infrastructures: intellectual frameworks and research challenges*. Ann Arbor: University of Michigan, 40.

Edwards PN, Mayernik MS, Batcheller AL, Bowker GC, Borgman CL. 2011. Science friction: data, metadata, and collaboration. *Social Studies of Science* **41(5)**:667–690 DOI 10.1177/0306312711413314.

European Space Agency. 2016. Gaia. *Available at http://sci.esa.int/gaia/* (accessed on 23 July 2016).

Hagen J. 2003. The statistical frame of mind in systematic biology from quantitative zoology to biometry. *Journal of the History of Biology* **36(2)**:353–384 DOI 10.1023/A:1024479322226.

Hammersley M, Atkinson P. 2007. *Ethnography: principles in practice*. 3rd edition (reprinted). London: Routledge.

Hey AJG, Trefethen A. 2003. The data deluge: an e-science perspective. In: Berman F, Fox G, Hey AJG, eds. *Grid computing: making the global infrastructure a reality*. West Sussex: Wiley, 809–824.

Integrated Ocean Drilling Program (IODP) Planning Sub Committee (IPSC) Scientific Planning Working Group. 2001. *Earth, oceans and life: IODP initial science plan*. Washington, D.C.: International Working Group Support Office.

IODP. 2014. International Ocean Discovery Program. *Available at http://iodp.org* (accessed on 13 June 2014).

Jackson SJ, Buyuktur A. 2014. Who killed waters? Mess, method, and forensic explanation in the making and unmaking of large-scale science networks. *Science, Technology & Human Value* **39(2)**:285–308 DOI 10.1177/0162243913516013.

Kay LE. 2000. *Who wrote the book of life? A history of the genetic code*. Stanford: Stanford University Press.

Keller EF. 1984. *A feeling for the organism, 10th aniversary edition: the life and work of Barbara McClintock*. London: Macmillan.

Kelty CM. 2012. This is not an article: model organism newsletters and the question of open science. *BioSocieties* **7(2)**:14–168 DOI 10.1057/biosoc.2012.

Kitchin R. 2014. Big data, new epistemologies and paradigm shifts. *Big Data & Society* **1**(**1**):2053951714528481 DOI 10.1177/205395171452848.

Kitchin R, Lauriault TP. 2014. Small data, data infrastructures and big data. *SSRN Electronic Journal* DOI 10.2139/ssrn.237614.

Kragh H. 1996. *Cosmology and controversy: the historical development of two theories of the universe*. Princeton: Princeton University Press.

Lenoir T. 1999. Shaping biomedicine as an information science. In: Mary EB, Trudi BH, Robert VW, eds. *Proceedings of the 1998 conference on the history and heritage of science information systems*. Medford: Information Today, Inc, 27–45.

Lenoir T, Hays M. 2000. Manhattan project for biomedicine. In: *Controlling our destinies: historical, philosophical, ethical, and theological perspectives on the human genome project*. South Bend Indiana: University of Notre Dame Press, 19–26.

Leonelli S. 2010. Documenting the emergence of bio-ontologies: or, why researching bioinformatics requires HPSSB. *History and Philosophy of the Life Sciences* **32**(**1**):105–125.

Leonelli S. 2012. When humans are the exception: cross-species databases at the interface of biological and clinical research. *Social Studies of Science* **42**(**2**):214–236 DOI 10.1177/030631271143626.

Leonelli S. 2014. What difference does quantity make? On the epistemology of Big Data in biology. *Big Data & Society* **1**(**1**):2053951714534395 DOI 10.1177/2053951714534395.

Leonelli S, Ankeny RA. 2012. Re-thinking organisms: the impact of databases on model organism biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* **43**(**1**):29–36 DOI 10.1016/j.shpsc.2011.10.003.

LSST Science Collaboration, Abell PA, Allison J, Anderson SF, Andrew JR, Angel JRP, Armus L, Arnett D, Asztalos SJ, Axelrod TS, Bailey S, Ballantyne DR, Bankert JR, Barkhouse WA, Barr JD, Barrientos LF, Barth AJ, Bartlett JG, Becker AC, Becla J, Beers TC, Bernstein JP, Biswas R, Blanton MR, Bloom JS, Bochanski JJ, Boeshaar P, Borne KD, Bradac M, Brandt WN, Bridge CR, Brown ME, Brunner RJ, Bullock JS, Burgasser AJ, Burge JH, Burke DL, Cargile PA, Chandrasekharan S, Chartas G, Chesley SR, Chu Y-H, Cinabro D, Claire MW, Claver CF, Clowe D, Connolly AJ, Cook KH, Cooke J, Cooray A, Covey KR, Culliton CS, De Jong R, De Vries WH, Debattista VP, Delgado F, Dell'Antonio IP, Dhital S, Di Stefano R, Dickinson M, Dilday B, Djorgovski SG, Dobler G, Donalek C, Dubois-Felsmann G, Durech J, Eliasdottir A, Eracleous M, Eyer L, Falco EE, Fan X, Fassnacht CD, Ferguson HC, Fernandez YR, Fields BD, Finkbeiner D, Figueroa EE, Fox DB, Francke H, Frank JS, Frieman J, Fromenteau S, Furqan M, Galaz G, Gal-Yam A, Garnavich P, Gawiser E, Geary J, Gee P, Gibson RR, Gilmore K, Grace EA, Green RF, Gressler WJ, Grillmair CJ, Habib S, Haggerty JS, Hamuy M, Harris AW, Hawley SL, Heavens AF, Hebb L, Henry TJ, Hileman E, Hilton EJ, Hoadley K, Holberg JB, Holman MJ, Howell SB, Infante L, Ivezic Z, Jacoby SH, Jain B, Jedicke R, Jee MJ, Jernigan JG, Jha SW, Johnston KV, Jones RL, Juric M, Kaasalainen M,

Kafka SS, Kahn SM, Kaib NA, Kalirai J, Kantor J, Kasliwal MM, Keeton CR, Kessler R, Knezevic Z, Kowalski A, Krabbendam VL, Krughoff KS, Kulkarni S, Kuhlman S, Lacy M, Lepine S, Liang M, Lien A, Lira P, Long KS, Lorenz S, Lotz JM, Lupton RH, Lutz J, Macri LM, Mahabal AA, Mandelbaum R, Marshall P, May M, McGehee PM, Meadows BT, Meert A, Milani A, Miller CJ, Miller M, Mills D, Minniti D, Monet D, Mukadam AS, Nakar E, Neill DR, Newman JA, Nikolaev S, Nordby M, O'Connor P, Oguri M, Oliver J, Olivier SS, Olsen JK, Olsen K, Olszewski EW, Oluseyi H, Padilla ND, Parker A, Pepper J, Peterson JR, Petry C, Pinto PA, Pizagno JL, Popescu B, Prsa A, Radcka V, Raddick MJ, Rasmussen A, Rau A, Rho J, Rhoads JE, Richards GT, Ridgway ST, Robertson BE, Roskar R, Saha A, Sarajedini A, Scannapieco E, Schalk T, Schindler R, Schmidt S, Schmidt S, Schneider DP, Zhan H, et al. 2009. *LSST Science Book, Version 2.0*. ArXiv preprint. arXiv:0912.0201.

Meyer ET. 2009. *Moving from small science to big science: social and organizational impediments to large scale data sharing (SSRN Scholarly Paper No. ID 2166245)*. Rochester: Social Science Research Network.

Mukerji C. 2014. *A fragile power: scientists and the state*. Princeton: Princeton University Press.

National Science Foundation. 2010. *NSF data management plans*. Washington, D.C.: National Science Foundation.

NOAA. 2016. National Oceanic and Atmospheric Administration. *Available at http://www.noaa.gov* (accessed on 23 July 2015).

O'Donoghue T, Punch K (eds.) 2004. *Qualitative educational research in action: doing and reflecting*. Abingdon: RoutledgeFalmer.

Orcutt BN, LaRowe DE, Biddle JF, Colwell FS, Glazer BT, Reese BK, Kirkpatrick JB, Lapham LL, Mills HJ, Sylvan JB, Wankel SD, Wheat CG. 2013. Microbial activity in the marine deep biosphere: progress and prospects. *Frontiers in Microbiology* **4**: Article 189 DOI 10.3389/fmicb.2013.00189.

Ortega C. 2003. ARGOS capabilities for global ocean monitoring. In: Dahlin H, Flemming NC, Nittis K, Petersson SE, eds. *Building the European capacity in operational oceanography: Proceedings of the third international conference on EuroGOOS*. Amsterdam: Elsevier, 317–324.

Paul NW. 2009. Rationalitäten der Wissenproduktion: Über Transformationen von Gegenständen, Technologien und Information in Biomedizin und Lebenswissenschaften. *Berichte Zur Wissenschaftsgeschichte* **32(3)**:230–245 DOI 10.1002/bewi.200901351.

Porter TM. 1996. *Trust in numbers: the pursuit of objectivity in science and public life*. Princeton: Princeton University Press.

Ribes D. 2014. Ethnography of scaling, or, how to a fit a national research infrastructure in the room. In: *Proceedings of the 17th ACM conference on computer supported cooperative work & social computing*. New York: ACM, 158–170. *Available at http://dl.acm.org/citation.cfm?id=2531624*.

Ribes D, Bowker GC. 2008. Organizing for multidisciplinary collaboration: the case of the geosciences network. In: Olson GM, Zimmerman AS, Bos N, eds. *Science on the Internet*. Cambridge: MIT Press, 311–330.

Ribes D, Finholt TA. 2008. Representing community: knowing users in the face of changing constituencies. In: *Proceedings of the 2008 ACM conference on computer supported cooperative work*. New York: ACM, 107–116.

Ribes D, Polk JB. 2015. Organizing for ontological change: the kernel of an AIDS research infrastructure. *Social Studies of Science* **45(2)**:214–241 DOI 10.1177/0306312714558136.

Sawyer S. 2008. Data wealth, data poverty, science and cyberinfrastructure. *Prometheus* **26(4)**:355–371 DOI 10.1080/08109020802459348.

Sloan Digital Sky Survey: Home. 2016. *Available at http://www.sdss.org/* (accessed on 08 November 2015).

Southern California Earthquake Center. 2016. Southern California Earthquake Center | Studying earthquakes and their effects in California and beyond. *Available at https://www.scec.org/* (accessed on 23 July 2016).

Star SL. 1999. The ethnography of infrastructure. *American Behavioral Scientist* **43(3)**:377–391 DOI 10.1177/00027649921955326.

Star SL, Ruhleder K. 1996. Steps toward an ecology of infrastructure: design and access for large information spaces. *Information Systems Research* **7(1)**:111–134 DOI 10.1287/isre.7.1.111.

Strauss MA. 2014. Mapping the Universe: surveys of the sky as discovery engines in astronomy. *Daedalus* **143(4)**:93–102 DOI 10.1162/DAED_a_00309.

Suárez-Díaz E, Anaya-Munoz VH. 2008. History, objectivity, and the construction of molecular phylogenies. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Science* **39(4)**:451–468.

Teske A, Biddle JF, Schrenk M. 2011. *2011 workshop: deep biosphere sediment microbiology. Available at http://www.darkenergybiosphere.org/RCN/meetings/2011meeting/docs/2011DEBI_Sedi mentMeetingSummary2.pdf* .

UNOLS. 2015. University-National Oceanographic Laboratory Services. *Available at https://www.unols.org/* (accessed on 16 July 2015).

Whitman WB, Coleman DC, Wiebe WJ. 1998. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America* **95(12)**:65786583.