Int J Digit Libr (2015) 16:61–77
DOI 10.1007/s00799-015-0137-3

What lies beneath?: Knowledge infrastructures in the subseafloor biosphere and beyond

Peter T. Darch · Christine L. Borgman · Sharon Traweek · Rebekah L. Cummings · Jillian C. Wallis · Ashley E. Sands

Received: 19 January 2014 / Revised: 10 October 2014 / Accepted: 6 January 2015 / Published online: 15 February 2015

Abstract We present preliminary findings from a three-year research project comprised of longitudinal qualitative case studies of data practices in four large, distributed, highly multidisciplinary scientific collaborations. This project follows a 2×2 research design: two of the collaborations are big science while two are little science, two have completed data collection activities while two are ramping up data collection. This paper is centered on one of these collaborations, a project bringing together scientists to study subseafloor microbial life. This collaboration is little science, characterized by small teams, using small amounts of data, to address specific questions. Our case study employs participant observation in a laboratory, interviews ($n = 49$ to date) with scientists in the collaboration, and document analysis. We present a data workflow that is typical for many of the scientists working in the observed laboratory. In particular, we show that, although this workflow results in datasets apparently similar in form, nevertheless a large degree of heterogeneity exists across scientists in this laboratory in terms of the methods they employ to produce these datasets—even between scientists working on adjacent benches. To date, most studies of data in little science focus on heterogeneity in terms of the types of data produced: this paper adds another dimension of heterogeneity to existing knowledge about data in little science. This additional dimension makes more complex the task of management and curation of data for subsequent reuse. Furthermore, the nature of the factors that contribute to heterogeneity of methods suggest that this dimension of heterogeneity is a persistent and unavoidable feature of little science.

Keywords Data deluge · Big science · Little science · Multidisciplinary scholarship · Knowledge infrastructures

P. T. Darch (✉) · C. L. Borgman · S. Traweek · R. L. Cummings · J. C. Wallis · A. E. Sands
Knowledge Infrastructures Project, Department of Information Studies, UCLA, GSE&IS Building, Room 210, Box 951520, Los Angeles, CA 90095-1520, USA
e-mail: petertdarch@ucla.edu

1 Introduction

Long predicted by the science community [30], both *Nature* and *Science* have now heralded the opportunities and challenges presented by the scientific data deluge [19,20]. Universities themselves are assessing their rights, roles, and responsibilities for managing and for exploiting data from their researchers [4].

In addition to the sheer size of data generated, the heterogeneity of datasets is also increasing, even within individual domains. Scientific collaboration is becoming a more multidisciplinary, distributed endeavor [15]. As a result, approaches from multiple epistemological or social perspectives may be combined in the production of a dataset, and conversely a single dataset may be used in multiple contexts, crossing epistemological, cultural and social boundaries.

Contemporary digital scholarship is thus a rapidly changing and expanding undertaking. However, today's scientific methods and organization of collaborative work often do not scale well to today's volumes or diversity of data generated; qualitatively different approaches to scientific inquiry are required. As data are combined from multiple sources and are mined for new interpretations, the challenges of data management and curation multiply. Modern sensor networks, satellites, telescopes, and laboratory instruments can collect vastly more data, at far faster rates and far greater variety, than ever before. Scientists rely on their instruments, algorithms, and collaborators to clean, verify, visualize, and inter-

pret their data. Much can go wrong in the many steps involved in the design and deployment of instruments, collection and cleaning of data, and in the analysis and reporting of results. Data and responsibility pass through many hands, often over the course of many years, in the life cycles of collaborative data-driven science.

Scientific data management requires deep expertise in scientific theory, method, instrumentation, and interpretation. Skill sets are complex and are divided differently in each field and specialty. Each step in data handling requires knowledge and judgment of the steps that went before. Necessary details of data provenance often go undocumented, leaving researchers in the position of making multiparty inferences with insufficient information [42]. Minute differences in calibration, miniscule artifacts in a data stream, and other perturbations may be spotted by those closest to the research design—but these factors decrease in visibility the farther the interpreter lies from the source of the data.

The pressure from funding agencies such as the National Science Foundation (NSF) and the National Institute of Health (NIH) to share research data highlights the complexity of data-driven science. "Data" is a contested notion. Furthermore, competing views exist of research, innovation, and scholarship, disparate incentives for collecting and releasing data, the economics and intellectual property of research products, and public policy—and the requisite technical and human infrastructure. However, relatively few studies document consistent data release. Sharing research data is thus a conundrum—"an intricate and difficult problem" [11].

Research in both data-intensive *big science*, where data products are large in volume but typically homogeneous, such as astronomy projects centered on the building and operation of massive-scale instruments, and in *little science*, where data products are small in size but large in number, such as sensor network applications in ecology, marine biology, environmental engineering, and seismology, reveals a critical lack of infrastructure to support these new forms of scholarship. The promise of technology-enabled, data-driven digital scholarship in science is predicated upon available systems, services, tools, content, policies, practices, and human resources to discover, mine, and use research products, as well as to create those products in forms that are useful to others. Not only is this infrastructure not yet in place, it is not yet clear what should be built or how to build it [10,25].

However, this problem is becoming better recognized, as is the fact that sociotechnical research approaches can produce critical insights that inform the design, policy, and human resource requirements for scientific information infrastructure [25]. This paper presents preliminary findings from a three-year study of such infrastructures. *The Transformation of Knowledge, Culture, and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective* (henceforth known as the *Knowledge Infrastructures* project) involves longitudinal qualitative case studies of four large, distributed, multidisciplinary scientific collaborations. Two of these collaborations could be considered as big science, whereas the other two involve multiple research teams performing little science. Furthermore, two of these collaborations have been in the process of *ramping down* data collection and active research, while the two others are *ramping up* their activities. Beyond the preliminary results in this paper, the Knowledge Infrastructures project will continue to analyze these four distinct collaborations.

Here, we present an agenda for researching *knowledge infrastructures*, defined as "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds" [23, p. 17], through motivating and presenting the research questions that guide our Knowledge Infrastructures project team. Then, we present the methodologies of the Knowledge Infrastructures project, introduce the four case studies, and explain how our approach to research will make significant contributions in pursuing the research agenda. To demonstrate these contributions, we present preliminary findings from one of the case studies, a multidisciplinary, multi-institutional collaboration that involves studying microbial life beneath the seafloor. In particular, we explore and account for the diversity of practices in producing datasets that we observed across scientists, even in the same laboratory. To date, research on data practices has focused on heterogeneity of data types produced in a research setting; by contrast, here we demonstrate that for a single data type, there can be significant heterogeneity in how such datasets are produced even in the same research setting. This heterogeneity can multiply significantly the challenges involved in data management and curation.

2 A project for researching sociotechnical knowledge infrastructures

Our Knowledge Infrastructures project responds both to the needs of scientists in developing practices and infrastructures to manage the increasing volume and diversity of data in their work, and to substantial gaps in the existing social scientific literature that addresses scientists' data practices. In this section, we motivate the Knowledge Infrastructures project, introduce its features, and discuss how it will contribute to improved understandings of how scientists produce and manage their data.

2.1 Motivation for the Knowledge Infrastructures project

Research into knowledge infrastructures is a developing field. Current discourse around research data highlights the lack of institutional infrastructure comparable to the roles that libraries and publishers serve in access to scholarly publica-

tions. Infrastructure for research data is much more than disseminating resources; it must support data collection, analysis, use, and reuse for new scientific methods, and should democratize access in the process. The design of knowledge infrastructures rests on the ability to explicate the sociotechnical structures that are embodied in the data, in data practices, in technical arrangements, and in policies. These interdependencies are known to present significant risks to adoption and implementation of effective infrastructure [1,5].

The role of data as a scholarly research product is a growing concern, both practically and politically [4,44]. Our research recognizes the significant technical challenges that arise in managing research data, such as data granularity, data provenance, data structures, definitions of data, dataset identity, identifiers, and functions of data [2,6,29]. Some researchers have studied how authors of scientific journal articles cite reused datasets originally generated by other researchers, or develop more systematic ways and standards for citing these datasets [18,51]. However, reuse analysis covers only one part of the multistage data life cycle [53].

While countless policy reports call for the building of infrastructure and capacity for research data, only a handful of researchers consider how knowledge of data practices might inform design and policy processes [3,23,26,46,47,53]. Studies of data practices draw upon a larger body of work than can be enumerated here. Leigh Star and Karen Ruhleder were the first to assess infrastructure from a sociotechnical perspective [49], opening up a rich area to be mined by many others [9,24]. Included in studies of knowledge infrastructures are research on work practices, collaborations, virtual organizations, computer-supported collaborative work, project life cycles, and project time [15,34,37,39,45,50].

2.2 Research questions for the Knowledge Infrastructures project

In the Knowledge Infrastructures project, we address general research questions across the four research sites:

1. What new infrastructures, divisions of labor, knowledge, and expertise are required for data-driven science?
2. How are the infrastructures of multidisciplinary, data-driven scientific collaborations established and how are they dismantled?
3. How do data management, curation, sharing, and reuse practices vary among research areas?
4. What data are most important to curate, from whose perspective, and who decides?

2.3 Knowledge Infrastructures project case studies and methods

To address these research questions, we are conducting case studies of four large, distributed, collaborative, multidisciplinary projects. We selected case studies for a 2 × 2 research design (see Table 1), enabling the comparison of two research projects that produce large volumes of homogeneous data (in this case, the *Sloan Digital Sky Survey, or SDSS*, and the *Large Synoptic Survey Telescope*, or *LSST*) with two projects that produce smaller amounts of heterogeneous data (in this case, the *Center for Embedded Network Sensing*, or *CENS*, and the *Center for Dark Energy Biosphere Investigations, or C-DEBI*). It also allows us to compare projects that are in the earlier stages of their life cycles (LSST and C-DEBI) and are ramping up data production, and projects at later stages of their life cycles (SDSS and CENS) that have ramped down data production

Table 1 Site comparisons by data scope and by life cycle stage

	Big science	Little science
Ramping up data collection	LSST	C-DEBI
Ramping down data collection	SDSS	CENS

Comparisons of these sites enable us to assess the knowledge infrastructure requirements for a broad spectrum of scientific research and practice. These sites also allow us to understand processes of knowledge transfer, such as that between scientists, between scientists and information professionals, between research projects, and between science projects and the public. We are able to identify infrastructure practices that contribute to better strategies for data management and to make recommendations for policy and practice.

Research on the two ramping down projects, CENS and SDSS, began prior to the Knowledge Infrastructures project. Research on C-DEBI began in 2012, and on LSST in 2014. The CENS project involved the development of sensing technology in collaboration with teams from a variety of sciences, most notably ecology, marine biology, and seismology. CENS embodied little science, and the data tended to be heterogeneous and complex. CENS focused much less explicitly on transferring its data as part of its legacy. Some members of the Knowledge Infrastructures team were embedded in CENS for a decade both as observers and participants in the development of knowledge infrastructure. This infrastructure included: the CENS Deployment Center, to plan data collection campaigns and serve as reference metadata after the fact; the CENS publication repository; and the CENS data registry as part of an annual reporting system. Developing these systems also enabled us to identify good practice associated with multidisciplinary data management.

SDSS was a highly visible project in the domain of astronomy, embodying big science. Much is being learned about its practices, policies, successes and failures, and transfer of expertise to other sciences and projects. SDSS, on the surface, may appear to have exemplified a solved data curation environment. However, our closer inspection of SDSS is

revealing social and technical architectures contingent upon changing technologies.

As with CENS, C-DEBI also embodies little science. C-DEBI studies microbial subseafloor life. It was launched in late 2010, and affords opportunities to observe how the work of negotiating, challenging, building, and maintaining data management practices unfolds in a new collaborative setting. We are learning what partners in information handling they seek, at what stages, and how they compare to the lessons of CENS.

LSST, like SDSS, is in the field of astronomy and will produce data that are largely homogeneous and very large in scale. Although the telescope is not due to launch operations until 2022, the collaborative team involved is being assembled, and is already making critical design decisions about both the technology for data collection and the infrastructure for data management. LSST team members are building on the experiences of those involved in the SDSS and other large-scale astronomy projects.

2.4 A theoretical framework for analyzing knowledge infrastructures

Our study examines how scientists and engineers use both social and technical resources to accomplish their goals, taking into account individuals, groups, collaborations, organizations, ideas, techniques, and technologies. A large body of scholarship investigates relationships between the technical and the social. Many studies of technological and social change have accounted for the latter in terms of the former, seeing technological change as the driving force behind reconfigurations of social relations; this argument has been labeled *technological determinism* [32].

Scholars of technology became dissatisfied with the technological determinism approach because it does not explain why some technologies achieve wide acceptance, while other similar technologies fail to do so [21]. Some scholars have developed a *social construction of technology* (*SCOT*) approach that explored how the development and uses of technologies were shaped by social process, including how actors use and shape technologies in pursuit of specific social interests [7].

A subsequent development of SCOT was *actor-network theory (ANT)*, which focused on goals, agency, and interaction in knowledge making [16,38]. The underlying idea of ANT is that all actors, human and nonhuman, pursue interests or are goal-directed, and thus build networks of social and material resources to pursue these goals. ANT has been widely and successfully used in analyses of science, technology, and society.

In our analyses, scientists and engineers in a laboratory draw on their current networks of resources (social and material) to accomplish each step of the workflow they have established. They regard each step as necessary for answering their own research questions and those of the larger collaboration, leading to the accumulation of more resources: recognition and credit for the lab and its members in the form of publications, funding, and promotions. Furthermore, their networks of resources enable them to reevaluate and improve these processes [39].

These research networks include the expertise acquired during the members' education and experience at previous sites. They also include the multiple techniques and technologies (sample and data collection strategies, research equipment, protocols, handbooks, journal articles, funding) accessible from both within and beyond the laboratory. It is only when the scientists and engineers become aware of, and competent in, engaging with certain techniques and technologies and their affordances that they become part of the laboratory's network. Similarly, those technologies and techniques might be part of one network, but unknown in another. These networks are dynamic and might even change rapidly, as equipment, techniques, knowledge, personnel, and funding are introduced to, or removed from, the laboratory. The knowledge of how and when to access these network resources is also in flux.

Applying this theoretical framework to the research carried out in these laboratories not only helps us to perceive the heterogeneity of data practices observed across these sites in answering the similar research questions and producing datasets of similar form, but also to account for why this heterogeneity—even between researchers working on adjacent benches—occurs.

2.5 The Center for Dark Energy Biosphere Investigations

The Center for Dark Energy Biosphere Investigations (C-DEBI) is an NSF *Science and Technology Center* (STC) launched in September 2010. C-DEBI brings together scientists from the biological, chemical, and physical sciences to study subseafloor microbial life, in particular to study interactions between the composition of microbial communities and the physical environment they inhabit [22].

C-DEBI serves as an important case study for studying contemporary developments in digital scholarship. The project is massively distributed across institutions in the USA and Europe, and very highly multidisciplinary. As such, it is an exemplar of the complexity of data-driven science. It is also a complement to CENS, as is explicated in Subsect. 2.5.

2.5.1 Organization and work of C-DEBI

Scientists involved with C-DEBI work toward the project's scientific goals through the collection and analysis of physical samples, such as rocks from the seafloor (known as *cores*). Fundamental to scientists' work is the production, analysis

and correlation of data about the cores' microbial communities with the physical properties (such as geochemical or hydrological) of these samples.

The data life cycle may start in a number of contexts. One particularly important context is scientific ocean drilling cruises. During our period of observation, these were often conducted by the *Integrated Ocean Drilling Program (IODP)* which ran from 2003–2013 (it should be noted that the IODP was replaced with a new drilling program in 2013, namely the *International Ocean Discovery Program*, also known as IODP. For the purposes of this paper, the acronym "IODP" will be used to refer to the Integrated Ocean Drilling Program throughout). IODP organized regular research cruises bringing together scientists from a broad range of disciplines and institutions to visit a specific site to collect cores, which are subsequently analyzed both onboard the ship and in scientists' laboratories at their home research institutions. C-DEBI also supports scientists who participate in research cruises organized by organizations other than the IODP.

As well as providing some funding and equipment support for cruises, C-DEBI also distributes funding directly to scientists. This funding is generally characterized by being short term (typically one to three years in length), to individuals and small teams (usually of two or three) and across a very broad range of institutions. The main opportunity for funding is through the *Small Grants* program, through which grants are awarded to proposed projects that use existing datasets and samples (e.g., from cruises). Other grants are directed toward early career researchers, such as doctoral students and postdoctoral researchers. These grants are awarded on a regular basis following competitive calls for proposals. To date, these grants have supported approximately 90 scientists in more than 30 laboratories across the USA, Europe and East Asia [17]

2.5.2 C-DEBI infrastructure

C-DEBI has also implemented other measures to support the community of researchers, fostering connections and exchanges of knowledge. One important component of bringing the community together is the project website, which contains a wide range of C-DEBI-related information, including key project personnel, descriptions of the main scientific foci of the project, information about the various grants and fellowships, a list of C-DEBI-contributed scientific publications, and C-DEBI official documents such as the Proposal and Annual Reports. The project also communicates with community members, and any others who are interested, via a twice-monthly newsletter. Finally, the project also provides opportunities for affiliated scientists to come together, such as an annual project meeting.

2.6 C-DEBI and CENS

We are comparing data management, curation, and sharing practices across the four case studies in our 2 × 2 research design. The richest source of comparisons and contrasts for C-DEBI is with CENS. There are many similarities and differences between C-DEBI and CENS, which will extend and add to the extensive body of work we have already produced about our studies of CENS.

As with C-DEBI, CENS was an NSF STC. It was launched in 2002, and ceased operation in 2012. CENS was a distributed, multidisciplinary collaboration involving five research universities across California. Its focus was to bring technologists and domain scientists (terrestrial ecology, marine biology, environmental engineering, seismology, plus applications in urban settings and arts) together so that the technologists could develop networked sensing tools that would allow domain scientists to collect data at higher spatial and temporal resolution. Like C-DEBI, CENS was a federation of a number of small teams of technologists and scientists working together on such projects, funded by a mixture of internal and external grants.

In common with C-DEBI, CENS was little science, in the sense that it involved the generation of a large number of heterogeneous, small-scale datasets meant for consumption by those that generated these data [13]. However, C-DEBI differs from CENS in important ways. One is that CENS focused on developing emergent technologies to support scientific work, whereas C-DEBI foregrounds studying emergent scientific problems. Another is that C-DEBI involves the integration of samples and data produced in a domain (IODP cruises) that shares many features with big science with work of small, multidisciplinary teams in individual laboratories. How the similarities and differences between C-DEBI and CENS will augment our findings from CENS is explicated below.

2.6.1 Lack of shared interests across a project team

One key finding from our studies of CENS is the lack of shared interests that existed within individual project teams. Technologists were interested mainly in accomplishing the task of building networks of sensors that were technologically novel. Thus, technologists were primarily interested in data about how the technology operated. They were much less interested in the scientific data per se, regarding it as background context. The converse was true in the case of the domain scientists [14].

One implication of this finding is that the technologists would take data about how the sensors operated with them to their laboratories and manage them according to their own particular practices and standards, while the domain scientists would do the same with scientific data, sometimes dis-

carding data that were no longer in use. The separation of the data sources made it difficult to reproduce results. Although the technologists and the scientists were interdependent, the data practices did not support this interdependence.

Similar challenges are appearing in C-DEBI. C-DEBI produces both biological and physical data. However, where different types of data are produced by different scientists and managed in different contexts, there can be significant implications for the interoperability of these data. Furthermore, subsequent storage and curation can diverge due to cultural practices or formal requirements in different disciplines. These choices can have implications for subsequent reproduction and verification of scientific analyses.

2.6.2 Trust in data

Trust in data is essential for their use and reuse in the scientific endeavor. Our research on CENS found that scientists' ability to assess the integrity of data was essential for reuse. This ability depended on the knowledge the scientist possess of stages of the data life cycle—from research design to data storage and curation [54]. The life cycle of CENS data involved many steps, each dependent on preceding steps: the effect of decisions made at each step was cumulative throughout the life cycle [53].

Furthermore, a great deal of confusion and disagreement occurred amongst domain scientists and technologists about who was responsible for different types of data, and for different stages of the data life cycle for each type of data. Questions of who owned different types of data were frequently unresolved because some types of data or metadata did not implicate the interests of either the scientists or the technologists, and were thus frequently neglected by both [52].

Issues about who is responsible for certain types of data also arise in C-DEBI. For instance, the interests of very few scientists involved in C-DEBI projects seem to be implicated in the tools that support C-DEBI-related work. Information about these tools is important for the subsequent interpretation and reuse of C-DEBI-generated scientific data but it is difficult to see whose interests are served by collection, storage and curation of such information.

2.6.3 Successful data sharing

Enabling the widespread sharing of data promises many benefits for science [11]. The first step in facilitating data sharing is to ensure effective data management practices at every stage of the data life cycle. However, our research on CENS also exposes a number of other issues that complicate the sharing of data. CENS researchers were generally willing to share data, subject to a number of conditions: they were more willing to share data that they have already published, and are also more likely to be willing to share data that involved less effort to collect [12]. Other conditions included ensuring that the producer of the data received proper attribution, and that the amount of effort to share data was not burdensome. Given these conditions, and that few repositories for CENS data actually existed, data sharing was very rare across CENS [55].

The C-DEBI case study provides an ideal opportunity for us to extend these findings because the observation of many different types of interactions enables a better understanding of the particular contexts in which data sharing is more and less likely to take place. We are conducting analyses of the interplay of various technological, infrastructural, social and normative factors that facilitate data sharing.

2.6.4 Big science meets little science

Another point of comparison between C-DEBI and CENS is that while all the data produced and used by CENS researchers were characteristic of little science, C-DEBI also involves data that are produced in a context, namely the IODP, that shares many features with big science. The data generated on IODP expeditions about the physical properties of cores are highly structured, professionally curated according to stringent standards, and are archived in publicly accessible databases in the long term.

The case study of C-DEBI offers the possibility to see the interactions between the IODP standards and the day-to-day data practices of researchers. The addition of IODP to the data life cycle can complicate many of the factors outlined above. For instance, the involvement of an additional organization can introduce additional divergent interests to C-DEBI scientists. Adding more steps to the data life cycle can impact subsequent stages of the life cycle, contributing to the complexity of the tasks facing scientists as they judge the integrity of datasets and attempt to interpret them.

3 C-DEBI case study

Above, we presented C-DEBI as a research site, explaining how it is an important exemplar of contemporary developments in scientific digital scholarship and thus provides an excellent case study for understanding data practices in a little science project both in its own right and in comparison with the other case studies that comprise our Knowledge Infrastructures project. Here, we present findings from the first year of our case study.

3.1 Methods

We are conducting a longitudinal ethnographic case study of C-DEBI. An ethnographic study involves a range of qualitative research methods to provide a thorough account of the

organization under study [28]. Our methods include interviews, participant observation, online ethnography, and document analysis. Drawing on data from a range of sources allows for triangulation [43]. The content of texts (including interview transcripts, reports, and ethnographic notes) is highly contingent, rather than simply reflecting an underlying reality. Triangulation involves the crosschecking of data from different sources produced in different contexts, which helps to ensure that conclusions drawn from the data are not biased by the context in which the data are produced.

3.1.1 Participant observation

A key feature of this case study is long-term participant observation of C-DEBI. Our observations include being embedded for eight months in a laboratory headed by a leading figure in C-DEBI at a large US research university, observing scientists at work and in meetings. We also attend scientific meetings (conferences, workshops, seminars, and colloquia) of both C-DEBI and the broader scientific communities in which it is embedded. Participant observation has been successfully applied to the study of scientists and their practices since the 1970s [36,39,41,50], and has latterly been used in studies of geographically distributed, multidisciplinary collaborations [31]. Participant observation is particularly suitable for this case study as it affords a detailed understanding of the local, disciplinary, and institutional contexts in which scientists are working as well as relationships and networks amongst scientists. We have been able to observe how ideas, practices, and methods are communicated between collaborators.

3.1.2 Interviews

Our current interview sample comprises 49 people, including C-DEBI-affiliated scientists and scientists, curators, and managerial staff involved in related activities such as the scientific ocean research cruises (IODP). Our sample is detailed in Table 2, which distinguishes between respondents involved in C-DEBI and those working for IODP. The C-DEBI sample is broken down further by geographic location (USA or not), and career stage. The column "Involved with IODP" indicates which interviewees are involved in policy- or decision-making in the IODP. The IODP interviewees are further split into two groups: those in cruise operations, and those with the *Consortium for Ocean Leadership*, which was responsible for administering US involvement in the IODP.

C-DEBI-affiliated scientists are based in research institutions and laboratories across the USA, at a wide range of career stages, from a variety of disciplinary backgrounds, and are working on a range of projects. Potential interviewees were identified based on whether they had been observed at work in the laboratory, and their involvement in C-DEBI-funded projects or IODP operations. Recommendations were also sought from interviewees.

Table 2 The composition of our interview sample

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

Interviews ranged in length from 35 min to two hours and 30 min, with the majority being between one and two hours long. The scientists interviewed were questioned not only about their data practices and day-to-day scientific work but also about their academic and professional backgrounds, enabling us to understand how the scientists' multidisciplinary backgrounds impact observed data practices. The non-scientists interviewed were asked about their work within the C-DEBI project, including the building, implementation, and maintenance of C-DEBI infrastructure and policies.

3.1.3 Document analysis

We assembled a corpus of of documents for analysis. Some documents help to explain the work conducted by C-DEBI-affiliated scientists in their laboratories, for example scientific journal articles, instruction manuals for laboratory equipment, and published protocols for techniques observed in the laboratory. Other documents help us to interpret social contexts in which C-DEBI scientists operate. These include official C-DEBI documents such as the initial proposal, Annual Reports to the NSF, operating documents (e.g., the Strategic Implementation Plan), and calls for funding. Finally, we collected other documents to understand better

the broader contexts in which the C-DEBI project is operating, including NSF and IODP.

3.1.4 Data analysis

We analyzed data using a grounded theory approach [27]. We read interview transcripts and other documents closely, and a number of themes emerged. We then coded the documents according to these themes. Adopting a grounded theory approach meant that the findings in this paper are data-driven, in the sense that they emerge from the empirical research rather than being imposed upon the data in a top-down fashion.

3.2 Introducing the Jones laboratory

The majority of the participant observation has so far taken place in a single laboratory, the *Jones laboratory*, at a large research university in the USA (n.b. the name "Jones" and the names of the individual scientists are pseudonyms). The head of the laboratory is a senior figure in the leadership of C-DEBI, and the focus of the laboratory's work is on interactions between microbial life and physical processes in the deep subseafloor. Work in the laboratory is funded largely by the NSF, both directly and through IODP and C-DEBI.

The overall composition of the research group in this laboratory often changes, due to new PhD students and postdoctoral researchers joining the group, and others completing their doctorates or postdoctoral research projects and moving on to other laboratories or industry. During the period of observation, the laboratory personnel has comprised a tenured Professor who was the laboratory's leader, four postdoctoral researchers with between zero and five years' experience in the laboratory, six PhD students ranging from first year to fifth year, one visiting PhD student from another laboratory in the USA, one undergraduate student, and two short-term international research visitors.

3.3 A typical workflow in the laboratory

Here, we present a standard workflow within the laboratory. This workflow is a composite of many observed workflows. In particular, although the form of the resultant datasets appears similar across scientists in this laboratory, a high degree of heterogeneity nevertheless exists across the laboratory regarding the tools and methods used to produce these datasets. We account for this heterogeneity, discussing how different scientists—even those working on adjacent benches—have access to different and constantly changing configurations of social, material, and scientific resources to help them accomplish the different steps of the workflow.

We first set the scene by presenting the basic steps of this workflow (3.3.1). Then, in each of Sects. 3.3.2, 3.3.3 and 3.3.4, a single step in the workflow is examined in more depth. In particular, for each step, we compare the differing methods employed by two scientists and examine why these methods are used.

3.3.1 A biological workflow in the Jones laboratory

The workflow is outlined in Fig. 1. The central goal of a project incorporating this workflow is to understand the

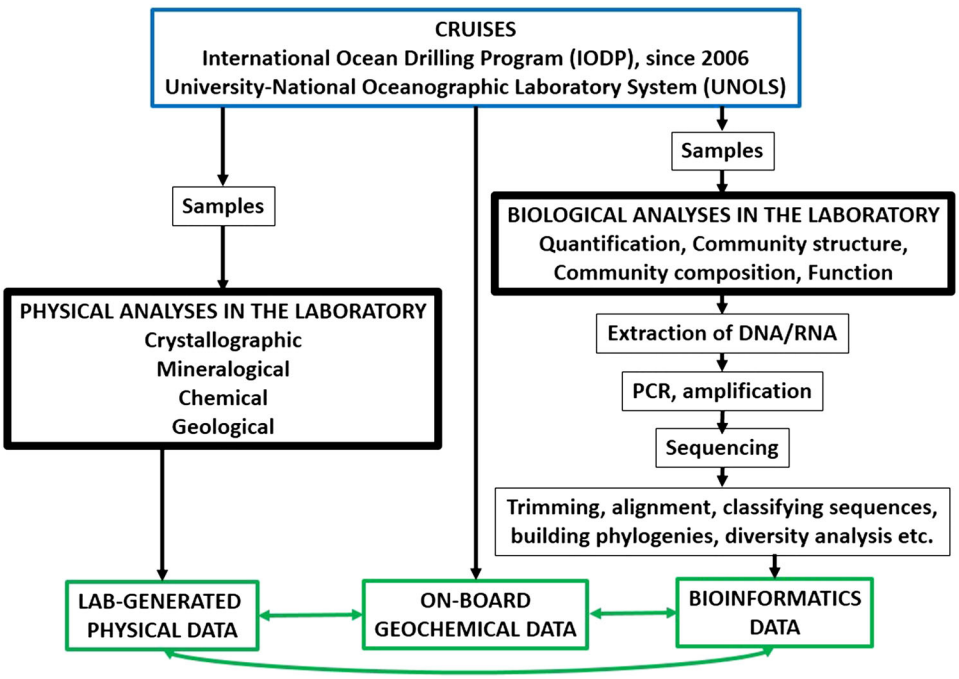

Fig. 1 A typical data workflow observed in the Jones laboratory

mutual shaping of the microbial community that exists under the seafloor at a particular site, and the physical composition of the surrounding seafloor.

The starting point for this data cycle is the collection of cores for analysis during scientific cruises. Some cores may be subject to onboard analyses, producing data about the cores' physical characteristics. When the cruise ends, some cores from IODP cruises are sent to one of the IODP core repositories, while other cores are distributed to various laboratories for biological and physical science analyses.

Within the laboratory, scientists typically specialize in one type of analysis or the other. Most of the members of the observed laboratory perform biological analyses, and then correlate with physical science data that are generated either onboard a cruise or by other scientists. For the sake of tractability, here we focus on biological analyses only.

The main focus of biological research in the laboratory is characterizing the ecology and function of microbes in cores. *Biomass* (matter from living, or recently living, microbes) is quantified, and microbes are classified into *operational taxonomic units (OTUs)*, namely groups of microbes with similar DNA sequences. These classifications are used to identify community members against previously characterized microbes found elsewhere, to see how different community members are related to each other, and to produce measures of community diversity. Scientists may also compare communities across sample sites.

Here, we focus on quantification of biomass, and classifying bacteria into OTUs, as these are common foci of projects in the observed laboratory. The first step in these analyses is the *extraction* of genetic material (DNA or RNA) from cores. One particularly common challenge is the relatively low biomass often found in deep-sea environments compared with biomass in other domains, due to the relatively low level of available nutrients. This challenge is critical because without an adequate biomass yield, the scientists cannot proceed with their analysis.

Following extraction, the scientist has DNA or RNA. In the case of RNA extraction, the the RNA sequence must then be transformed into its analog DNA sequence. The next step, known as *amplification*, involves the production of multiple copies of the DNA sequences of interest. When trying to characterize microbial communities, the standard portion of DNA sequenced is known as 16*S*. Primers, short sequences of nucleotide bases usually synthesized in the laboratory, are used to facilitate amplification.

The making of multiple copies of the target sequence is achieved through *polymerase chain reaction (PCR)*. Standard PCR results simply in multiple copies being made of the target sequence, while *quantitative PCR (qPCR)* also allows for the quantification of the levels of archaea, bacteria, and fungi (and their subtypes) in a sample. Following amplification and PCR (or qPCR), the next step is to generate a product that allows for *sequencing*, namely the process of determining what nucleotides are contained in each DNA sequence.

Sequencing can happen either within the laboratory or, more usually, is conducted by an external sequencing facility. The machine to carry out sequencing within the laboratory has only been recently acquired and so during the period of our fieldwork, external sequencing has been the predominant method used. These facilities produce DNA sequences that show the nucleotide bases of the 16S sequences that have been extracted from the cores.

Once sequences are acquired, there are multiple steps carried out in the laboratory to clean and process these sequences for analysis. Scientists receive two sequences corresponding to the same DNA sequence, namely a forward and backward sequence, and a scientist's first task is to marry these sequences together by matching nucleotides. Another important step is identifying and removing the part of the sequence that corresponds to the primers. A third stage of cleaning sequences involves checking that the correct nucleotide has been identified at each point along the sequence (or, *nucleotide-checking*), using data provided by the sequencing facility that shows the confidence with which each nucleotide has been identified.

Subsequently, sequences are aligned to allow for comparison (this is known as *sequence alignment*). Once aligned, sequences are then clustered into OTUs. OTUs in the sample are identified and classified by being compared with online databases of sequences of previously characterized bacteria.

Finally, the scientist seeks to produce representations of the microbial ecology they have found in their sample(s), and how this ecology compares to the microbial ecology found at other depths below the seafloor at the same site, or at other sites. One form of representation of the ecology in a single sample is pie charts, which show the relative proportions of archaea, bacteria, and fungi, and of their subtypes, in a sample. Another form of representation is *phylogenetic trees*. A phylogenetic tree shows how the OTUs in the sample may be related to each other. Finally, the scientist typically calculates numerical measures of the sample diversity.

The scientist may also compare the site analyzed with other sites. Making comparisons can involve producing *cladograms*, which are tree diagrams showing the relationships of different sites to each other. They may also produce Venn diagrams to illustrate overlaps between sites. Once these final steps are completed, the scientist may publish results in a journal or present at a conference. The biological data may then also be correlated with physical science data to understand how the physical environment shapes the microbial community, and vice versa.

Although resulting from a single workflow, there is nevertheless a great deal of heterogeneity of methods employed in producing this biological sequence data. It is toward this het-

erogeneity that we now turn. In particular, we focus on three steps: increasing the yield of nucleic acid; choosing how to sequence DNA; and the cleaning of 16S sequences.

3.3.2 Addressing the challenge of increasing nucleic acid yield

A critical challenge for scientists is to find methods that can increase biomass yield from cores. Different scientists improvise using different techniques, introducing an important level of heterogeneity across scientists in their methods for producing the final datasets and research outputs characterizing the communities of microbes that are found in the deep subsurface. We have observed at least four methods in this single laboratory. In this subsection, we focus on two in particular.

Adrian Adrian is a second-year PhD student, whose background was in microbiology domains other than the deep subseafloor. He encountered the problem of low biomass during the early days of his doctorate. In the laboratory at the time was a new postdoctoral researcher, George. Prior to joining the Jones laboratory, George had completed a PhD in which he investigated the microbial ecology of another environment in which there is very low biomass, and for which he learned a particular technique, called *multiple displacement algorithm (MDA)*, to increase the DNA yield. The various chemicals required to perform MDA are commercially available in a single kit.

Adrian, who had developed a strong rapport with George, turned to George for assistance. As a result, Adrian has become very conversant with the method of MDA. In addition to using George's expertise as a resource, Adrian is also able to secure financial resources to purchase the kit because the Jones laboratory is relatively well funded.

However, MDA is not a perfect solution, in the sense that it does not amplify all sequences with equal probability, which can foreclose the possibility of some of the subsequent steps of the analysis being performed, in particular quantitative measurements of different types of bacteria, archaea, and fungi. However, both Adrian and George agree that this trade-off is worthwhile because they only have access to limited quantities of physical samples for analysis. Given that they have found a technique that works for them, they are reluctant to waste scarce physical samples by attempting to use other methods with which they may be unfamiliar.

Jenny Jenny is a postdoctoral researcher who joined the Jones laboratory following completion of a PhD studying microbial ecology in another low biomass environment. Jenny also has an academic background in chemistry and soil science. Jenny does not use MDA to increase nucleic acid yield. Instead, she prefers a method that she developed in conjunction with her doctoral supervisor, as part of her doctoral research. When she encountered the challenge of how to increase nucleic acid yield during her doctoral research, Jenny was able to draw on her expertise in soil science to adapt existing techniques from studying soil microbiology to studying seafloor sediments.

Jenny was able to develop this method by drawing on a number of resources available to her at that time. Her Masters degree in soil science meant that she was conversant with much of the soil science literature, and was thus able to discover the existence of the paper presenting this method. With the assistance of her supervisor's expertise, Jenny was able to grasp the potential application of this method to seafloor sediments, and was given the encouragement to do so. Finally, Jenny's educational background in chemistry gave her expertise to draw on when developing and refining this method.

Jenny continues to use this method, because she does not like using commercially available kits. This dislike is not simply personal taste: Jenny finds that companies usually do not give sufficiently detailed information about the individual components of kits, limiting her ability to modify these kits. Instead, she is able to adapt the methods she has developed to different contexts of use. Her ability to do so is a direct result of her expertise gained through her academic background.

3.3.3 Making decisions about sequencing

Once nucleic acid yield has been increased, the scientists then undertake steps of PCR and cloning so that they can then subsequently sequence the DNA in the sample. A number of different options are available for outsourcing the production of sequence data, including private companies and other research institutions such as university laboratories or hospitals. The choice of which sequencing facility to use is generally up to the individual. An individual's choice is influenced by the interplay of a number of technical, scientific, economic, and social factors.

Two graduate students Diane and Mike are both PhD students in the Jones laboratory, and frequently use the same company to sequence their samples. Neither had a background in biological sciences prior to embarking on their PhDs.

Upon joining the Jones laboratory, Diane chose to use the same company that most other laboratory members were using, primarily because the laboratory already had an account set up with them and because, as a new member of the laboratory, she was reluctant to create additional administrative work for the laboratory manager. The laboratory manager orders chemicals and equipment on behalf of the scientists, and Diane's choice of sequencing facility

can be thus be seen as motivated by helping to ensure the laboratory manager would be willing to assist Diane as she moved forwards with her doctoral work. In other words, the way Diane uses resources available in her network not only help her to accomplish the immediate task of sequencing, but also help her to configure the network of resources available to her in anticipation of accomplishing future tasks.

Mike, too, used this same company when he first joined the laboratory on the advice of Richard, a postdoctoral researcher in the laboratory. As his background was in chemical engineering, he was keen to follow the expertise of others in the laboratory. When Mike first joined the laboratory, he approached the laboratory leader for assistance with many different technical issues, and the laboratory leader advised Mike that Richard would be able to help him. Mike was able to access the laboratory leader's social expertise regarding who in the laboratory possessed sufficient expertise to help Mike. In turn, Richard effectively became part of Mike's network, meaning Mike was then able to access Richard's expertise.

Jenny Jenny uses a different sequencing facility than the company used by Mike and Diane. Her chosen facility is one that she started to use while a PhD student. Jenny looked to the expertise of others when making her initial decision to use this facility. However, during our interview with Jenny, she also discusses the details of some different types of sequencing, and their scientific implications. Jenny uses her personal scientific knowledge and expertise to evaluate her current choice of sequencing facility and the particular services that she requires of the facility. Furthermore, Jenny's decisions about sequencing are also influenced by what is considered credible by the broader scientific community, i.e., the length of sequence that meets the standards of evidence required by this community. As with Mike and Diane (above), when Jenny was completing her doctorate, she drew on the advice of other scientists in her network. However, now that she has acquired more experience and knowledge, she is confident to make her own evaluations of the various types of sequences and sequencing facilities.

Of particular note here is that Jenny is drawing not only on her own scientific expertise to evaluate different options but also her social expertise regarding what the broader scientific community regards as credible. It is those in this broader community who are reviewers of the journal articles that Jenny writes, authors who may choose to cite Jenny's work in their own papers, possible future collaborators, potential future employers, or gatekeepers to future funding opportunities. In other words, Jenny is showing her awareness of the importance of building and sustaining networks in this broader community that may provide access to future resources, in turn influencing her choice of sequencing services.

3.3.4 Cleaning sequences

Once sequencing has been completed, the scientist receives the sequences in a file from the sequencing facility. However, before they are able to perform analyses on these sequences, the scientist needs to perform a number of steps to clean and prepare the sequences. We observed a number of differing configurations of computational tools that were employed in the laboratory to perform these steps. Two of these tools are presented here.

One tool is a piece of software called *Geneious* [35]. Geneious has a graphical user interface that allows the user to inspect and manage sequences. For each sequence, it displays the confidence with which the sequencing facility was able to identify each nucleotide. The user can manually delete or change individual nucleotides, or they can automate Geneious to remove all nucleotides or sequences falling below a certain confidence level. Acquiring a license for Geneious is expensive. The laboratory owns a license, and scientists usually access Geneious using the laboratory computer on which it is hosted, or by logging in remotely. Geneious works on the Apple interface only.

A second tool is *mothur* [48], which is available freely and uses a command-line interface. mothur automates all stages of sequence management, from cleaning through to analysis and production of graphical and pictorial representations of results. mothur can handle very large numbers (tens of thousands or even higher) of sequences.

Diane Diane, the graduate student encountered above in Sect. 3.3.4, does not use Geneious or mothur to clean sequences. She has spent a great deal of time living remotely from the laboratory where she is not able to access the computer in the Jones laboratory to use Geneious. Furthermore, Diane is not able to access this computer remotely as she owns a PC with Windows interface, which is not compatible with Geneious. The functionality of mothur for cleaning sequences was only added once Diane had completed a substantial portion of cleaning her sequences and Diane judges that the acquisition of expertise in using mothur would take more effort than it would save in terms of cleaning sequences.

Instead, Diane has embarked on some of the tasks involved in cleaning sequences by hand. However, because Diane performs these tasks at home, her husband has been able to see how time consuming some of these tasks are. Her husband has a background in computer science and suggested he write a program to automate the removal of primers, which he subsequently has done.

The network of resources Diane was able to access when she started processing sequences has determined the methods Diane employs to accomplish the task of cleaning sequences. Neither Geneious nor mothur formed part of this network at that time. Instead, Diane's only available resource was her

ability to complete the tasks by hand, becoming her approach by default. Initially Diane was not aware that her husband was in a position to help. She only became aware that he would be able to after he suggested to her that he write a program: it was only at this point that her husband's expertise has become part of the network of resources accessible to her.

The network of resources available to Diane has been dynamic over time as she becomes aware of her ability to access new resources. Further, we can see that it is Diane's perception of her husband's ability to help that has determined whether and when he is in her network of resources: he has possessed the technical ability all along to write a program, but it was not until he made Diane aware of this ability that he has become part of her network.

George George, the postdoctoral researcher who is presented in Sect. 3.3.3, performs most of the cleaning of sequences using Geneious. However, he does not like to use the features of Geneious that would allow him to automate all stages of the sequence-cleaning process, preferring instead to perform steps such as nucleotide-checking manually. In particular, George regards his approach as resulting in better quality sequences that may enable him to better identify species in the data, even though it is more time consuming.

George is able to access Geneious through the Apple computer in the laboratory and has an Apple laptop computer that means he is able to access the laboratory's copy of Geneious when he is working from home. He prefers to perform tasks, such as nucleotide-checking, manually to improve the quality of sequences, and is able to accomplish these tasks manually due to the affordances of Geneious.

Working manually, in turn, better enables him to identify species within these sequences, which promises to secure him greater scientific credibility and recognition in the eyes of the broader scientific community, and thus promises to increase his ability to build the network of resources available to him in the future.

However, George has started to perform a type of sequencing known as *tag sequencing*, which has much higher throughput and results in datasets comprising tens of thousands (rather than hundreds) of sequences. To check each sequence manually would be intractable. Instead, George has begun to use mothur. mothur was recommended to him by Lee, a new doctoral student in the laboratory who organizes mothur tutorials at the university and is a source of advice on how to use mothur within the laboratory.

Changes in the scale of datasets force George both to reconsider how he uses the resources he is able to access through his network and how to reconfigure his network to access other resources to complete the task of sequence cleaning. In particular, George was able to access Lee's expertise to learn how to use mothur, so mothur is now part of his network of resources.

3.4 Discussion

In Subsect. 3.3, we present examples of heterogeneity observed in data production practices in our case study of C-DEBI, with scientists in the same laboratory and even working on adjacent benches using a diversity of approaches to accomplish similar tasks and produce datasets similar in form and intent. Indeed, the heterogeneity presented above is only a fraction of the total heterogeneity that we have observed. For example, scientists perform analysis of sequences using a disparate range of tools and software including commercially available and open-source software, and sequence databases.

Heterogeneity of data practices has long been regarded as a hallmark of little science [46]. To date, this heterogeneity has been understood in terms of the types of dataset produced [13,46]. The analysis presented in Subsect. 3.3 introduces heterogeneity along another dimension, namely scientists using a diversity of practices to produce datasets similar in purpose and form.

As discussed above, our CENS research demonstrates that decisions made at each stage of the data life cycle have a cumulative effect on data [53]. In the case of the subseafloor biosphere, decisions regarding the choice of methods can have a significant impact on the results of scientific analyses, with important implications for the reuse of datasets. For example, the quantification of global subseafloor biomass is foundational to the study of the subseafloor biosphere, and attempts to quantify this biomass involve aggregating datasets from a wide range of studies [33]. However, a recent meta-analysis of studies of subseafloor life found that the method employed to quantify biomass can have a major impact on the results [40], with significant implications for the quantification of global biomass.

Furthermore, the ability of a scientist to assess the integrity and trustworthiness of a dataset tends to be enhanced when the scientist has greater knowledge about the factors—both social and technical—involved in the different stages of the dataset's production and curation [54]. Our CENS findings show that the production and use of multiple types of datasets significantly complicate these issues. Different people from different disciplinary backgrounds are involved in the production of different datasets, using different methods for producing these data. The task of tracking, documenting, and maintaining access to all of the datasets in a single workflow becomes extremely complicated. Adding another dimension of heterogeneity as described in Subsect. 3.3 can only complicate this task further.

3.4.1 Why does heterogeneity come about?

By focusing on individual scientists as the unit of analysis, we can understand how different scientists accomplish the same tasks in different ways. By viewing the process of accom-

plishing each task as a case of the scientist drawing on the sociotechnical networks of resources available to them at the time of carrying out the tasks, we are better able to account for this heterogeneity. Some of the factors that shape these networks are discussed here.

Disciplinary background We find that differences in disciplinary background promote heterogeneity of data practices along two dimensions in particular. The first is that some scientists may be aware of the existence of a particular tool or method due to their background whereas other scientists may not. For example, both George and Jenny employ techniques for increasing nucleic acid yield that they had learned or developed during their doctorates prior to joining the Jones Laboratory.

The second dimension is that different scientists may be aware of the same tool or method, but each evaluates its usefulness differently according to their particular knowledge and experience. For instance, when choosing sequencing facilities, Jenny considers some of their scientific advantages and disadvantages. Mike, on the other hand, had little prior experience of biological research and so trusts the judgment of others.

Career stage We also find that differences in career stages, in particular issues of social status related to career stage, can drive differences in how scientists make choices about which methods to pursue. For instance, we see that Diane's choice of sequencing facility as a newly arrived PhD student in the laboratory was influenced by her reluctance to cause additional burden for the laboratory manager. Instead, Diane's priority was ensuring a good working environment in which to pursue her PhD.

In the cases of both Jenny and George—both more senior scientists pursuing postdoctoral positions—we see that a concern with producing scientific work that is recognized as credible and significant by the broader scientific community is critical in shaping their choices of certain methods. In the case of Jenny, this concern impacted her decisions regarding sequencing. George chooses to eschew Geneious's ability to automate certain tasks involved in cleaning sequences to increase his chance of identifying novel species.

Social networks within and without the laboratory Another factor contributing to heterogeneity of methods is that different scientists have access to different social networks inside of the laboratory and outside. For instance, Adrian's use of MDA for increasing nucleotide yield was learned from George. Mike has been able to learn about which sequencing facility to use first by accessing the expertise of Professor Jones about who might have the expertise to help (i.e., Richard) and then approaching Richard. However, we also saw that Diane has been able to access the expertise of her husband—outside of the laboratory and of her scientific domain—to write a program to assist her with sequence cleaning.

Physical access to tools Although all members of the laboratory were in theory able to access all tools available at the time, circumstances mean that Diane is not able to access the computer in the lab, and thereby Geneious. Furthermore, her possession of a Windows laptop rather than an Apple Macintosh laptop means she is unable to access Geneious remotely. As a result, her sequence cleaning, unusually, does not involve Geneious.

Shifting networks of resources Another feature of the networks of resources to which people have access is that these networks are not static, but dynamic. Over time, scientists may change the way in which they accomplish certain tasks, or different scientists may perform the same task differently if they joined the laboratory at different points in time.

New tools or people being introduced to the laboratory can drive this dynamism. For instance, mothur has become available to George once the functionality of sequencing cleaning was added, and once Lee joined the laboratory. However, the network of resources available to a particular scientist also depends on the scientist's awareness of what resources exist. For instance, as Mike was made aware of Richard's expertise on various scientific matters, Mike then approached Richard for advice on sequencing facilities. Similarly, Diane's access of her husband's expertise has only occurred after her husband told her he would be able to help her.

Heterogeneity as a permanent feature The above discussion shows that the heterogeneity observed is not just happenstance but is instead a consequence of the interplay of multiple social, cultural, technical, and scientific factors. The dynamic nature of these factors suggests that the heterogeneity—and the challenges for data management that result—will be a persistent feature of this laboratory.

For example, new personnel will continue to enter the laboratory from a variety of disciplinary backgrounds with new expertise or approaches that others in the laboratory may learn from. Social networks—both within and without the laboratory—will continue to change, which will impact how knowledge about methods and tools will spread. The laboratory will acquire new tools and technologies that scientists may incorporate into their own workflows. Scientists will also move through career stages—from being a new doctoral student anxious not to cause disruption in their new laboratory through to a more senior doctoral student or postdoctoral scholar taking into account how the work they conduct will impact on their reputation in the broader academic

community. In short, challenges for successful data management and curation that result from heterogeneity of practices are likely to remain during the course of C-DEBI and beyond.

3.4.2 Implications of heterogeneity for assessing data integrity

A scientist's understanding and knowledge of what was involved in producing a dataset can have a major impact on the extent to which they are able to assess the dataset's integrity and trustworthiness. One example of where such knowledge can be useful is that the choice of technique for increasing nucleic acid yield can bias results (for example, the use of MDA), thus impacting on subsequent stages of the data life cycle, for instance by foreclosing subsequent analyses. Scientists who reuse such a dataset in their own work need to know not only the methods involved in producing the dataset but also that these methods render the dataset unsuitable for certain tasks. Not knowing one or the other of these could have major scientific implications.

Working out how to supply this knowledge is a complicated process: what granularity of details needs to be recorded? For instance, MDA is a kit and the protocols for its use are available on the company's website which make them easily accessible, whereas Jenny's method is contained within her publications and thus may be more difficult to locate. On the other hand, as discussed by Jenny, commercially available kits are often opaque about their precise components whereas she knows the types and quantities of chemicals she used and is willing and able to supply these details when asked.

The heterogeneity of methods significantly complicates the task of effective data management and curation for multiple purposes, such as checking and verifying scientific analyses and potential future reuse of datasets. At the same time, however, the heterogeneity observed in the laboratory underlines, and indeed makes more critical, the importance of curating not just datasets themselves but also information about their provenance (e.g., methods used in producing these, how these methods have been derived and adapted, and implications of the particular methods used for possible future uses of the datasets).

4 Conclusions

Scientists across a wide range of scientific disciplines are being confronted by the challenges of managing volumes of data increasing in both scale and diversity. To exploit the potential of digital scholarship to its fullest, it is vital to study existing data practices to inform the development of research infrastructures.

The Knowledge Infrastructures project presented in this paper has already made significant progress toward understanding data practices, and will continue to fill existing gaps in the literature. To date, studies have focused on heterogeneity in terms of the types of datasets being produced. In this paper, however, we find that even within a single workflow in a single laboratory, the practices, methods, and techniques used in the production of datasets can be highly heterogeneous, with many implications for data storage, curation, integration with other sources of data, and potential data sharing and reuse.

5 Future work in the knowledge infrastructures project

The empirical research presented in this paper provides a starting point for other themes that we are investigating. The heterogeneity of methods discussed above simultaneously makes more critical and more difficult different components of data management practice. Here, we briefly outline findings that will be covered in greater depth in future publications.

5.1 Recordkeeping in the laboratory

The first stage in effective data management is to ensure that records are kept about the production of data. Records kept at the sites of data production by those who produced the data can play an important role in generating effective metadata and in establishing provenance. In the case of the data workflow presented above, it is important to capture the heterogeneity of methods.

Scientists record details of their methods in laboratory notebooks. We have found that the notebook practices of scientists within the laboratory vary substantially in terms of the granularity of detail and the types of detail recorded. This variety is related to a number of sociotechnical factors, including scientists' disciplinary backgrounds and training received during undergraduate degrees, career stage, and personal preferences regarding detail. In other words, while the heterogeneity of methods makes effective recordkeeping more important, the very factors that drive this heterogeneity also contribute to the heterogeneity of practices in recordkeeping.

5.2 Storage and curation of laboratory-generated data

We are also finding that there are many differences across scientists in terms of the fate of laboratory-generated datasets, with many different points of data loss. There are multiple sociotechnical factors involved. For instance, the journals in which C-DEBI-funded scientists publish mandate that bio-

logical datasets supporting the arguments of articles must be deposited to external databases; conversely, there currently is no such requirement for physical science data. Thus, the data currently deposited in online repositories represents only a fraction of all data generated in the laboratory.

Furthermore, data may be lost when scientists leave the laboratory, when they take their own computers, hard drives, memory stick and other backup media with them. If they move into another domain of microbiology, or even leave the field or academia altogether, it becomes even more difficult for others to track and discover the scientists' data. The short-term nature of much of C-DEBI's funding contributes to this occurrence.

5.3 Where does data get shared?

Apart from the data that are deposited in databases, there appears to be little data shared within the collaboration. However, there are two particular circumstances in which data sharing has been observed within C-DEBI. The first is where a scientist discovers the existence of another's dataset through reading this latter scientist's paper. We are currently charting the processes by which this dataset might be eventually shared, from the initial steps of discovery, through negotiation (for instance, regarding crediting the dataset's originator), and eventual integration with other datasets. We are identifying many sources of friction—social, technical, and scientific—in these processes.

The other instances of data sharing observed in this paper are the result of serendipitous encounters between scientists from different institutions. These instances are infrequent. In particular, the sharing of data between researchers in different institutions or disciplines is a rare and fragile accomplishment that involves the alignment of multiple factors, including high levels of trust between researchers, alignment of researchers' interests, and opportunism in exploiting possibilities afforded by infrastructures.

5.4 Why is sharing data important?

Through our studies of C-DEBI, we are also developing a richer understanding of why data sharing can be important and beneficial to science, which will extend existing rationales for data sharing [8]. The data generated in the Jones laboratory, and the C-DEBI collaboration more generally, are often expensive and difficult to obtain. Furthermore, they are also made scarcer due to the relative novelty of the domain of study. Thus, sharing data promises many economic benefits for this domain.

Furthermore, the loss of datasets, and of information about workflows, can have significant implications on the ability to reproduce and validate the analyses of others. The ability to validate and reproduce such analyses is valuable in the context of C-DEBI: as outlined above, scientists bring many different approaches to such an analysis. It could be very useful for others to reproduce analyses that are based on unfamiliar or new methods and tools to test reliability of novel methods.

5.5 Big science meets little science: IODP cruises and laboratory practices

Case studies of the challenges and efforts in building knowledge infrastructures to support scientific work tend to follow the big/little science dichotomy, generally characterizing data life cycles as unfolding entirely in one context or the other.

At first sight, C-DEBI seems to exemplify little science. However, the C-DEBI data life cycle unfolds across both big and little science contexts. For example, the life cycle starts with the collection of physical samples on board scientific ocean drilling cruises, typically large-scale collaborations with expensive infrastructure and a large budget. Cruise samples and data are collected, managed, and made publicly accessible according to established standards. C-DEBI is an excellent opportunity to study how big and little science contexts shape each other, via the flow of individuals, physical samples, data, and data practices.

Acknowledgments The work in this paper has been supported by the Sloan Foundation Award #20113194, *The Transformation of Knowledge, Culture and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective*. We also acknowledge the contributions of Milena Golshan, Irene Pasquetto, and Laura A. Wynholds for commenting on drafts of this paper, and Elaine Levia for technical and administrative support.

References

1. Altman, M.: Digital preservation through archival collaboration: the data preservation alliance for the social sciences. Am. Arch. **72**(1), 169–182 (2009)
2. Anderson C.: The long tail. Wired Mag., **12**(10) (2004, October). http://www.wired.com/wired/archive/12.10/tail_pr.html
3. Aronova, E., Baker, K.S., Oreskes, N.: 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. Hist. Stud. Nat. Sci. **40**(2), 183–224 (2010). doi:10.1525/hsns.2010.40.2.183
4. Association of Research Libraries: The research library's role in digital repository services: final report of the ARL digital repository issues task force. Association of Research Libraries. Washington, DC (2009b). www.arl.org/bm~doc/repository-services-report.pdf
5. Bechhofer, S., Ainsworth, J., Bhagat, J., Buchan, I., Couch, P., Cruickshank, D., Sufi, S.: Why linked data is not enough for scientists. In: Sixth IEEE e-science conference. Brisbane, Australia (2010). http://eprints.ecs.soton.ac.uk/21587/
6. Berman, F., Lavoie, B., Ayris, P., Choudhury, G. S., Cohen, E., Courant, P., Van Camp, A.: Sustaining the digital investment: issues and challenges of economically sustainable digital preservation (Interim Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access). San Diego (2008). http://brtf.sdsc.edu/publications.html

7. Bijker, W.E., Hughes, T.P., Pinch, T.J.: The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology. MIT Press, Cambridge (1987)
8. Borgman, C. L.: Big data, little data, no data: scholarship in the networked world. MIT Press, Cambridge, MA (2015)
9. Borgman, C. L.: The premise and promise of the global information infrastructure. First Monday, **5** (2000). http://www.firstmonday.dk/issues/issue5_8/borgman/index.html
10. Borgman, C.L.: Scholarship in the Digital Age: Information, Infrastructure, and the Internet. MIT Press, Cambridge (2007)
11. Borgman, C.L.: The conundrum of sharing research data. J. Am. Soc. Inf. Sci. Technol. **63**(6), 1059–1078 (2012). doi:10.1002/asi.22634
12. Borgman, C.L., Wallis, J.C.: Building digital libraries for scientific data: an exploratory study of data practices in habitat ecology. In: Gonzalo, J., Thanos, C., Verdejo, M.F., Carrasco, R.C. (eds.) Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries, pp. 170–183. Springer, Berlin, Heidelberg, Alicante, Spain (2006)
13. Borgman, C.L., Wallis, J.C., Enyedy, N.D.: Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. Int. J. Digit. Libr. **7**(1–2), 17–30 (2007). doi:10.1007/s00799-007-0022-9
14. Borgman, C.L., Wallis, J.C., Mayernik, M.S.: Who's got the data? Interdependencies in science and technology collaborations. Comput. Support. Coop. Work **21**(6), 485–523 (2012). doi:10.1007/s10606-012-9169-z
15. Bozeman, B., Fay, D., Slade, C.P.: Research collaboration in universities and academic entrepreneurship: the-state-of-the-art. J. Technol. Transf. **38**(1), 1–67 (2013). doi:10.1007/s10961-012-9281-8
16. Callon, M.: The sociology of an actor–network: the case of the electric vehicle. In: Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World, pp. 19–34. Macmillan, London (1986)
17. Center for Dark Energy Biosphere Investigations: Center for dark energy biosphere investigations STC annual report 2013 (2014). http://www.darkenergybiosphere.org/internal/docs/C-DEBI-Annual-Report-2013.pdf
18. CODATA-ICSTI Task Group on Data Citation Standards Practices: Out of cite, out of mind: the current state of practice, policy, and technology for the citation of data. Data Sci. J., **12**, CIDCR1–CIDCR75 (2013). doi:10.2481/dsj.OSOM13-043
19. Data's shameful neglect. Nature, **461**(7261), 145 (2009). doi:10.1038/461145a
20. Dealing with data. Science, **331**(6018), 692–729 (2011)
21. Deuten, J. J.: Cosmopolitanising technologies: a study of four emerging technological regimes. Twente University Press, Enschede (2003).http://doc.utwente.nl/38695/1/t0000007.pdf
22. Edwards, K.: Center for dark energy biosphere investigations (C-DEBI): a center for resolving the extent, function, dynamics and implications of the subseafloor Biosphere (2009). http://www.darkenergybiosphere.org/internal/docs/2009C-DEBI_FullProposal.pdf
23. Edwards, P.N.: A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. MIT Press, Cambridge, MA (2010)
24. Edwards, P. N., Jackson, S. J., Bowker, G. C., Knobel, C. P.: Understanding infrastructure: dynamics, tensions, and design: report of a workshop on history and theory of infrastructure, lessons for new scientific cyberinfrastructures. National Science Foundation, Washington, DC (2007). http://hdl.handle.net/2027.42/49353
25. Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., Calvert, S.: Knowledge infrastructures: intellectual frameworks and research challenges (p. 40). University of Michigan, Ann Arbor, MI (2013). http://deepblue.lib.umich.edu/handle/2027.42/97552
26. Faniel, I.M., Jacobsen, T.E.: Reusing scientific data: how earthquake engineering researchers assess the reusability of colleagues' data. J. Comput. Support. Coop. Work **19**(3–4), 355–375 (2010). doi:10.1007/s10606-010-9117-8
27. Glaser, B.G., Strauss, A.L.: The Discovery of Grounded Theory: Strategies for Qualitative Research. Aldine Pub. Co, Chicago (1967)
28. Hammersley, M., Atkinson, P.: Ethnography: Principles in Practice. Routledge, London (2007)
29. Helland, P.: If you have too much data, then "good enough" is good enough. Commun. ACM **54**, 40–47 (2011). doi:10.1145/1953122.1953140
30. Hey, A.J.G., Trefethen, A.: The data deluge: an e-science perspective. In: Berman, F. Fox, G., Hey, A.J.G. (Eds.) Grid computing: making the global infrastructure a reality, pp. 809–824. Wiley, West Sussex, England (2003). http://www.rcuk.ac.uk/escience/documents/report_datadeluge.pdf
31. Hine, C.: Connective ethnography for the exploration of e-science. J. Comput. Media. Commun. **12**(2), 618–634 (2007). doi:10.1111/j.1083-6101.2007.00341.x
32. Hughes, T.P.: Technological momentum. In: Smith, M.R., Marx, L. (eds.) Does Technology Drive History? The Dilemma of Technological Determinism. pp. 101–113. MIT Press, Cambridge, MA (1994)
33. Kallmeyer, J., Pockalny, R., Adhikari, R.R., Smith, D.C., D'Hondt, S.: Global distribution of microbial abundance and biomass in subseafloor sediment. Proc. Natl. Acad. Sci. **109**(40), 16213–16216 (2012)
34. Karasti, H., Baker, K.S., Millerand, F.: Infrastructure time: long-term matters in collaborative development. Comput. Support. Coop. Work (CSCW) **19**(3–4), 377–415 (2010). doi:10.1007/s10606-010-9113-z
35. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al.: Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28**(12), 1647–1649 (2012)
36. Knorr-Cetina, K.: The manufacture of knowledge. Pergamon Press Oxford, (1981). http://sites.google.com/site/sciencestudies09/reader/Knorr-Cetina_ManKnow-Chapter1.doc
37. Knorr-Cetina, K.: Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press, Cambridge (1999)
38. Latour, B.: Science in Action: How to Follow Scientists and Engineers through Society. Harvard University Press, Cambridge (1987)
39. Latour, B., Woolgar, S.: Laboratory Life: The Construction of Scientific Facts, 2nd edn. Princeton University Press, Princeton (1986)
40. Lloyd, K.G., May, M.K., Kevorkian, R.T., Steen, A.D.: Meta-analysis of quantification methods shows that archaea and bacteria have similar abundances in the subseafloor. Appl. Environ. Microbiol. **79**(24), 7790–7799 (2013)
41. Lynch, M.: Art and artifact in laboratory science: a study of shop work and shop talk in a research laboratory. Routledge & Kegan Paul, London (1985)
42. Meng, X.-L.: Multi-party inference and uncongeniality. In: Lovric, M. (ed.), International Encyclopedia of Statistical Science, pp. 884–888. Springer, Berlin Heidelberg (2011). http://link.springer.com/referenceworkentry/10.1007/978-3-642-04898-2_381
43. O'Donoghue, T., Punch, K.: Qualitative Educational Research in Action: Doing and Reflecting. Routledge, London (2004)
44. Office of Science and Technology Policy: Harnessing the power of digital data for science and society: Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. Washington, D.C. (2009). http://www.nitrd.gov/About/Harnessing_Power.aspx
45. Østerlund, C., Carlile, P.: Relations in practice: sorting through practice theories on knowledge sharing in complex organizations. Inf. Soc. **21**(2), 91–107 (2005)

46. Palmer, C. L., Cragin, M. H., Heidorn, P. B., Smith, L. C.: Data curation for the long tail of science: the case of environmental studies. In: Presented at the 3rd International Digital Curation Conference, Washington, DC (2007). https://apps.lis.uiuc.edu/wiki/download/attachments/32666/Palmer_DCC2007.rtf?version=1
47. Ribes, D., Bowker, G.C.: Between meaning and machine: learning to represent the knowledge of communities. Inf. Org. **19**(4), 199–217 (2009). doi:10.1016/j.infoandorg.2009.04.001
48. Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., et al.: Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. **75**(23), 7537–7541 (2009)
49. Star, S.L., Ruhleder, K.: Steps toward an ecology of infrastructure: design and access for large information spaces. Inf. Syst. Res. **7**(1), 111–134 (1996). doi:10.1287/isre.7.1.111
50. Traweek, S.: Beamtimes and Lifetimes: The World of High Energy Physicists (1st Harvard University Press pbk.). Harvard University Press, Cambridge (1988)
51. Uhlir, P. F. (Ed.): For attribution-developing data attribution and citation practices and standards: summary of an International Workshop. The National Academies Press, Washington, D.C (2012). http://www.nap.edu/catalog.php?record_id=13564
52. Wallis, J.C., Borgman, C.L.: Who is responsible for data? An exploratory study of data authorship, ownership, and responsibility. In: Annual meeting of the American Society for Information Science and Technology (Vol. 48, pp. 1–10). New Orleans, LA. Information (2011). doi:10.1002/meet.2011.14504801188
53. Wallis, J.C., Borgman, C.L., Mayernik, M.S., Pepe, A.: Moving archival practices upstream: an exploration of the life cycle of ecological sensing data in collaborative field research. Int. J. Digital Curation **3**(1), 114–126 (2008). doi:10.2218/ijdc.v3i1.46
54. Wallis, J.C., Borgman, C.L., Mayernik, M.S., Pepe, A., Ramanathan, N., Hansen, M. A.: Know thy sensor: trust, data quality, and data integrity in scientific digital libraries. In: Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries, Vol. **LINCS 4675**, pp. 380–391. Springer, Budapest, Hungary:Berlin (2007). doi:10.1007/978-3-540-74851-9_32
55. Wallis, J.C., Rolando, E., Borgman, C.L.: If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PLoS ONE **8**(7), e67332 (2013). doi:10.1371/journal.pone.0067332