Limits to the Pursuit of Reproducibility: Emergent Data-Scarce Domains of Science

Peter T. Darch(✉)

School of Information Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, USA
ptdarch@illinois.edu

Abstract. Recommendations and interventions to promote reproducibility in science have so far largely been formulated in the context of well-established domains characterized by data- and computationally-intensive methods. However, much promising research occurs in little data domains that are emergent and experience data scarcity. This paper presents a longitudinal study of such a domain, deep subseafloor biosphere research. Two important challenges this domain faces in establishing itself are increasing production and circulation of data, and strengthening relationships between domain researchers. Some potential interventions to promote reproducibility may also help the domain to establish itself. However, other potential interventions could profoundly damage the domain's long-term prospects of maturation by impeding production of new data and undermining critical relationships between researchers. This paper challenges the dominant framing of the pursuit of reproducible science as identifying, and overcoming, barriers to reproducibility. Instead, those interested in pursuing reproducibility in a domain should take into account multiple aspects of that domain's epistemic culture to avoid negative unintended consequences. Further, pursuing reproducibility is premature for emergent, data-scarce domains: scarce resources should instead be invested to help these domains to mature, for instance by addressing data scarcity.

Keywords: Reproducibility · Data reuse · Little data · Open code
Open data

1 Introduction

Many key stakeholders (such as funding agencies, professional societies, researchers, and members of the information professions) regard pursuit of reproducibility as an urgent concern for all domains of science [1–4]. These stakeholders are concerned with promoting scientific integrity, and the ability to reproduce published scientific findings by replicating steps in the original analysis can detect error and malpractice.

To date, interventions and recommendations to promote reproducibility have largely been devised in the context of well-established data-intensive domains [5]. However, there are many other domains of science that are new and emergent, and that face a critical scarcity of data that hinders their prospects of maturation. These domains are culturally distinct from well-established data-intensive domains. Interventions to

G. Chowdhury et al. (Eds.): iConference 2018, LNCS 10766, pp. 164–174, 2018.
https://doi.org/10.1007/978-3-319-78105-1_21

advance reproducibility formulated in the context of well-established data-intensive domains may be unsuitable or even damaging if implemented in emergent data-scarce domains. Rather than investing their limited resources in interventions to promote reproducibility, emergent data-scarce domains should instead prioritize addressing data scarcity, for instance by investing in infrastructure for data production and reuse.

Through presenting a longitudinal case study of an emergent data-scarce domain, deep subseafloor biosphere research, this paper addresses the following questions:

(1) How feasible is pursuing reproducibility in emergent data-scarce domains?
(2) How desirable is pursuing reproducibility in emergent data-scarce domains?

2 Background

To frame subsequent discussions of reproducibility, this section first covers the concept of epistemic cultures, particularly in relation to data and software. Next, it considers efforts to promote reproducibility. This section concludes by discussing challenges facing deep subseafloor biosphere research as an emergent data-scarce domain.

2.1 Epistemic Cultures in Science

The *epistemic cultures* [6] of different scientific domains can vary in many ways, including how research activities are organized (such as the size of teams involved), and what counts as evidence of scientific phenomena. Domains also differ according to degree of institutionalization: markers of a well-established domain can include its own journals, conferences, professional societies, university departments or research institutes, and dedicated streams within funding agencies [7].

Other major differences between domains' epistemic cultures relate to data and software [1]. Some domains, such as astronomy and computational social science, are characterized by the use of highly standardized computationally- and data-intensive methods. Data and software sharing in these *big data* domains is typically supported by sophisticated digital infrastructure.

By contrast, *little data* domains are characterized by access to much smaller quantities of data that are often heterogeneous both in type and by method of production [1]. In these domains, such as ecology, data and software sharing is frequently inhibited by patchy or inadequate standards for data production, analysis, and management; inconsistent policies; and uneven provision of digital infrastructure [8]. Successful data sharing is often facilitated by personal contact between the original data producer, and the potential data user.

2.2 Computational Reproducibility: "Barriers" and Interventions

Stodden [9] distinguishes different types of reproducibility. *Empirical reproducibility* refers to provision of details about a non-computational experiment that allows another researcher to carry out the experiment. *Computational reproducibility* refers to availability of code and data used to produce a piece of research. As each type of

reproducibility has different requirements and faces distinct challenges for its realization, this paper will focus on computational reproducibility.

The pursuit of reproducible science is often framed as a process of identifying and overcoming "barriers to reproducibility" [10, p. 73]. Frequently identified barriers include a lack of digital infrastructure for making code and data publicly accessible, and a lack of policies to encourage use of infrastructure where it exists [1]: these barriers can be addressed by building new infrastructure, and devising and enforcing new policies. Another barrier is use of proprietary software [11], which inhibits reproducibility for many reasons: its source code is often not publicly accessible; researchers are not able to extract and share workflows they produce using this software; and prospective reproducers may have to pay to use this software. Researchers are instead encouraged to use open source software, or to write and publicly share their own code [12]. Other scholars argue for new cultural norms to advance reproducibility, such as reproducibility "etiquette", where the prospective reproducer of a piece of research contacts the author who originally conducted that research [13, p. 310].

Interventions and recommendations to promote reproducibility often require substantial investments of resources in building infrastructure, devising policies, and changing practices. So far, these interventions and recommendations have been mainly formulated in the context of well-established big data domains [1]. Recently, attention has shifted to reproducibility in fields that are not usually considered big data, such as archaeology [14], although the focus is typically on the specific areas of those fields where computationally- and data-intensive practices are the norm [15].

2.3 Deep Subseafloor Biosphere Research: An Emergent Data-Scarce Domain

One type of little data domain is the *data-scarce* domain [16], characterized by not having enough data to pursue the domain's major objectives. Data scarce domains are often new and emergent, multidisciplinary, and struggle for resources as they attempt to establish themselves. Addressing data scarcity is a critical step for helping these domains to mature and raise their status. One example is deep subseafloor biosphere research, whose researchers integrate physical science and bioinformatics data to answer questions about relationships between microbial communities in the seafloor and the physical environment they inherit.

Since studies of the deep subseafloor biosphere began in the late 1990s, two infrastructures have been instrumental to this domain's emergence. One is the *Center for Dark Energy Biosphere Investigations* (*C-DEBI*), a ten-year NSF Science and Technology Center launched in 2010, providing short-term funding to over 150 researchers across the US and Europe. Since 2015, C-DEBI has operated an online data portal. C-DEBI requires recipients of its funding to upload data they produce to an openly-accessible public database or (where no relevant database exists) to its own portal.

The second infrastructure is the *International Ocean Discovery Program* (*IODP*), which operates five scientific ocean drilling cruises per year to procure physical

samples (*cores*) of the seafloor for analysis. IODP serves multiple domains besides deep subseafloor biosphere research, such as studies of plate tectonics.

Besides C-DEBI and IODP, deep subseafloor biosphere research has little institutional strength: no journals exist that are dedicated to this domain, and its researchers are distributed across multiple university departments (including departments of biological sciences, of earth sciences, and of oceanography). A key objective of C-DEBI is to foster links between these researchers, and it provides significant funding to promoting research collaborations between distributed researchers.

The rarity of cruises, requirements to share IODP resources with other domains, and the domain's relative newness means deep subseafloor biosphere research has access to small quantities of data. This domain is data-scarce in that researchers wish to address the domain's research topics in a more statistically intensive manner than is afforded by current data [16]. Domain leaders seek to transition the domain from *discovery-driven* science, where researchers describe microbial communities in cores, to *hypothesis-driven* science, where researchers test statistical hypotheses about microbial activity. This transition would bring the domain in line with domains that study microbes in other environments. Domain leaders also hope deep subseafloor biosphere research contribute to key open questions in science through producing and integrating datasets about microbial communities in different geographic locations.

By addressing data scarcity, the domain's leaders seek not only to produce more and better science, but also to help the domain mature and increase its institutional strength [16]. Through shifting to hypothesis-driven science and addressing high profile questions, domain leaders hope the domain will become more credible and better-established. Thus, improving production of new data and encouraging circulation and reuse of extant data are critical priorities for the domain's leaders. C-DEBI also allocates a great deal of resources towards pursuing these priorities.

3 Methods

This paper presents findings from a longitudinal qualitative case study of deep subseafloor biosphere researchers, focussing on C-DEBI and IODP. Research methods comprised long-term participant observation, interviews, and document analysis, following standard ethnographic practices [17]. Fieldwork included eight months embedded in a laboratory headed by a leading figure in C-DEBI at a large US research university, weeklong observation trips to two other laboratories and to IODP headquarters, and observations of a research expedition and scientific conferences.

The interview sample comprises 55 people, including C-DEBI-affiliated scientists (n = 41), and curators and managerial staff involved in IODP (n = 14). Interviews ranged in length from 35 min to two hours and 30 min, with the majority being between one and two hours. Documents analysed include official C-DEBI documents such as Annual Reports, and documents about IODP operations.

4 Findings

This section presents a typical workflow in deep subseafloor biosphere research. Although domain researchers conduct a growing range of analyses, with different purposes, the workflow discussed here is widely used by researchers in many laboratories. This sections first describes key steps in this workflow, and the choices made by researchers during these steps, and then discuss the challenges that would be faced by a researcher seeking to reproduce a project incorporating this workflow.

4.1 A Typical Workflow in Deep Subseafloor Biosphere Research

The central aim of a project incorporating this workflow is to characterize the composition of the microbial community (type and quantity of microbes) in a particular part of the seafloor, and to understand this community's relationship to the physical environment it inhabits. The workflow is summarized in Fig. 1.

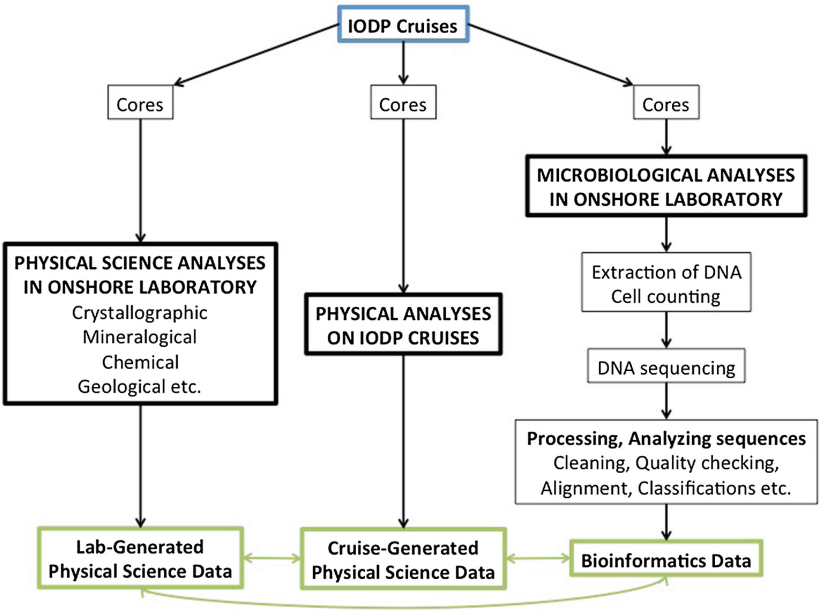

Fig. 1. A typical workflow in deep subseafloor biosphere research

The first step in this workflow is the collection of cores on IODP cruises. Some cores are subject to onboard analyses that yield baseline data of their physical characteristics. These data are made available through an online IODP database. Other cores are distributed among cruise participants, who take them to their onshore laboratories to analyze their physical characteristics, and the microbial communities they contain.

Here, we will focus more heavily on microbiological analyses. The first steps in a microbiological analysis are counting the number of cells and extracting DNA from cores. Researchers, even those in the same laboratory, display a high degree of methodological heterogeneity when conducting these steps (see [18] for more details). The reason for this heterogeneity is that, given the subseafloor biosphere is a low biomass environment, traditional methods for cell counting and DNA extraction do not work. Instead, researchers adapt methods they learned prior to embarking on deep subseafloor biosphere research. The type of method used, however, has implications for the bioinformatics data that is subsequently generated: some methods are biased in the sense they lead to overrepresentation of some types of microbes in the subsequent steps of the workflow, and some methods are more efficient than others, resulting in a greater yield of DNA.

Next, DNA is prepared for sequencing. Sequencing is carried out either in the laboratory itself or, more commonly, by an external sequencing facility. The outcome of sequencing is a file comprising a series of DNA sequences, representing the microbes in the core. Each sequence comprises a series of nucleotides, and the sequencing facility typically returns the sequences with probability estimates of how accurately they were able to identify each nucleotide (known as *quality scores*). Next, the researcher processes and analyzes these sequences. The first step is to use quality scores to check and clean sequences. Next, sequences are aligned. Similar sequences are clustered into *Operational Taxonomic Units*, which are then compared with publicly accessible bioinformatics databases of already-known microbes.

Researchers use a range of computational methods process and analyze sequences. Some researchers write their own code. Other researchers use a piece of open source software called mothur. Finally, researchers who are less comfortable with computational methods often choose to use proprietary software called Geneious, with a graphical user interface that researchers find intuitive and easy to use.

Next, the researcher correlates the microbial community's composition with certain characteristics (such as geochemical or mineralogical) of the physical environment it inhabits, with the aim of understanding how these physical characteristics shape the microbial community and vice versa. These physical science data may come from the IODP database, or from analyses of cores in onshore laboratories. Once the researcher has completed their analysis, they will prepare an article for publication. This article presents brief information about the methods used. Journals often require the researcher to upload supporting DNA sequences to a publicly accessible bioinformatics database before article publication: once uploaded, sequences are assigned accession numbers by the database, and these numbers are published alongside the journal article. C-DEBI also now requires all physical science data produced by its researchers to be uploaded to a relevant publicly accessible database.

However, not all research products resulting from the described workflow are made publicly accessible, such as pre-cleaned DNA sequences and quality scores, and code written by researchers. These research products often eventually got lost, for instance, when a graduating doctoral student takes up a position in industry.

4.2 Reproducing This Workflow: Accessing Data, Software, and Code

To reproduce this workflow in its entirety requires access to physical samples, data, and code and software. Reproducibility of the steps that involve handling core samples is highly infeasible: given the expense and rarity of IODP cruises, the IODP personnel interviewed explained that cores would only be given to researchers to produce new science, and not to reproduce previous analyses. Instead, the focus here is on later, post-sequencing, steps in the workflow.

A prospective reproducer of the workflow is likely to be able to access sequence data and physical science data used in the analysis, given the policies and digital infrastructure currently in place. However, the reproducer is unlikely to easily access data received from the sequencing facility (pre-cleaned sequences, and quality scores), posing a significant challenge in reproducing sequence-cleaning and quality-checking steps. Further, the reproducer may also lack the information necessary required to interpret sequence data, such as detailed accounts of the methods used to produce the sequences, to understand whether these data may contain biases.

The prospective reproducer may therefore need to contact the researcher who originally conducted the project for access to some data, and help in interpreting these data. However, a number of interviewed researchers expressed reservations about doing so, as they do not want to undermine relationships with their domain colleagues by implying they did not trust these colleagues' competence and honesty. The deep subseafloor biosphere domain is relatively small: maintaining good relationships is very important to researchers, particularly junior researchers who rely on senior researchers for patronage and employment opportunities.

A prospective reproducer may or may not be able to access the software or code necessary to reproduce the workflow, depending on the computing choices made by the researcher who conducted the original research. If open source software was used, the reproducer should be able to access this software. If the research was conducted using proprietary software, the prospective researcher is unlikely to be able to reproduce the workflow: the software costs money to use (an annual license for Geneious currently costs $395), its source code is not openly available, and it does not allow users to extract and share workflows.

Finally, if the research involved code written by the researcher themselves, the prospective reproducer may not be able to access this code. Occasionally, subseafloor biosphere researchers who produce their own code do make this code openly available, for instance via an online repository or their own website. Otherwise, the prospective reproducer would have to approach the researcher for the code: however, as with data, some subseafloor biosphere researchers expressed reservations about approaching colleagues for code for the purpose of reproducing research.

5 Discussion and Conclusions

In common with other emergent data-scarce domains, the deep subseafloor biosphere has two important objectives for establishing itself in the long-term. One objective is maintaining and deepening interpersonal relationships between domain researchers as a

necessary precursor to increasing the domain's institutional strength. The second objective is addressing data scarcity to raise the domain's scientific profile. Although some proposed interventions to promote reproducibility may also aid pursuit of these two objectives, other commonly-proposed interventions are potentially fundamentally incompatible with these objectives. Emergent data-scarce domains should focus their scarce resources on addressing data scarcity rather than on pursuing reproducibility.

5.1 An Intervention that Promises to Benefit Emergent Data-Scarce Domains

While C-DEBI has made major strides in policy and infrastructure towards ensuring some data produced by its researchers are made openly accessible, other data and code necessary to fully reproduce deep subseafloor biosphere workflows remain inaccessible. Investing in better infrastructure for data and code is a key step in pursuing reproducible science, as is devising and enforcing policies that require researchers to use this infrastructure [1]. These steps are compatible with addressing data scarcity by promoting circulation and reuse of data and software.

5.2 Interventions that Risk Damaging Emergent Data-Scarce Domains

Some practices that are promoted as fundamental to reproducibility seem to be incompatible with the interests of emergent data-scarce domains. These practices have the potential to make data scarcity more acute or to undermine critical relationships between domain researchers.

Risk of Making Data Scarcity More Acute. A key requirement for promoting reproducibility is that code or software used in research should be openly accessible to others [12]. This requirement means researchers should avoid using proprietary software and instead either write and make openly available their own code, or use open source software. This requirement could conflict with production of new data and science in emergent data-scarce domains. In deep subseafloor biosphere research, computational skills are patchy. Researchers have experienced disparate amounts of computational training prior to joining the domain, depending on their disciplinary backgrounds. Researchers with lower levels of comfort with computational methods exhibited a strong preference for using a piece of proprietary software. The use of this software enables them to produce and process data more rapidly than the alternatives. Requiring these researchers to switch away from their preferred software would be likely to slow down production of new data, potentially exacerbating data scarcity.

In the long-term, this source of conflict between pursuing reproducibility and pursuing the domain's objectives is likely to lessen. As coding becomes more widespread in scientific curricula, more researchers are likely to enter the domain able to write their own code, or contribute to development of open source software. However, in the shorter term, demands of reproducibility will need to be balanced with the domain's critical need to produce new data and science.

Risk of Undermining Critical Interpersonal Relationships. Contact between the researcher who conducted the original research, and the prospective reproducer of that

research, may be an integral part of reproducible science. Some advocates of reproducibility have argued that this contact is good etiquette and should become a cultural norm [13]. Even if this practice does not become an integral part of reproduction for the purposes of good manners, such contact may nevertheless be necessary when reproducing research in emergent data-scarce domains. However, such contact also risks undermining these domains' prospects of maturation.

Research on data and code sharing and reuse demonstrates that, in these domains, direct contact between the producer and potential reuser is often necessary so that the potential reuser can better understand and interpret a dataset or piece of code – even when this dataset or code is made openly accessible via a digital repository [8]. Likewise, a potential reproducer may well need to contact the researcher who conducted the original research for help in understanding data or code. For instance, the methodological heterogeneity in the deep subseafloor biosphere means that a prospective reproducer with a background in one scientific discipline may need help in understanding a dataset produced by a method that originated in another discipline.

However, contact between the researcher who conducted the original research, and the prospective reproducer risks damaging the strength of emergent data-scarce domains. Many deep subseafloor biosphere researchers expressed their concern at approaching a colleague for the resources necessary to reproduce this colleague's research, believing it would degrade their relationships. Unlike well-established big data domains, deep subseafloor biosphere research lacks institutional strength. Instead, the strength of the deep subseafloor biosphere domain relies on the strength of interpersonal relationships between researchers. Maintaining and deepening these ties is critical for the domain, and is a necessary precursor to increasing the domain's institutional strength. While sharing data and code for reuse can reinforce these ties by implying collegiality and forming the basis for future collaboration, sharing for reproducibility threatens instead to undermine these ties.

5.3 Implications for Pursuing Reproducible Science

This paper has two implications for pursuing reproducibility. One implication is to challenge the dominant framing of the pursuit of reproducibility as identifying, and then devising interventions to overcome, "barriers to reproducibility" [10]. This framing lends itself to a narrow focus on evaluating a possible intervention from the perspective of whether it advances reproducibility in a particular domain. However, this intervention could have far-reaching and harmful unintended consequences for the domain that go well beyond reproducibility.

Interventions devised in the context of established big data domains should instead only be rolled out to other domains with due care. The pursuit of reproducibility in a domain should involve understanding and analyzing that domain's epistemic culture [6] in its entirety, to better anticipate potential consequences of specific proposed interventions. The case study in this paper suggests that particularly relevant dimensions of an epistemic culture to consider include the domain's objectives, the domain's institutional strength, the role and scale of data in domain research, methodological heterogeneity in the domain, domain researchers' software/coding preferences, the disciplinary backgrounds and training of domain researchers, available digital

infrastructure, the nature of relationships between domain researchers, and existing norms regarding sharing of data and software within the domain.

A second implication of this paper is that pursuing reproducibility should not be a priority for emergent data-scarce domains. C-DEBI focuses its scarce resources on addressing data scarcity and cultivating relationships between researchers, activities that help the domain to mature. Pursuing reproducibility prematurely could risk the long-term prospects of emergent data-scarce domains by directing scarce resources away from activities that help these domains establish themselves, and towards activities that instead hinder their maturation. Future work will examine the extent to which domains should mature before they pursue reproducibility. Reproducibility is important for scientific integrity, and its realization should be a major long-term goal for all scientific domains. However, its pursuit must not be at the expense of the development of promising emergent data-scarce domains.

Acknowledgements. This work is funded by the Alfred P. Sloan Foundation (Awards #20113194, #201514001). Thank you to current members of UCLA Center for Knowledge Infrastructures (CKI) for comments on earlier drafts of this paper (Christine L. Borgman, Bernie Boscoe, Milena S. Golshan, Irene Pasquetto, and Michael J. Scroggins), to past members of CKI (Ashley E. Sands and Sharon Traweek) for discussion of ideas, and to Rebekah L. Cummings for assistance with data collection. Thank you also to the C-DEBI and IODP personnel who were observed and interviewed.

References

1. Borgman, C.L.: Big Data, Little Data, No Data: Scholarship in the Networked World. The MIT Press, Cambridge (2015)
2. Vitale, C.R.: Is research reproducibility the new data management for libraries? Bull. Assoc. Inf. Sci. Technol. **42**(3), 38–41 (2016)
3. Baker, M.: 1,500 scientists lift the lid on reproducibility. Nat. News **533**(7604), 452 (2016)
4. Pellizzari, E., Lohr, K.N., Blatecky, A., Creel, D.: Reproducibility: A Primer on Semantics and Implications for Research, 1st edn. RTI Press/RTI International, Research Triangle Park (2017)
5. Stodden, V., Leisch, F., Peng, R.D. (eds.): Implementing Reproducible Research. CRC Press, Boca Raton (2014)
6. Knorr-Cetina, K.: Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press, Cambridge (1999)
7. Lenoir, T.: Instituting Science: The Cultural Production of Scientific Disciplines. Stanford University Press, Stanford (1997)
8. 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)
9. Stodden, V.: Resolving irreproducibility in empirical and computational research. IMS Bull. Online (2013)
10. Ram, K., Marwick, B.: Building towards a future where reproducible, open science is the norm. In: Kitzes, J., Turek, D., Deniz, F. (eds.) The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences, pp. 69–78. University of California Press, Oakland (2018)

11. Ince, D.C., Hatton, L., Graham-Cumming, J.: The case for open computer programs. Nature **482**(7386), 485–488 (2012)
12. Stodden, V., et al.: Enhancing reproducibility for computational methods. Science **354** (6317), 1240–1241 (2016)
13. Kahneman, D.: A new etiquette for replication. Soc. Psychol. **45**(4), 310 (2014)
14. Marwick, B.: Computational reproducibility in archaeological research: basic principles and a case study of their implementation. J. Archaeol. Method Theory **24**(2), 424–450 (2017)
15. Kitzes, J., Turek, D., Deniz, F. (eds.): The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences. Univ of California Press, Oakland (2018)
16. Darch, P.T., Borgman, C.L.: Ship space to database: emerging infrastructures for studies of the deep subseafloor biosphere. PeerJ Comput. Sci. **2**, e97 (2016)
17. Hammersley, M., Atkinson, P.: Ethnography: Principles in Practice, 3rd edn. Routledge, London (2007). Reprinted
18. Darch, P.T., Borgman, C.L., Traweek, S., Cummings, R.L., Wallis, J.C., Sands, A.E.: What lies beneath?: knowledge infrastructures in the subseafloor biosphere and beyond. Int. J. Digit. Libr. **16**(1), 61–77 (2015)