Beyond Big or Little Science: Understanding Data Life Cycles in Astronomy and the Deep Subseafloor Biosphere

Peter T. Darch, University of California, Los Angeles
Ashley E. Sands, University of California, Los Angeles

Abstract
For decades, the big science and little science dichotomy has served as a starting point for many analyses of scientific research and data practices, including studies used to inform the construction of scientific knowledge infrastructures. We challenge this dichotomy by presenting findings from longitudinal, qualitative case studies of data life cycles in two scientific domains, each centered around a large, distributed scientific collaboration. One is astronomy and the Sloan Digital Sky Survey (SDSS). The other is the deep subseafloor biosphere and the Center for Dark Energy Biosphere Investigations (C-DEBI). We show that some critical stages of the data life cycle in each domain unfold in big science contexts while other critical stages occur in little science contexts. Furthermore, these big and little science contexts shape each other dynamically. This challenging of the big and little science dichotomy has implications for the building of scientific knowledge infrastructures, including those supporting data management.
Keywords: Big science, little science, microbiology, astronomy, knowledge infrastructures
Citation: Darch, P.T., Sands, A.E. (2015). Beyond Big or Little Science: Understanding Data Lifecycles in Astronomy and the Deep Subseafloor Biosphere. In *iConference 2015 Proceedings*.

Acknowledgements: This research is funded by the Alfred P. Sloan Foundation (*The Transformation of Knowledge, Culture, and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective*, #20113194, P.I. Christine L. Borgman, Co-P.I. Sharon Traweek). Earlier data collection was funded by the U.S. National Science Foundation (*Data Conservancy*, #OCI0830976, P.I. Sayeed Choudhury, and *Knowledge & Data Transfer: the Formation of a New Workforce*, #1145888, P.I. Christine L. Borgman, Co-P.I. Sharon Traweek). We thank in particular Christine L. Borgman and Sharon Traweek for their guidance and mentorship. We also acknowledge the contributions of Milena Golshan and Irene Pasquetto for commenting on earlier drafts of this paper, and Rebekah Cummings, Laura A. Wynholds, and David S. Fearon for assistance with conducting the case studies. We are deeply grateful to the C-DEBI, IODP, and SDSS personnel, and other astronomers, who we interviewed and observed at work.
Research Data: In case you want to publish research data please contact the editor.
Contact: petertdarch@ucla.edu, ashleysa@ucla.edu

1 Introduction

Novel digital technologies enable the collection of vastly more data, at faster rates, than ever before across a wide range of scientific disciplines (Borne, 2013; Hey, Tansley, & Tolle, 2009). However, the promise of these technologies is predicated upon the availability of *knowledge infrastructures*, defined as "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds" (P. N. Edwards, 2010, p. 17), in order to enable the production, management, curation, and accessibility of data. The provision of such infrastructures is uneven both across and within scientific domains. Much uncertainty exists around what should be built or how to build it, particularly in dynamic contexts of ever-changing scientific technologies (Bell, Hey, & Szalay, 2009; Borgman, 2007, 2015; P. N. Edwards et al., 2013).

Studies of scientific data practices are conducted to inform the construction of these infrastructures. The demarcation of scientific work into *big science* and *little science* is a powerful trope in studies of scientific practices generally (Furner, 2003a, 2003b; Price, 1963), and as a starting point in studies of data practices more specifically (Bicarregui et al., 2013; Cragin, Palmer, Carlson, & Witt, 2010).

To explore the extent to which this boundary holds, we present findings from case studies of data life cycles in two scientific domains. One is astronomy, centered on the creation and management of data in the *Sloan Digital Sky Survey* (*SDSS*)[1], and the subsequent use of these data by individual and small groups of astronomers. The other case study focuses on scientists studying the ecology of the deep subseafloor biosphere (microbial life under the seafloor) as part of the *Center for Dark Energy Biosphere Investigations* (*C-DEBI*)[2].

Although astronomy is often regarded as an exemplar of big science (Borne, 2013), and ecology as an exemplar of little science (Borgman, Wallis, & Enyedy, 2007), we find that data life cycles in each case study unfold across both big and little science contexts. Furthermore, these big and little science contexts shape each other. The binary of big and little science, while a helpful categorization in some

[1] http://www.sdss.org/
[2] http://darkenergybiosphere.org/

ways, may prove unsuitable to the development of infrastructure for scientific projects and domains that do not neatly fit within only the criteria of big or of small science.

2 Life Cycles for Scientific Data

The term *life cycle*, as applied to science, can be defined as "the socio-technical ensemble of activities of a particular field of practice and the associated artifacts", including stages of planning, facilitating, carrying out, and disseminating results of a scientific project (Pepe, Mayernik, Borgman, & Van de Sompel, 2010, p. 571).

Life cycle models focusing more specifically on scientific data tend to foreground stages involving the preservation and access of data, i.e. stages after data have already been produced (Greenberg, 2009). A life cycle model that gives a fuller account of all stages in a scientific project, including planning and data collection, is developed by Wallis, Borgman, Mayernik & Pepe (2008). This model comprises nine stages (see Figure 1):

a) **Experimental design;**
b) **Calibration and ground-truthing,** involving testing and refinement of equipment;
c) **Data capture,** which may involve measurements of physical phenomena, and collection and processing of physical samples;
d) **Cleaning data,** which can include application of calibration data, or removal of outliers;
e) **Deriving numerical data,** involving transforming observational data and samples into more meaningful data points;
f) **Integrating data from multiple sources,** either the researchers' own data or from external sources;
g) **Data analysis** to test and generate hypotheses, and to draw conclusions;
h) **Publication** of findings in, for example, journals or conference papers;
i) **Storage and preservation,** which may include local storage on personal computers or laboratory servers or in publicly accessible databases.

Although presented sequentially, in practice these stages are not discrete or linear: a particular stage may occur at multiple times during the course of a project.

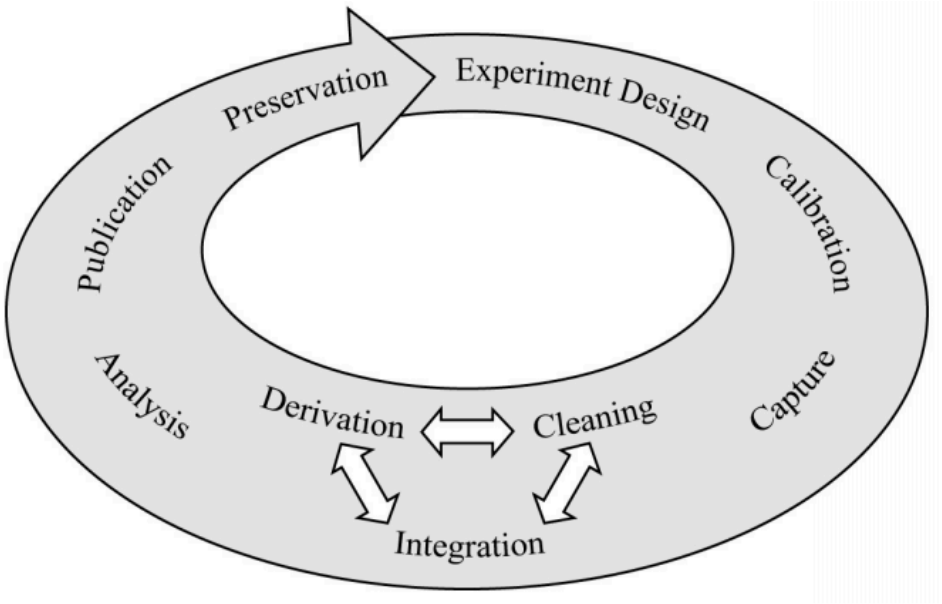

Figure 1. The scientific data life cycle (image from Wallis et al. (2008))

3 Big and Little Science, and Knowledge Infrastructures

The terms "big science" and "little science" are used to characterize and contrast two styles in the organization of scientific work (Price, 1963, p. 1). Both terms have been subsequently subject to a wide range of definitions. In recent decades, the dichotomy of big and small science remains a highly influential paradigm within a wide range of academic disciplines (most notably Information Studies, Science and Technology Studies, History of Science, and Sociology) for understanding and analyzing scientific work (Furner, 2003a, 2003b).

3.1 Little Science

Little science research is defined in relation to the small scale of projects along multiple dimensions. Little science is also referred to variously as *small science* (Cragin et al., 2010; Onsrud & Campbell, 2007) and *long-tail science* (Heidorn, 2008; Palmer, Cragin, Heidorn, & Smith, 2007). Little science projects are short-term, usually spanning a matter of months or at most a few years. They require little external funding, typically up to tens or hundreds of thousands of US dollars (Heidorn, 2008).

Little science projects are generally carried out by individuals or small teams (normally a few in number) of scientists working in a single laboratory (Chompalov, 2014). As a result, there is little role specialization: a scientist carrying out little science projects must conduct all of the steps involved in the life cycle (see Figure 1) (Wallis et al., 2007, 2008). Resultant scientific papers are usually single-authored, or involve at most a handful of co-authors (Knorr-Cetina, 1999).

The structure and organization of little science projects has important implications for the data and knowledge management practices in these projects. The data produced are generally small in volume, but may be heterogeneous in type and form (Bicarregui et al., 2013; Borgman et al., 2007; Karasti & Baker, 2008; Karasti, Baker, & Halkola, 2006). There may be little standardization of methods across the scientific domain, even to the extent that each scientist may use different tools and techniques to generate datasets similar in form and intent (Darch et al., Forthcoming).

The responsibility for data management falls to the scientists who produced the data. The data are typically managed according to localized, ad hoc standards, and usually only for the immediate purposes of the scientists (Borgman, Wallis, & Mayernik, 2012; Wallis & Borgman, 2011). As a result, data are often neglected after they are no longer needed by the scientist, and may be lost (Bicarregui et al., 2013).

3.2 Big Science

Big science projects are defined as large in scale along multiple dimensions (Chompalov, 2014; Galison & Hevly, 1992), and have been reported in a range of scientific disciplines, such as physics (Traweek, 1988), astronomy (Smith, 1992), and human genome research (Lenoir & Hays, 2000; Vermeulen, 2010). Big science projects involve funding that is often on the scale of tens or hundreds of millions or even billions of US dollars, from multiple government agencies, private benefactors or foundations (Lambright, 2008). This level of funding facilitates the construction and operation of large-scale facilities.

Big science collaborations usually comprise hundreds or thousands of members (Chompalov, 2014), and are often international undertakings (Knorr-Cetina, 1999). There is typically a high level of bureaucratization (Chompalov, 2014; Galison & Hevly, 1992). Documents governing the work and organization include formalized agreements between partner institutions, extensive reporting to funding bodies, and detailed work plans (Collins, 2003; N. Gray, Carozzi, & Woan, 2012). The division of labor involves a large degree of specialization, with a collaboration member typically focusing on a narrow range of routinized tasks (Capshew & Rader, 1992; Collins, 2003). The scale of big science collaborations and the nature of the work involved have given rise to a corporate model of authorship, in which journal articles from big science projects can involve many tens, hundreds, or even thousands of co-authors (Galison, 2003; Knorr-Cetina, 1999; Wray, 2006).

The nature of data and data practices in big science both characterize and result from these endeavors. The routinized nature of big science work facilitates the production of large volumes of homogenous data (Bicarregui et al., 2013; N. Gray et al., 2012). Data practices also tend to be routinized as official documents set out standards for the conditions under which data are to be collected, stored, managed, curated, and made accessible (Borgman, 2015; Borne, 2013).

3.3 Knowledge Infrastructures to Support Big and Little Science

Case studies of the challenges of building infrastructures for data collection, processing, management, curation, and access, tend to cast data life cycles as unfolding entirely in either a big science context or a little science context.

Those studies that focus on little science highlight various social and technical factors characteristic of little science, and discuss how these factors complicate the establishment of knowledge infrastructures. In particular, these studies discuss challenges arising from the heterogeneity of datasets, as well as the diversity of contexts in which these datasets are produced, used, and reused. Some studies focus on data life cycles within a single domain, including ecology (Baker & Millerand, 2007; Karasti & Baker, 2008), environmental science (Palmer et al., 2007), biology and biomedicine (Leonelli, 2013), and distributed network sensing (Borgman et al., 2007, 2012; Mayernik, Wallis, & Borgman, 2013). Other studies involve a range of domains in order to draw out common factors across various little science contexts (Cragin et al., 2010; Witt, Carlson, Brandt, & Cragin, 2009).

Studies that focus on data life cycles in big science pay particular attention to the volume and homogeneity of data, and relate specific challenges and opportunities for data production and management to the organization of the contexts in which data life cycles take place (Bicarregui et al., 2013). In particular, astronomy is often studied as an exemplar of big science (Borne, 2013; N. Gray et al., 2012).

However, the demarcation between big and little science may not be quite so clean in practice. Price states, "it is clear Little Science contained many elements of the grandiose. And tucked away in some academic corners, modern Big Science probably contains shoestring operations by unknown pioneers" (1963, p. 3). In a similar vein, Flannery et al. identify *facilities-based science*, defined as small-scale, short-term projects carried out by individuals and small teams using large-scale, expensive infrastructures (Bicarregui et al., 2013, p. 31; Flannery et al., 2009). One example of facilities-based science is individual astronomers using large telescope facilities to collect data in order to work on their own research questions (Smith, 1992).

Thus, scientific data life cycles may not take place solely in a big or a little science context, but instead across a mixture of contexts. Decisions taken at each step of the data life cycle have a cumulative impact (Wallis et al., 2008), and so it is critical that the design, implementation, and operation of scientific knowledge infrastructures take into account all contexts if these infrastructures are to support successfully the needs of scientists. Hence, it is important to understand the extent to which data life cycles unfold across multiple big and little science contexts, as well as the relationships between these contexts.

4 Case Studies

The above discussion motivates our four research questions:

a) What are the different contexts across which a single data life cycle unfolds?
b) How do data practices vary across these contexts?
c) How can these contexts be characterized as big or little science?
d) How do these contexts shape each other and the data practices within each context?

We address these research questions via longitudinal, qualitative case studies of data life cycles in two scientific domains, each focused around a large scientific project. The case studies and methods are introduced here.

4.1 Astronomy and the Sloan Digital Sky Survey

The Sloan Digital Sky Survey (SDSS) is a large telescope project built and operated by a consortium of hundreds of astronomers, software engineers, instrument builders, and managers, based at multiple sites in the USA. SDSS produces data that are unprecedented in astronomy in terms of scope and quality. SDSS data are made available through public *Data Releases* and have been used by astronomers globally to answer questions at the cutting edge of research.

The first phase of SDSS, *SDSS-I*, ran from 2000-2005, the second, *SDSS-II*, from 2005-2008, and subsequent SDSS projects continue today[3]. This case study focuses on SDSS-I and SDSS-II (henceforth referred to collectively as *SDSS-I/II*). The seventh, and final, Data Release of SDSS-I/II occurred in 2009 (Abazajian et al., 2009). SDSS-I/II data are some of the most used astronomy data and continue to be accessed millions of times each month (Madrid & Macchetto, 2009; Zhang et al., 2012)[4].

4.2 Deep Subseafloor Biosphere and the Center for Dark Energy Biosphere Investigations

The Center for Dark Energy Biosphere Investigations (C-DEBI) is a ten-year *National Science Foundation Science and Technology Center* (*NSF STC*) launched in 2010 (K. Edwards, 2009). The project aims to build a community of researchers to study the ecology of the deep subseafloor biosphere. C-DEBI brings microbiologists together with a variety of physical scientists, including geologists, hydrologists, and geochemists. These researchers are geographically distributed, with the Principal Investigator (PI) and four co-PIs based at five US universities distributed coast-to-coast. C-DEBI funds small projects conducted by over 100 scientists in more than 50 universities and research institutions across the USA, Europe, and Asia (Center for Dark Energy Biosphere Investigations, 2014).

[3] http://www.sdss.org/sdss-surveys/
[4] http://skyserver.sdss.org/log/en/traffic/

Scientists involved with C-DEBI pursue their scientific goals through the collection and analysis of rock samples, known as *cores*, from the seafloor. The most significant source of cores during the period of our case study are scientific ocean drilling cruises, or expeditions, that were conducted by the *Integrated Ocean Drilling Program* (*IODP*), which ran from 2003-2013[5].

4.3 **Studying Big and Little Science, and Data Practices**

Our deep subseafloor biosphere and astronomy case studies provide ideal opportunities to address our research questions. At first sight, astronomy might be regarded as an exemplar of big science, while the deep subseafloor biosphere might be regarded as an exemplar of small science. Together, our two case studies initially appear to exemplify the dichotomy of big science and little science, and thus are suitable for exploring the extent to which the dichotomy holds.

5 Methods

We present selected findings from an eighteen-month study of scientists affiliated with C-DEBI, and a five-year study of astronomy and SDSS. These case studies include participant observation, semi-structured interviews, and document analysis.

5.1 **Astronomy and SDSS**

Our interview sample for astronomy and SDSS comprises 118 interviews with SDSS collaboration team members and external users of SDSS-I/II data. Interviews ranged from 45 to 150 minutes. We conducted participant observation for 45 days across five key SDSS sites. We also assembled a corpus of documents for analysis, including journal articles, draft papers, reports, project plans, policy documents, websites, and other official documents.

5.2 **Deep Subseafloor Biosphere and C-DEBI**

Our interview sample for the deep subseafloor biosphere and C-DEBI comprises 49 people, including C-DEBI-affiliated scientists, IODP curators, and managerial staff. Interviews ranged in length from 35 to 150 minutes. We were embedded for eight months in a laboratory headed by a leading C-DEBI figure at a large US research university, conducted week-long observational work in two other laboratories, and attended and participated in scientific meetings (Darch & Cummings, 2013). We have also assembled a corpus of documents including official documents from C-DEBI, IODP, and NSF.

6 Findings

Data life cycles in both case studies unfold across multiple contexts. In the case of SDSS-I/II and astronomy, data were produced and managed by the SDSS-I/II collaboration itself, and are subsequently used and processed by individuals and small teams of astronomers in many institutions worldwide. In the deep subseafloor biosphere and C-DEBI case study, the data life cycle began on scientific ocean drilling expeditions with the collection and processing of cores. Data and cores from the expeditions are then analyzed by individuals and small teams, globally distributed across a wide range of laboratories.

6.1 **Astronomy and SDSS**

The life cycle involving SDSS-I/II data occurs across two distinct contexts. In the first context, the SDSS-I/II collaboration itself collected, processed, and released the data. Subsequently, individual and small groups of astronomers retrieve, process further, analyze, and use these data to conduct research at the cutting edge of astronomy. The SDSS-I/II findings below are summarized in Figure 2.

6.1.1 **Production of data within SDSS-I/II**

SDSS-I/II comprised hundreds of researchers internationally and included 25 member organizations. The hundreds of collaboration members varied some over the years, with the final Data Release authored by 204 individuals (Abazajian et al., 2009).

[5] http://www.iodp.org/history

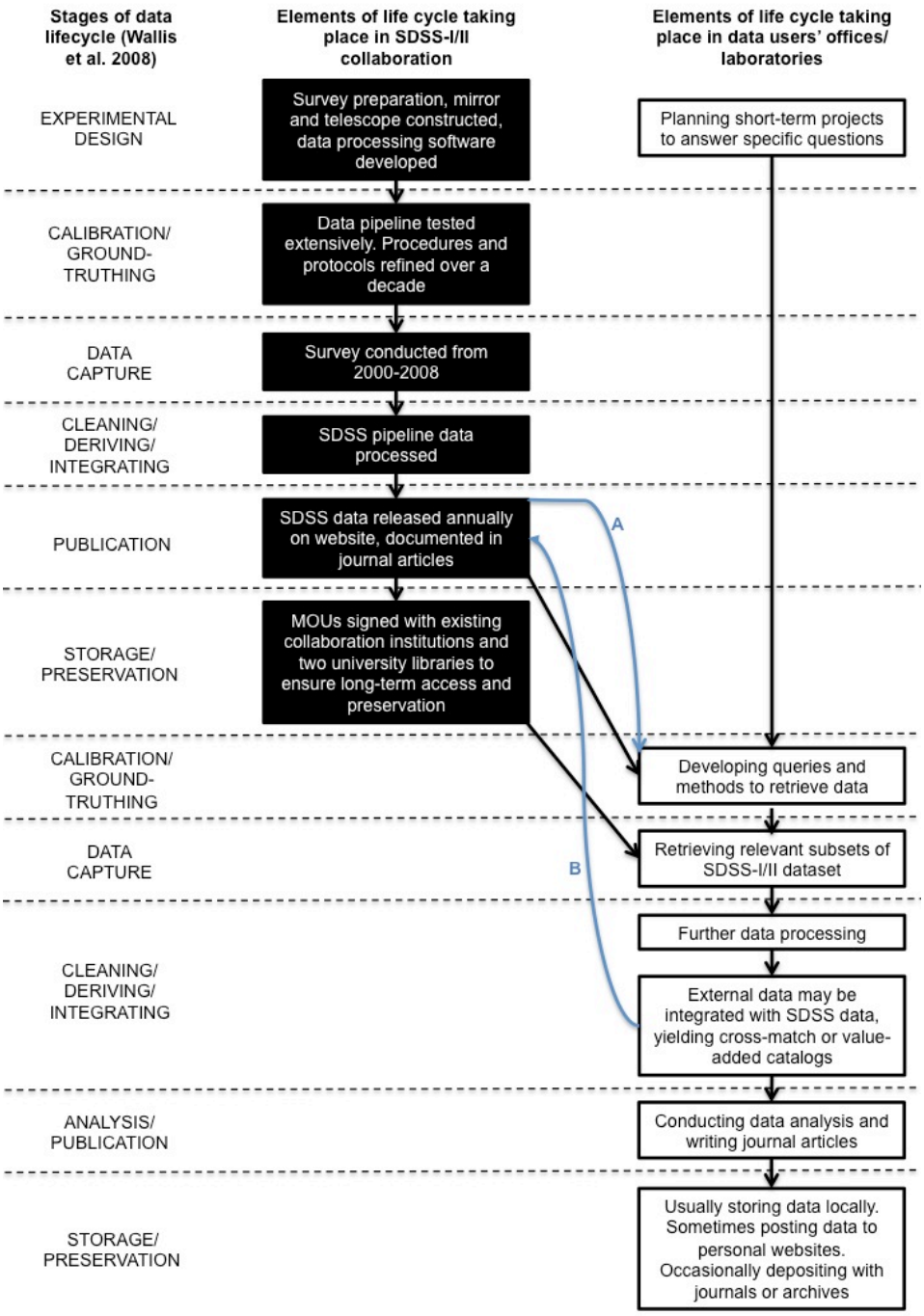

Figure 2: Data life cycle for SDSS-I/II split across the SDSS-I/II collaboration and SDSS data users' offices and laboratories, and how it relates to the data life cycle model of Wallis et al. (2008)

The straight, black arrows denote the flow of work.
The wavy, blue arrows denote some instances where the context of the SDSS-I/II collaboration shapes data users' offices or laboratories, or vice versa:

A The scale of SDSS data promotes the adoption or development of new tools, and learning of new skills by data users

B Cross-match or value-added catalogs developed by data users may sometimes be incorporated into subsequent SDSS data releases

Planning for SDSS-I/II began in the late 1980s (Finkbeiner, 2010). Survey preparation, mirror and telescope construction, and data processing software development, took another decade to design and construct. The data collection survey was conducted from 2000-2008.

SDSS received hundreds of millions of US dollars from multiple sources, including core funding from the Alfred P. Sloan Foundation[6]. Funding also came from the National Science Foundation, the U.S. Department of Energy, and the participating institutions. The significant financial investment and size of the project resulted in a plethora of documentation and agreements to govern the collaboration, including *Principles of Operation*, public and private *Memoranda of Understanding* (*MOUs*) which define terms of collaboration between partners, *Publication Policies*, and other written agreements (Astrophysical Research Consortium, 2000; "SDSS Scientific and Technical Publication Policy," 2003).

Dedicated staff and facilities supported SDSS-I/II. The project governance included a director, board, working groups, and advisory council (Astrophysical Research Consortium, 2000). The survey was carried out using a dedicated 2.5-meter telescope at Apache Point Observatory in New Mexico (Gunn et al., 2006).

The SDSS-I/II dataset is a large, complex aggregation of materials representing multiple elements of the international project. This dataset includes four kinds of data, namely a *Photometric Catalog*, a *Spectroscopic Catalog*, images, and spectra (Szalay, Kunszt, Thakar, & Gray, 1999). In total, the SDSS-I/II archive forms a collection between 100 and 200 terabytes in size (Sands, Borgman, Traweek, & Wynholds, 2014).

SDSS-I/II astronomers and computer scientists spent more than a decade constructing data management procedures and protocols in order to ensure a standardized and consistent SDSS-I/II data product over time and across multiple Data Releases. Our interviews show that data management involved many steps beginning with collection of raw data by the instruments, software pipeline processing, data distribution, and project archiving. The construction of the data processing pipeline was a critical aspect of the entire project and accounted for approximately 25% of the survey's total "cost and effort" (J. Gray et al., 2002, p. 2).

6.1.2 Uses of SDSS-I/II data by individuals and small teams

Internationally, many astronomers continue to make use of the publicly available SDSS-I/II data. These astronomers generally work with the data alone, or in collaboration with one or two other senior researchers, graduate students, or postdoctoral researchers. Our interviews reveal that these astronomers use the data to investigate their own research questions distinct from the SDSS I/II collaboration. The research practices typically involve retrieving relevant subsets of the large SDSS-I/II dataset and processing these data. Many astronomers we interviewed also combine SDSS-I/II data with data from other sources, such as surveys from different wavelengths, optical data, and spectroscopic findings from personal data collections. For example, one of our interviewees has created a *cross-match catalog* combining SDSS-I/II data, infrared data from a different project, and their own personal source lists. The resultant datasets are heterogeneous as they contain data from multiple sources, developed on distinct instruments.

Cross-match catalogs are an example of processed datasets, distinct from the primary SDSS-I/II dataset. Our interviews uncovered a pattern to the way these processed datasets are managed. They are usually created and stored on university computer networks or personal computers. The enhanced datasets are rarely archived. Much of the derived data made by individuals and small groups cannot be found or easily re-created, even after only one or two years, and even by the researchers who created them. Much research data are lost as hard-drives and laptops are replaced, website URLs become out of date, and graduate students move on to careers in new locations.

6.1.3 Relationships between SDSS-I/II collaboration and data users

The volume of the SDSS-I/II data impacts the research contexts of individuals and small teams of users, for example in terms of the tools and methods they use to obtain, process, manage, and analyze data. These astronomers have to discover or create tools in order to manage the SDSS-I/II data and develop expertise in databases. One interviewee explained the importance of such a tool to enable "working with tabular and catalog data, for doing plots in selected columns and subset-cutting…" (SDSS Data User 1). Another interviewee described having to improve their own expertise in databases:

> "There are hurdles I would say between the typical scientist and using a database in their research. Right now you really have to understand databases more than you want to as a researcher" (SDSS Data User 2).

[6] http://www.sloan.org/about-the-foundation/

Even for those familiar with databases, existing tools and techniques do not always scale to the size of the SDSS data. One interviewee explained how they instead had to write their own code:

> "Because we had [a] hundred million objects we couldn't do it with those tools. So what we did was to download Sloan data [and a different project's data] on our local disks here, and write a piece of code… and then match the two" (SDSS Data User 3).

Conversely, the research conducted by SDSS-I/II data users can also impact the SDSS-I/II dataset itself over time. Some of our interviewees have created *value-added catalogs*, which are supplemental datasets that build on subsets of the larger SDSS-I/II dataset. SDSS data users can devote an extensive amount of resources to developing these catalogs, which often includes spending many years refining algorithms. These tasks involve a huge amount of manual work, perfecting specific components of the data. Once the catalog is finalized and released, the data are often incorporated into to the SDSS-I/II dataset itself and made available to other astronomers through SDSS infrastructure.

The large volume of SDSS-I/II data provides an example of how the SDSS-I/II collaboration context impacts individuals' research contexts. Value-added catalogs illustrate how the work contexts of individuals and small teams can impact the larger SDSS infrastructure.

6.2 Deep Subseafloor Biosphere and C-DEBI

The data life cycle in our deep subseafloor biosphere case study takes place across two contexts. The first are IODP expeditions, where cores were collected. The second comprises various onshore laboratories where C-DEBI-funded scientists analyze these cores to characterize subseafloor microbial communities and the environments they inhabit. Figure 3 summarizes the deep subseafloor biosphere findings presented below.

6.2.1 Production of knowledge products on IODP expeditions

Cores were collected on IODP expeditions. IODP operated from 2003-2013, and was preceded by the *Deep Sea Drilling Program* (*DSDP*) and *Ocean Drilling Program* (*ODP*), which ran from 1968-1983 and 1983-2003 respectively. IODP has been succeeded by the *International Ocean Discovery Program* (henceforth referred to as *IODP 2*)[7]. IODP brought together scientists from 26 countries and 12 scientific disciplines, including microbiology and a range of physical sciences. MOUs were signed between various countries that governed the precise terms of each nation's involvement (including financial contributions and the amount of space allocated to each country on expeditions).

IODP expeditions were conducted on one of two ocean-drilling cruise ships, both hundreds of meters in length and expensive to build and operate. For example, one of these ships, the *Chikyū*, cost 60 billion yen (approximately 600 million US dollars) to build in 2001-02 (Liu, Wang, & Zhou, 2004).

Each expedition would visit one site in the ocean and have specific scientific objectives. The focus and location of a particular expedition was determined following a competitive proposal process. Each expedition was governed by its *Science Prospectus*, a document usually around 100 pages in length detailing the objectives of the expedition and how they were to be accomplished (for example, D'Hondt, Inagaki, & Alvarez Zarikian (2010)).

Scientists would apply to sail on each expedition. Typically, an expedition comprised 25-30 scientists, and one IODP scientist explained to us how they applied formal criteria, codified in official documents, for selecting the applicants who would sail.

An expedition's *Sampling Plan* determined how cores were to be allocated amongst expedition participants to take back to their own onshore laboratories. In particular, the Plan would set out how to allocate samples for microbiological and for physical science purposes to avoid biological contamination:

> "Sample planning is completely different for microbiology, so we have to have special tools, we have to have sterilized tools, sterilized syringes, sterilized bags" (IODP Personnel 1)

Cores were subjected to a wide range of analyses of physical properties onboard the ships: these analyses were consistent across all expeditions, conducted according to standardized procedures. Resultant data, accompanied by rich metadata, were stored in the publicly accessible IODP database[8].

The conduct of work on each expedition involved a significant amount of role specialization. IODP curators oversaw core processing and data production. Expedition scientists occupied very specific roles for the entire expedition. IODP cruise personnel explained to us how this division of labor enabled high-throughput processing of cores and data collection.

[7] http://iodp.org/
[8] http://iodp.tamu.edu/database/

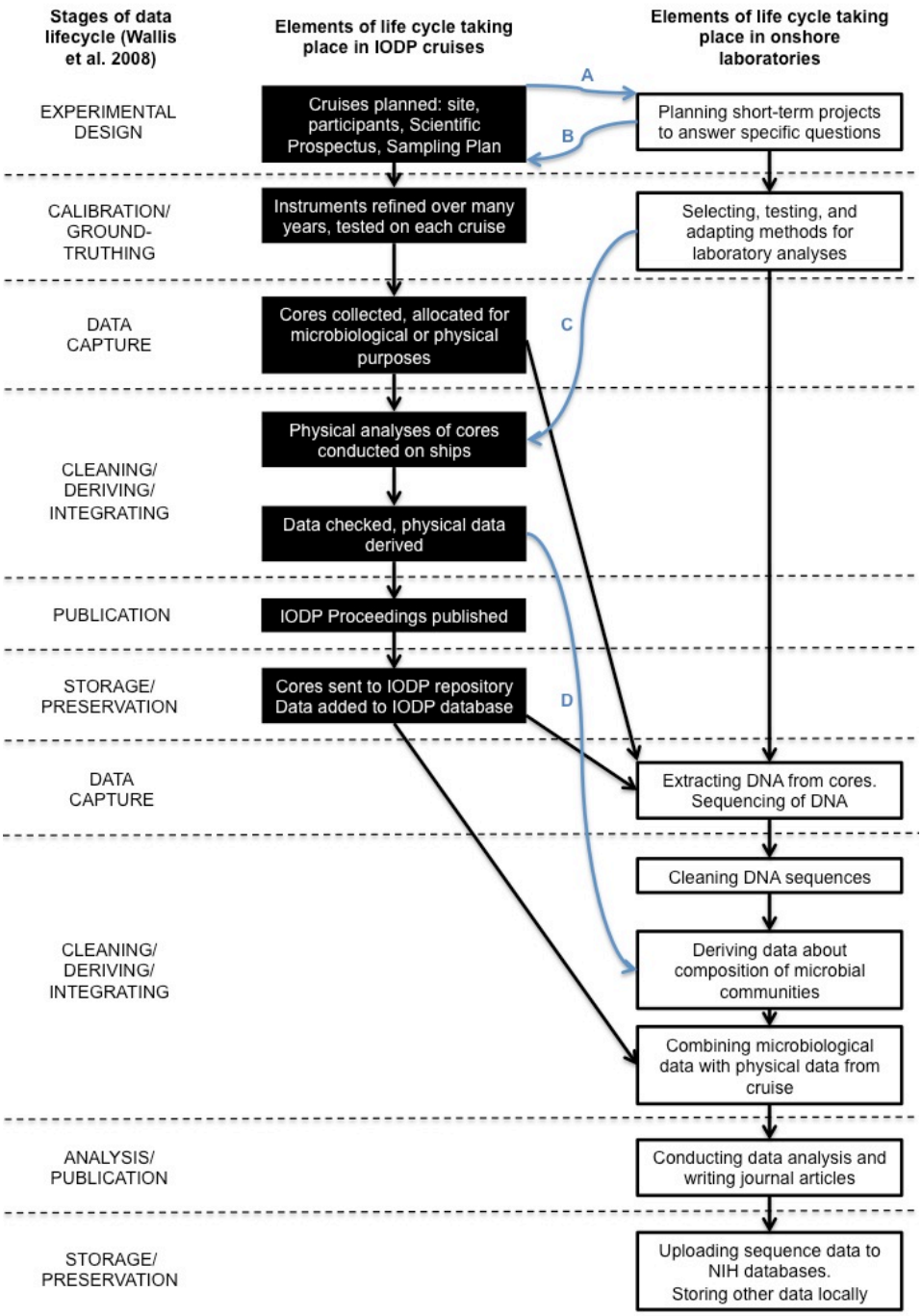

Figure 3: Data life cycle for C-DEBI split across IODP cruises and onshore laboratories, and how it relates to the data life cycle model of Wallis et al. (2008)

The straight, black arrows denote the flow of work.
The wavy, blue arrows denote some instances where the context of the onshore laboratory shapes the context of IODP cruises, or vice versa:

A Individual and small group plans are constrained by the cruise plans
B Cruise plans are comprised of multiple individual and small group plans
C Heterogeneity of methods means no microbiological data were collected on IODP cruises
D The lack of microbiological data collected on cruises means microbiological analyses must be carried out in onshore laboratories, reducing the time available for more advanced analyses

After expedition completion, and once cores had been allocated to scientists, remaining core samples were distributed to one of three IODP core repositories located in the USA, Germany, and Japan, where they continue to be stored (Committee on the Review of the Scientific Accomplishments and Assessment of the Potential for Future Transformative Discoveries with U.S.-Supported Scientific Ocean Drilling, 2012). For each cruise, a volume of *Proceedings of the Integrated Ocean Drilling Program* would be published, detailing what samples were collected and what data were generated onboard a particular expedition. All scientists participating in an expedition would be co-authors of this volume (for example, see D'Hondt, Inagaki, Alvarez Zarikian & The Expedition 329 Scientists (2011)).

6.2.2 Uses of IODP knowledge products by individuals and small teams

C-DEBI funds individuals and small teams (two to three) of scientists to work on short-term projects (typically one to three years in length) to process cores from IODP expeditions in onshore laboratories. The funding is relatively small-scale, usually tens of thousands of dollars per project (Center for Dark Energy Biosphere Investigations, 2010).

The aim of many of these projects is to characterize the microbial communities found in the cores. To do so, scientists produce and use a range of physical and biological datasets. One example of a particularly important biological workflow is determining the composition and function of a microbial community. A scientist first extracts DNA from microbial cells in the rock, and then quantifies and sequences these nucleic acids. Sequences are then subjected to a range of computationally-intensive analyses. Our laboratory-based observations reveal a significant degree of heterogeneity across C-DEBI scientists at almost all stages of the process regarding the techniques and tools used (Darch et al., Forthcoming). Sometimes, data are also combined with data from other sources, such as physical science data from IODP cruises, to form new, hybrid datasets.

We observed recordkeeping practices in the laboratory varying from scientist to scientist in terms of the granularity and types of detail recorded in laboratory notebooks. One of our interviewees compared their notebook practices to a colleague's:

> "We both definitely have different notebook styles…he's very comfortable writing a few things that he knows are necessary… I'm really explicit with everything I do in the notebook" (C-DEBI Scientist 1)

Finally, the curation of laboratory-generated datasets can also be highly variable. Genetic sequence data that support the conclusions of published journal articles must be uploaded to a publicly accessible NIH-operated database, such as GenBank[9]. However, there is no such requirement for other biological and hybrid datasets.

As a result, the management of these other data is generally localized and ad hoc, stored on a scientist's computer or the laboratory server only for as long as they are useful to the immediate needs of the particular project for which they were created. Often, other scientists are not aware of these data; as a result, these data can be lost when scientists leave the laboratory, or move into other domains of study.

6.2.3 Relationships between IODP expeditions and onshore laboratories

The data life cycle detailed above started on IODP expeditions and is completed in multiple onshore laboratories. These two contexts, however, are not independent of each other but influence and shape one another.

For example, there is interplay between the processes of planning projects in onshore laboratories and of planning cruises. On the one hand, the former is constrained by the latter as scientists take into account what samples they are likely to procure from a particular cruise when devising research projects. On the other hand, cruise plans are shaped by what analyses individual scientists aspire to carry out in their onshore laboratories. These scientists propose sites and scientific objectives for cruises, apply to sail on cruises, and negotiate allocation of samples, all in order to pursue their own particular research questions.

Another example of the interplay between the two contexts relates to the heterogeneity of methods across onshore laboratories used to generate microbiological data from cores. This heterogeneity has had a number of implications for the conduct of IODP expeditions. One implication is that, in contrast to physical science data, very little microbiological data were routinely collected aboard IODP expeditions, and thus IODP databases do not contain microbiological data. In turn, in onshore

[9] http://www.ncbi.nlm.nih.gov/genbank/

laboratories, scientists often have to perform time-consuming basic analyses themselves, holding back their research progress relative to other disciplines involved with IODP. An interviewee explained:

> "As the biologist, we have…to now process all of our samples, do all the sequence analysis, do the bulk labor of all of our work on the equipment that we already have to have in our lab versus what everybody else is doing on the ship" (C-DEBI Scientist 3)

A second implication is that microbiologists were often unwilling to participate in IODP expeditions because microbiologists would have little opportunity to conduct work on expeditions. As a result, a number of IODP expeditions had no microbiologist onboard. In negotiations about core allocation, the absence of a microbiologist meant no one advocated for the interests of microbiologists, in turn restricting the quantity and quality of samples available for microbiological analysis in onshore laboratories. This is explained by one of our interviewees:

> "I requested 150 samples…Due to limitations and then lack of lobbying 'cause I was not out at sea and there was no microbiologist out at sea, I only got about 50 samples of my 150…And there were no contamination checks done on any of the cores, so the integrity of each one of my samples is then questioned" (C-DEBI Scientist 3)

In order to improve the quality and quantity of cruise data and cores for microbiologists in the context of IODP 2, an international group of senior deep subseafloor biosphere scientists are advocating for standardization across the research community of methods for producing microbiological data (Orcutt et al., 2013). These scientists secured funding for a weeklong workshop, which was carried out in Seoul in August 2014[10]. The aim of this workshop was to produce recommendations about method standardization. In other words, in order to reconfigure data and sample collection practices onboard IODP 2 expeditions, data practices of deep subseafloor biosphere researchers in their onshore laboratories are now in the process of being reconfigured, providing another example of the mutual shaping of the two contexts.

7 Discussion

Studies of scientific domains and projects often follow the big/little science dichotomy (Price, 1963), usually characterizing the scientific domain under study as either big science (Galison & Hevly, 1992; Vermeulen, 2010) or little science (Borgman et al., 2007; Cragin et al., 2010; Heidorn, 2008). Our study challenges this dichotomy. We have shown here that a data life cycle often does not unfold solely in a big or a little science context, but across both. Further, these big and little science contexts can shape one another in multiple ways. These considerations have important implications for the design of knowledge infrastructures across the data life cycle.

7.1 **Big Science: SDSS and IODP**

Both SDSS-I/II and IODP exhibited many features that characterize big science projects (Galison & Hevly, 1992). They both received large quantities of funding from federal agencies (Lambright, 2008). Both organizations brought together consortia of many universities and private institutions, crossing international boundaries (Knorr-Cetina, 1999). SDSS-I/II and IODP were long-term undertakings on the order of decades (Bicarregui et al., 2013).

Another big science feature exhibited by both organizations is that of large facilities (McCray, 2000; Pestre & Krige, 1992). The work of SDSS-I/II focused on a massive telescope. IODP involved two ships and three core repositories. In both cases, work regarding these facilities were governed by significant bureaucratization (Chompalov, 2014; Collins, 2003; N. Gray et al., 2012; Hevly, 1992). For instance, SDSS-I/II was governed by documents including Principles of Operation, MOUs, and Publication Policies. IODP expeditions were governed by MOUs, a Sampling Plan, a Science Plan, multiple criteria for choosing expedition participants, and procedures for processing and managing cores and data.

The organization of work conducted in these large facilities has also exemplified many features of big science. SDSS-I/II involved hundreds of personnel. IODP employed dozens of staff and also involved nearly 1000 scientists (Chompalov, 2014). These personnel have usually occupied very specialized roles, with specific routinized tasks (Capshew & Rader, 1992; Collins, 2003). The number of scientific personnel involved impacted the authorship of publications detailing work conducted in the contexts of SDSS-I/II

[10] http://www.ecord.org/pdf/Biosphere-Paleoclimate_flyer.pdf

(Data Releases) and IODP (Post-Cruise Reports), echoing corporate authorship practices typical of big science (Galison, 2003; Knorr-Cetina, 1999; Wray, 2006).

The knowledge products of both SDSS-I/II and IODP are also characteristic of big science. SDSS-I/II datasets are large in scale and homogeneous (Borgman, 2015; N. Gray et al., 2012), managed by a specialized staff (Bicarregui et al., 2013) in a highly standardized manner according to codified standards (Borne, 2013), and made publically accessible through a coordinated infrastructure (Borgman, 2015). Although the data and cores produced onboard IODP expeditions were much smaller in volume and more heterogeneous in terms of the disciplines they covered, the scale was nevertheless very large relative to the deep subseafloor biosphere domain. Furthermore, these cores and data were produced and processed according to well-defined standards and overseen by IODP curators. Finally, a coordinated infrastructure has facilitated the management and accessibility of cores and data produced onboard IODP expeditions.

7.2 Little Science: Astronomers and C-DEBI

As exemplars of facilities-based science, SDSS-I/II and IODP expeditions were large infrastructures enabling the conduct of smaller-scale science (Flannery et al., 2009). Data life cycles begun in the context of SDSS-I/II and IODP subsequently moved into contexts that exemplify little science (see Figures 2 and 3).

The knowledge products of both IODP expeditions and SDSS-I/II projects are often used in ways characteristic of little science. Large numbers of disparate individuals and small teams of scientists conduct projects using these knowledge products. We observed astronomers using SDSS-I/II data and C-DEBI-affiliated scientists using cores and data collected on IODP expeditions (Chompalov, 2014). These projects require little external funding and last a matter of months or (at most) a year or two (Heidorn, 2008).

The division of labor in these projects is minimal, with a scientist often conducting all stages of the project in their laboratory or office, from initial planning to analysis and dissemination of results (Wallis et al., 2007, 2008). Results are disseminated in papers that are either single-authored or have at most a handful of co-authors (Knorr-Cetina, 1999).

The form and management of the data produced in these contexts exhibit many characteristics of little data. Astronomers use a variety of tools to transform portions of SDSS-I/II data into new datasets. They also increase the heterogeneity of datasets by combining SDSS-I/II data with data from other sources. C-DEBI-affiliated scientists employ multiple techniques, even when producing datasets similar in form and intent, and combine microbiological data with physical science data to create heterogeneous datasets (Bicarregui et al., 2013; Borgman et al., 2007; Karasti & Baker, 2008; Karasti et al., 2006).

In the cases of both astronomers and C-DEBI scientists, data management is the responsibility of individuals or small teams. They use localized, ad hoc standards and manage data for as long as it is immediately useful, as is typically found in little science (Borgman et al., 2012).

7.3 Implications for the Design of Knowledge Infrastructures

Understanding all stages of the data life cycle is critical for the design of knowledge infrastructures to support data management (P. N. Edwards et al., 2013; Wallis et al., 2008). Our findings challenge the assumption embedded in the design of many knowledge infrastructures that data life cycles take place in contexts that primarily exhibit features characteristic of only little science (Borgman et al., 2007; Cragin et al., 2010; Karasti & Baker, 2008; Leonelli, 2013; Palmer et al., 2007), or characteristic of only big science (Bicarregui et al., 2013; Borne, 2013; N. Gray et al., 2012).

The building of knowledge infrastructures should take into account the differences between big and little science contexts across the same life cycle. For instance, the same dataset may be produced and managed according to well-established standards in a big science context, but subject to more localized, ad hoc management in a little science context. Data or other knowledge products can be large in volume and homogeneous in a big science context, and be transformed into smaller, more heterogeneous datasets in a little science context. Finally, the same individuals may play very different roles in different contexts, and therefore have distinct relationships to the data across these contexts. For instance, in a big science context, a scientist may be assigned a very narrow specialized role in the overall production and processing of data, while they may carry out all steps of data processing and analysis in a little science context.

The big and little science contexts discussed above are not static and independent of each other; they dynamically shape each other over time. For instance, the volume of data from SDSS-I/II (and subsequent iterations of SDSS) is driving changes in the tools, techniques, and expertise of astronomers

in their local work. We witnessed how the challenges and opportunities of ocean-drilling cruises are driving scientists to standardize methods in onshore laboratories, in turn to change practices on these cruises. Awareness of these relationships is critical to the building of knowledge infrastructures to support data practices.

8 Conclusions and Future Work

We have challenged the demarcation of boundaries between big and little science in the study of scientific data practices. At first sight, our astronomy case study appears to exemplify big science, while deep subseafloor biosphere research appears to exemplify little science. However, closer examination reveals that some important stages of the data life cycles in each domain unfold in big science contexts while others occur in little science contexts. Furthermore, the big and little science contexts shape each other over time. The entangled presence of both big and little science affects the planning of knowledge infrastructures to support all stages of scientists' data life cycles.

However, the accounts given in this paper are necessarily only partial. There are likely to be many other examples of relationships between big and little science contexts. For instance, the authors of this paper have recently commenced a case study of the building of the Large Synoptic Survey Telescope (LSST), a large telescope project currently in its construction phase (Connolly, 2014; Ivezic et al., 2011). We are observing how the project incorporates the anticipated needs of future users of LSST data into decisions about the construction of LSST infrastructure (LSST Science Collaboration et al., 2009). Additionally, features of the telescope's camera are being tested in a variety of little science contexts conducted by small teams of researchers; results of these investigations will be incorporated into the development of LSST infrastructure (Tyson et al., 2014).

Furthermore, there are clearly differences in scale along many dimensions (such as funding, datasets, and participants) between the various contexts we characterized here as big science. Similarly, there are many differences between those contexts we characterized as little science. Future work that breaks down the big/little science binary further by exploring the multiple scales at which science is conducted will allow for better and richer characterizations of data practices in contemporary science.

References

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., Prieto, C. A., An, D., … Zucker, D. B. (2009). The Seventh Data Release of the Sloan Digital Sky Survey. *The Astrophysical Journal Supplement Series*, *182*(2), 543–558. doi:10.1088/0067-0049/182/2/543

Astrophysical Research Consortium. (2000). *Principles of Operation for the Sloan Digital Sky Survey*. Retrieved from http://www.sdss.org/policies/sdss_poo.html

Baker, K. S., & Millerand, F. (2007). Scientific Infrastructure Design: Information Environments and Knowledge Provinces. *Proceedings of the American Society for Information Science and Technology*, *44*(1), 1–9. doi:10.1002/meet.1450440370

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

Bicarregui, J., Gray, N., Henderson, R., Jones, R., Lambert, S., & Matthews, B. (2013). Data Management and Preservation Planning for Big Science. *International Journal of Digital Curation*, *8*(1), 29–41. doi:10.2218/ijdc.v8i1.247

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

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge MA: MIT Press.

Borgman, C. L., Wallis, J. C., & Enyedy, N. D. (2007). Little Science Confronts the Data Deluge: Habitat Ecology, Embedded Sensor Networks, and Digital Libraries. *International Journal on Digital Libraries*, *7*(1-2), 17–30. doi:10.1007/s00799-007-0022-9

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

Borne, K. D. (2013). Virtual Observatories, Data Mining, and Astroinformatics. In T. D. Oswalt & H. E. Bond (Eds.), *Planets, Stars and Stellar Systems* (pp. 403–443). Springer Netherlands. Retrieved from http://link.springer.com/referenceworkentry/10.1007/978-94-007-5618-2_9

Capshew, J. H., & Rader, K. A. (1992). Big Science: Price to the Present. *Osiris*, *7*, 2–25. Retrieved from http://www.jstor.org/stable/301765

Center for Dark Energy Biosphere Investigations. (2010). *C-DEBI Strategic Implementation Plan, 2010-2015*. Retrieved from http://www.darkenergybiosphere.org/internal/docs/C-DEBI_SIP_2010Sep.pdf

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

Chompalov, I. (2014). Lessons Learned from the Study of Multi-organizational Collaborations in Science and Implications for the Role of the University in the 21st Century. In M. Herbst (Ed.), *The Institution of Science and the Science of Institutions* (pp. 167–184). Springer Netherlands. Retrieved from http://link.springer.com/chapter/10.1007/978-94-007-7407-0_9

Collins, H. M. (2003). LIGO Becomes Big Science. *Historical Studies in the Physical and Biological Sciences*, *33*(2), 261–297. doi:10.1525/hsps.2003.33.2.261

Committee on the Review of the Scientific Accomplishments and Assessment of the Potential for Future Transformative Discoveries with U.S.-Supported Scientific Ocean Drilling. (2012). *Scientific Ocean Drilling: Accomplishments and Challenges*. Washington, D.C.: National Academies Press.

Connolly, A. (2014). LSST Data Management: Prospects for Processing and Archiving Massive Astronomical Data Sets. Retrieved from http://www.lsst.org/files/docs/lsst_data_man_prospects.pdf

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data Sharing, Small Science, and Institutional Repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023–4038. doi:10.1098/rsta.2010.0165

Darch, P. T., Borgman, C. L., Traweek, S., Cummings, R. L., Wallis, J. C., & Sands, A. E. (Forthcoming). What Lies Beneath?: Knowledge Infrastructures in the Subseafloor Biosphere and Beyond. *International Journal on Digital Libraries*.

Darch, P. T., & Cummings, R. L. (2013, December 9). *Buried Deep: How Data About Subseafloor Life Becomes Dark and Why*. Presented at the American Geophysical Union 46th Annual Fall Meeting, San Francisco, CA. Retrieved from http://www.academia.edu/7427702/Buried_deep_How_data_about_subseafloor_life_becomes_dark_and_why

D'Hondt, S., Inagaki, F., & Alvarez Zarikian, C. (2010). *South Pacific Gyre Microbiology* (Vol. 329). Retrieved from http://publications.iodp.org/scientific_prospectus/329/index.html

D'Hondt, S., Inagaki, F., Alvarez Zarikian, C., & The Expedition 329 Scientists. (2011). *South Pacific Gyre Subseafloor Life* (Vol. 329). Tokyo: Integrated Ocean Drilling Program Management International, Inc. Retrieved from http://publications.iodp.org/proceedings/329/329title.htm

Edwards, K. (2009). *Center for Dark Energy Biosphere Investigations (C-DEBI): A Center for Resolving the Extent, Function, Dynamics and Implications of the Subseafloor Biosphere*. Retrieved from http://www.darkenergybiosphere.org/internal/docs/2009C-DEBI_FullProposal.pdf

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

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

Finkbeiner, A. K. (2010). *A Grand and Bold Thing: the Extraordinary New Map of the Universe Ushering in a New Era of Discovery*. New York: Free Press.

Flannery, D., Matthews, B., Griffin, T., Bicarregui, J., Gleaves, M., Lerusse, L., … Kleese, K. (2009). ICAT: Integrating Data Infrastructure for Facilities Based Science. In *Fifth IEEE International Conference on e-Science, 2009. e-Science '09* (pp. 201–207). doi:10.1109/e-Science.2009.36

Furner, J. (2003a). Little Book, Big Book: Before and After Little Science, Big Science: A Review Article, Part I. *Journal of Librarianship and Information Science*, *35*(2), 115–125. doi:10.1177/0961000603352006

Furner, J. (2003b). Little Book, Big Book: Before and After Little Science, Big Science: A Review Article, Part II. *Journal of Librarianship and Information Science*, *35*(3), 189–201. doi:10.1177/0961000603353006

Galison, P. (2003). The Collective Author. *Scientific Authorship: Credit and Intellectual Property in Science*, 325–355.

Galison, P., & Hevly, B. W. (1992). *Big Science: The Growth of Large-Scale Research*. Stanford, Calif.: Stanford University Press.

Gray, J., Slutz, D., Szalay, A. S., Thakar, A. R., vandenBerg, J., V, J., … Slutz, D. (2002). *Data Mining the SDSS SkyServer Database* (No. MSR-TR-2002-01) (pp. 189–210). Retrieved from http://inspirehep.net/record/609525?ln=en

Gray, N., Carozzi, T. D., & Woan, G. (2012). Managing Research Data in Big Science. *arXiv:1207.3923*. Retrieved from http://www.astro.gla.ac.uk/users/norman/projects/mrd-gw/report.html

Greenberg, J. (2009). Theoretical Considerations of Lifecycle Modeling: An Analysis of the Dryad Repository Demonstrating Automatic Metadata Propagation, Inheritance, and Value System Adoption. *Cataloging & Classification Quarterly*, *47*(3-4), 380–402. doi:10.1080/01639370902737547

Gunn, J. E., Siegmund, W. A., Mannery, E. J., Owen, R. E., Hull, C. L., Leger, R. F., … Wang, S. (2006). The 2.5 m Telescope of the Sloan Digital Sky Survey. *Astronomical Journal*, *131*, 2332–2359.

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, *57*(2), 280–299. doi:10.1353/lib.0.0036

Hevly, B. W. (1992). Reflections on Big Science and Big History. *Big Science: The Growth of Large-Scale Research*, 355–363.

Hey, A. J. G., Tansley, S., & Tolle, K. (Eds.). (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/contents.aspx

International Ocean Discovery Program. (2014). Retrieved June 13, 2014, from http://iodp.org/

Ivezic, Z., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., AlSayyad, Y., … Collaboration, for the L. (2011, June 7). LSST: from Science Drivers to Reference Design and Anticipated Data Products (Version 2.0). Retrieved from http://arxiv.org/abs/0805.2366

Karasti, H., & Baker, K. S. (2008). Digital Data Practices and the Long Term Ecological Research Program Growing Global. *International Journal of Digital Curation*, *3*(2), 42–58. doi:10.2218/ijdc.v3i2.57

Karasti, H., Baker, K. S., & Halkola, E. (2006). Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network. *Journal of Computer-Supported Cooperative Work*, *15*(4), 321–358. doi:10.1007/s10606-006-9023-2

Knorr-Cetina, K. (1999). *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge Mass.: Harvard University Press.

Lambright, W. H. (2008). Government and Science: A Troubled, Critical Relationship and What Can Be Done about It. *Public Administration Review*, *68*(1), 5–18. doi:10.1111/j.1540-6210.2007.00830.x

Lenoir, T., & Hays, M. (2000). Manhattan Project for Biomedicine. In *Controlling Our Destinies: Historical, Philosophical, Ethical, and Theological Perspectives on the Human Genome Project* (pp. 19–46). South Bend Indiana: University of Notre Dame Press.

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

Liu, X., Wang, Q., & Zhou, Z. (2004). IODP in Japan. *Advance in Earth Sciences*, *4*, 010.

LSST Science Collaboration, Abell, P. A., Allison, J., Anderson, S. F., Andrew, J. R., Angel, J. R. P., … Zhan, H. (2009). *LSST Science Book, Version 2.0* (arXiv e-print). Retrieved from http://www.lsst.org/lsst/scibook

Madrid, J. P., & Macchetto, F. D. (2009). High-Impact Astronomical Observatories. *Bulletin of the American Astronomical Society.*, *38*(4), 1286.

Mayernik, M. S., Wallis, J. C., & Borgman, C. L. (2013). Unearthing the Infrastructure: Humans and Sensors in Field-Based Research. *Computer Supported Cooperative Work*, *22*(1), 65–101. doi:10.1007/s10606-012-9178-y

McCray, W. P. (2000). Large Telescopes and the Moral Economy of Recent Astronomy. *Social Studies of Science*, *30*(5), 685–711. doi:10.1177/030631200030005002

Onsrud, H., & Campbell, J. (2007). Big Opportunities in Access to “Small Science” Data. *Data Science Journal*, *6*, OD58–OD66. doi:10.2481/dsj.6.OD58

Orcutt, B. N., LaRowe, D. E., Biddle, J. F., Colwell, F. S., Glazer, B. T., Reese, B. K., … Wheat, C. G. (2013). Microbial Activity in the Marine Deep Biosphere: Progress and Prospects. *Extreme Microbiology*, *4*, 189. doi:10.3389/fmicb.2013.00189

Palmer, C. L., Cragin, M. H., Heidorn, P. B., & Smith, L. C. (2007). Data Curation for the Long Tail of Science: The Case of Environmental Sciences. Presented at the 3rd International Digital Curation Conference, Washington, DC. Retrieved from https://apps.lis.uiuc.edu/wiki/download/attachments/32666/Palmer_DCC2007.rtf?version=1

Pepe, A., Mayernik, M. S., Borgman, C. L., & Van de Sompel, H. (2010). From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web. *Journal of the American Society for Information Science and Technology*, *61*(3), 567–582. doi:10.1002/asi.21263

Pestre, D., & Krige, J. (1992). Some Thoughts on the Early History of CERN. *Big Science: The Growth of Large-Scale Research*, 78–99.

Price, D. J. de S. (1963). *Little Science, Big Science*. New York, NY, USA: Columbia University Press.

Sands, A. E., Borgman, C. L., Traweek, S., & Wynholds, L. A. (2014). We're Working On It: Transferring the Sloan Digital Sky Survey from Laboratory to Library. *International Journal of Digital Curation*, *9*(2), 98–110. doi:10.2218/ijdc.v9i2.336

SDSS Scientific and Technical Publication Policy. (2003, October). Retrieved February 2, 2010, from http://www.sdss.org/policies/pub_policy.html

Smith, R. W. (1992). The Biggest Kind of Big Science: Astronomers and the Space Telescope. In P. Galison & B. W. Hevly (Eds.), *Big science: the growth of large-scale research* (pp. 1–20). Stanford, Calif.: Stanford University Press.

Szalay, A. S., Kunszt, P., Thakar, A. R., & Gray, J. (1999). Designing and Mining Multi-Terabyte Astronomy Archives: The Sloan Digital Sky Survey. *cs/9907009*. Retrieved from http://arxiv.org/abs/cs/9907009

Traweek, S. (1988). *Beamtimes and Lifetimes: The World of High Energy Physicists* (1st Harvard University Press pbk.). Cambridge, Mass.: Harvard University Press.

Tyson, J. A., Sasian, J., Gilmore, K., Bradshaw, A., Claver, C., Klint, M., … Resseguie, E. (2014). LSST optical beam simulator. In *SPIE Astronomical Telescopes+ Instrumentation* (Vol. 9154, pp. 915415–1 to 915415–9). International Society for Optics and Photonics. doi:10.1117/12.2055604

Vermeulen, N. (2010). *Supersizing Science: On Building Large-Scale Research Projects in Biology*. Universal-Publishers.

Wallis, J. C., & Borgman, C. L. (2011). Who is Responsible for Data? An Exploratory Study of Data Authorship, Ownership, and Responsibility. In *Annual Meeting of the American Society for Information Science & Technology* (Vol. 48, pp. 1–10). New Orleans, LA: Information Today. doi:10.1002/meet.2011.14504801188

Wallis, J. C., Borgman, C. L., Mayernik, M. S., & Pepe, A. (2008). Moving Archival Practices Upstream: An Exploration of the Life Cycle of Ecological Sensing Data in Collaborative Field Research. *International Journal of Digital Curation*, *3*(1), 114–126. doi:10.2218/ijdc.v3i1.46

Wallis, J. C., Borgman, C. L., Mayernik, M. S., Pepe, A., Ramanathan, N., & Hansen, M. A. (2007). Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries. In *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries* (Vol. LINCS 4675, pp. 380–391). Budapest, Hungary: Berlin: Springer. doi:10.1007/978-3-540-74851-9_32

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing Data Curation Profiles. *International Journal of Digital Curation*, *4*(3), 93–103. doi:10.2218/ijdc.v4i3.117

Wray, K. B. (2006). Scientific authorship in the age of collaborative research. *Studies in History and Philosophy of Science Part A*, *37*(3), 505–514. doi:10.1016/j.shpsa.2005.07.011

Zhang, J., Chen, C., Vogeley, M. S., Pan, D., Thakar, A., & Raddick, J. (2012). SDSS Log Viewer: visual exploratory analysis of large-volume SQL log data. In P. C. Wong, D. L. Kao, M. C. Hao, C. Chen, R. Kosara, M. A. Livingston, … I. Roberts (Eds.), (Vol. 8294, p. 82940D–82940D–13). doi:10.1117/12.907097

Table of Figures