July 25, 2015

Knowledge Infrastructures in Science: Data, Diversity, and Digital Libraries

Christine L Borgman, *University of California, Los Angeles*
Darch T Peter, *University of California, Los Angeles*
Sands E Ashley, *University of California, Los Angeles*
Pasquetto V Irene, *University of California, Los Angeles*
Golshan S Milena, *University of California, Los Angeles*, et al.

Knowledge Infrastructures in Science: Data, Diversity, and Digital Libraries

Invited paper for Special Issue of the International Journal on Digital Libraries
Submitted 26 November 2014; Revised 17 March 2015; Final 10 June 2015;
Proofs July 15, 2015; forthcoming in IJDL

Christine L. Borgman, Peter T. Darch, Ashley E. Sands, Irene V. Pasquetto,
Milena S. Golshan, Jillian C. Wallis, & Sharon Traweek
Center for Knowledge Infrastructures
Department of Information Studies, University of California, Los Angeles
GSE&IS Building, Room 235, Box 951520, Los Angeles, California 90095-1520
+1(310)825-6164
christine.borgman@ucla.edu, petertdarch@ucla.edu, ashleysa@ucla.edu,
irenepasquetto@ucla.edu, milenagolshan@ucla.edu, jwallisi@ucla.edu,
traweek@history.ucla.edu

Table of Contents

Abstract 2
Introduction 2
Knowledge Infrastructures in Science 4
  Data, Digital Libraries, and Stewardship 5
  Data in the Life Cycles of Science 6
  Big Science, Little Science, and Scale 6
  Data: Open and Closed 8
Research Methods 8
Findings 12
  Research Sites: Digital Libraries and Degrees of Openness 12
    Center for Embedded Networked Sensing (CENS) 12
    Sloan Digital Sky Survey (SDSS) 13
    Center for Dark Energy Biosphere Investigations (C-DEBI) 15
    Large Synoptic Survey Telescope (LSST) 17
  Comparing Sites: Dimensions of Diversity 18
    Smaller-Scale Science: CENS and C-DEBI 20
    Larger Scale Science: SDSS and LSST 23
    Earlier Stages of the Life Cycle: C-DEBI and LSST 25
    Later Stages of the Life Cycle: CENS and SDSS 26
Discussion 29
Conclusions 30
  Knowledge Infrastructures at Scale 31
  Knowledge Infrastructures in Rhythm 32
  Future Research Directions 32
Acknowledgements 33

Abstract

Digital libraries can be deployed at many points throughout the life cycles of scientific research projects from their inception through data collection, analysis, documentation, publication, curation, preservation, and stewardship. Requirements for digital libraries to manage research data vary along many dimensions, including life cycle, scale, research domain, and types and degrees of openness. This article addresses the role of digital libraries in knowledge infrastructures for science, presenting evidence from long-term studies of four research sites. Findings are based on interviews (n=208), ethnographic fieldwork, document analysis, and historical archival research about scientific data practices, conducted over the course of more than a decade. The *Transformation of Knowledge, Culture, and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective* project is based on a 2x2 design, comparing two "big science" astronomy sites with two "little science" sites that span physical sciences, life sciences, and engineering, and on dimensions of project scale and temporal stage of life cycle. The two astronomy sites invested in digital libraries for data management as part of their initial research design, whereas the smaller sites made smaller investments at later stages. Role specialization varies along the same lines, with the larger projects investing in information professionals, and smaller teams carrying out their own activities internally. Sites making the largest investments in digital libraries appear to view their datasets as their primary scientific legacy, while other sites stake their legacy elsewhere. Those investing in digital libraries are more concerned with the release and reuse of data; types and degrees of openness vary accordingly. The need for expertise in digital libraries, data science, and data stewardship is apparent throughout all four sites. Examples are presented of the challenges in designing digital libraries and knowledge infrastructures to manage and steward research data.

Introduction

Knowledge infrastructures are most simply defined as "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds" [43:17]. Infrastructures are not engineered or fully coherent processes. Rather, they are best understood as ecologies or complex adaptive systems. They consist of many parts that interact through social and technical processes, with varying degrees of success. Knowledge infrastructures include technology, intellectual activities, learning, collaboration, and distributed access to human expertise and to documented information [44]. Digital libraries, usually understood as information retrieval systems that support text, images, numeric data, and other formats [11], are an essential component of knowledge infrastructures. They may be deployed throughout the research life cycle, from capturing observations through cleaning, analysis, and interpretation of data, to management, curation, and stewardship of research products. We take a broad view of digital libraries, spanning the range from local systems for managing research data to large-scale data repositories, and applications from the initial stages of data collection through archiving and preservation.

As scientific technologies such as genomic sequencing, sensor networks, astronomy instruments, and laboratory tools collect data at significantly faster rates, the need for digital library services also grows. Adapting scientific methods to greater volumes of data, often with greater diversity, poses new challenges for science and for data management. Combining data from multiple sources for new interpretations requires yet new kinds of systems and services.

Digital libraries tend not to be generic systems. Rather, they are most effective when designed for specific communities and types of content. To design digital libraries for the management of scientific data requires expertise in scientific theory, method, instrumentation, interpretation, and knowledge organization. Whether a single digital library can support the entire life cycle of a given project is an open question, given the range of expertise and types of data handling involved. Scientific expertise is complex and divided differently within each field and specialty. Each step in data handling requires knowledge of the steps that went before. Details about data provenance that are necessary for interpretation may go unrecorded, so that future researchers lack the information necessary for reusing those data [10]. Researchers who design and carry out data collection activities may become aware of very small differences in calibration, minute artifacts in a data stream, and other perturbations – but these potential sources of error may be imperceptible to researchers further away from the data source. Digital libraries to manage research data may require far more metadata, provenance information, and other documentation than is required for most other retrieval applications.

Data management has become a much higher priority in the research process due to requirements of funding agencies and journals to release research data at the time of article publication. Significant challenges to implementing data management arise from the complexities of modern scientific collaboration: competing notions of what counts as data, disparate views of research and innovation, differing incentives for data sharing and release, debates around economics and intellectual property issues relating to knowledge products, and public policy. If the potential of data-intensive science is to be realized, then appropriate systems, services, tools, content, policies, practices, and human resources are required to discover and exploit research products. However, it is not yet clear what infrastructure should be built or how to build it. Digital libraries are a small but important part of the solution [12,13,16,44].

Managing research data is difficult, and making research data useful to unknown others, for unanticipated purposes, is far harder. As researchers approach the data management limits of available tools and resources, they hit the scaling problem. Having more data requires not just larger or faster tools, but indeed different tools and different modes of inquiry. The scaling problem is playing out differently in each field, lab, project, and research site. Only by comparing multiple cases over long periods of time can the array of data management challenges and the roles of digital libraries be identified.

Data management and digital library requirements vary considerably by the scale of the scientific project. Data collected by large instruments at large facilities are characterized by international, collaborative efforts that produce vast amounts of data. These data often are big in volume and velocity, but may be homogeneous in form and structure. Data collected at distributed sites by small teams are typified by heterogeneous methods, diverse forms of data, and by local control and analysis. Price [98] and others categorized these distinctions as "big science" and "little science," respectively. Data management concerns and practices appear to differ greatly along these and other dimensions [19,35,86]. "Big data" is equally difficult to define. From a management perspective, distinctions between volume, variety, and velocity are useful [78]. From a research perspective, "bigness" is relative to the available methods and tools for interpretation [90]. Degrees of homogeneity and heterogeneity may influence research data management more than size, per se [13].

Socio-technical research approaches can inform design, policy, and human resource requirements for infrastructure at all scales of science and scholarship. The *Transformation of Knowledge, Culture, and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective* project (henceforth known as the *Knowledge Infrastructures* project) compares four large, distributed, multidisciplinary scientific endeavors: two of the research projects collected data from very large instruments at large facilities, classified as big science in our research design, and two projects collected data at distributed sites by small teams, classified as little science for the purposes of comparison. Two sites are in the early stages of their research life cycle, ramping up their research activities, while the other two are in later stages, ramping down their data collection and active research.

This article presents an overview of the Knowledge Infrastructures project, a long-term exploration of processes related to data practices, one component of which is the use of digital libraries. In the next sections, we explain the research questions, outline the research methods, and discuss the data practices we observed at each site. More detailed analyses of the sites are provided in prior publications, which are referenced throughout this article. Initial comparisons of the four sites were presented in the conference paper on which this article is based[1]. Here we offer fuller analyses, conclusions, and directions for future research.

Knowledge Infrastructures in Science

Knowledge infrastructures for data-intensive science must do much more than disseminate resources – they must support data collection, analysis, use, and access to information. Knowledge infrastructures are expensive to construct and maintain. The value proposition and burden of costs are much debated [4,8,23,46,64,103].

[1] Articles for this special issue were invited from the "best papers" nominations from DL2014, the IEEE/ACM Joint Conference on Digital Libraries, London.

The design of successful knowledge infrastructures for science depends on successful explication of the socio-technical structures embedded in research data practices, technical configurations, and policies. These interdependencies pose significant barriers to establishing and adopting of effective infrastructures. Among the digital library challenges in managing research data are granularity, provenance, structures, identity, identifiers, and functions of data [7,31,100].

While countless policy reports call for building infrastructure and capacity for managing research data, only a handful of studies have addressed how understanding data practices might inform design and policy. Included in studies of knowledge infrastructures are research on work practices, collaborations, virtual organizations, computer supported collaborative work, project life cycles, and temporal factors [13,43,48,56,75,76,79,101,109,117].

Data, Digital Libraries, and Stewardship

Digital libraries originated with textual content and expanded quickly to include multi-media resources and research data. Design requirements vary by content type, user community, and other factors. Digital libraries, whether for data or documents, are most often deployed at the end of the research process to provide access to publications and to data that are released for purposes of reuse, verification, or reproducibility. A broader conception of digital libraries that can support the entire information life cycle is not a new idea [14]. However, this more inclusive notion requires a different architecture than systems designed for publications or other documents. Rarely are data self-describing, nor do they stand alone as independent units. Data are best viewed in relationship to papers, protocols, analytical tools, instruments, software, workflows, and other components of research practice. Thus, expertise in organizing and retrieving complex research objects has become critical to the management of data [13,44,95,99].

Digital libraries can support active use of data during the research process, products to be managed for later reuse or repurposing, and access to data and results. The activities and expertise will vary along this continuum. Data management is the general rubric of ensuring the integrity, access, and usability throughout the research process and beyond. Data curation is a form of digital curation, which "involves maintaining, preserving and adding value to digital research data throughout its lifecycle" [41]. Data preservation, in contrast, ensures long-term integrity but not necessarily availability for active scientific use. Dark archives provide preservation but not access. Stewardship addresses the need for long-term sustainability of data, including integrity and access. Some distinguish further between integrity, accessibility, and stewardship of research data [34]. Digital libraries can support any of these roles. The design challenges are to understand the data practices of the research community, to make policy decisions on how to sustain access, and to plan for curation. These decisions are inseparable from the workforce requirements for managing research data, including data scientists and data stewards.

Data in the Life Cycles of Science

"Life cycle" is a term commonly used in archives, libraries, records management, and digital libraries to reflect changes in the form and character of objects over time, usually from their origin to their ultimate disposition, which may be preservation or destruction [14,25,62,65,66,67,117]. When used in the context of scholarly communication, life cycles may refer to stages of research projects, from origin of the ideas through to publication and dissemination. These life cycles intersect, in that different research products are created and used in each stage of research [97]. Despite the wide use of the term, "life cycle" remains problematic as it suggests that research proceeds stepwise with a beginning, middle, and end, rather than iterating through many steps [12,24].

Scientific activities involve multiple, intersecting, and often disparate cycles. "Rhythms of collaboration" better captures the complexity of how data, practices, collaborations, and activities flow through any project [70,71]. Data collection may involve access to instruments, research sites, or people, and can be driven by sampling rates and other rhythms. Deadlines for grant proposals, conference papers, publishing venues, and personnel reviews often are in conflict. Data management may be central to planning long term projects – or may be an afterthought addressed as project funding ends or as deadlines for data release loom large. Large projects may consist of many smaller activities, each with its own life cycle of data practices. Conflicts over data practices often contribute to "science friction" [45].

Big Science, Little Science, and Scale

The terms "little science" and "big science" have been used since the 1950s to characterize and contrast styles of the organization of scientific work [98]. Both terms have been subject to a wide range of interpretations. The dichotomy of little and big science remains a highly influential paradigm within a wide range of academic disciplines that study scientific work, most notably Information Studies, Science and Technology Studies, History of Science, and Sociology [52,53]. However, these distinctions also are highly problematic, attempting to reduce the complexity of science to a simple dichotomy. In our research, we prefer to investigate how scale matters in data practices, rather than assume the distinction.

In our current work, we are finding a complex array of scalar relationships between aspects of individual projects [16,17,37,39]. These lead us to pose new questions about these relationships: To what extent is big science an aggregation of little science contexts? When and how do little science contexts cohere into a single infrastructure or big science context? When and how do small team practices cohere into a larger, single infrastructure? How do the scale of instruments and infrastructures shape each other? How do data practices vary by these factors?

Scientific research can be viewed in terms of several scale dimensions. Research conducted by small teams at distributed sites over the short term has been referred to variously as "small science" [35,92] and "long-tail science" [63,94]. Funding of such

projects is typically on the scale of tens or hundreds of thousands of U.S. dollars [63]. The structure and organization of short-term, distributed research projects conducted by small teams has important implications for their data management practices. Such data are generally small in volume, but may be heterogeneous in type and form [9,19,73,74].

Little science projects, by definition, tend to be carried out by individuals or small teams of scientists working in either a single laboratory or at distributed sites. As a result, role specialization is minimal. Resulting scientific papers are usually single-authored, or involve at most a handful of co-authors [76]. Standardization of methods across the scientific domain is minimal, to the extent that each scientist may use different tools and techniques to generate datasets similar in form and intent [38]. Responsibility for data management falls to the scientists who produced the data. As a result, data tend to be managed by localized, ad hoc practices for the immediate purposes of the scientists [20,116]. Consequently, data often are neglected after their immediate usefulness and may be lost [9].

At the other extreme of scale is scientific research conducted with large instruments, often at large facilities [30,55]. Big science projects of these types require massive funding, often on the scale of tens, millions, or even billions of U.S. dollars from multiple government agencies, private benefactors, or foundations [77]. This level of funding is necessary to construct and operate large facilities, whether telescopes, linear colliders, or data repositories. Such projects are found in a range of disciplines in the physical sciences, including physics, astronomy, physical oceanography, fusion physics, and space sciences [89,105,109]. Although human genome research can be conducted at smaller facilities with less expensive instruments, widely distributed research requires large budgets [80,114]. The greater the funding, the greater the oversight from the agencies providing those funds [30,55].

Collaborations using large instruments at large facilities may have hundreds or thousands of members and often are international undertakings, although the team members usually are distributed worldwide [76]. Documents governing the work and organization include formalized agreements between partner institutions, extensive reporting to funding bodies, detailed work plans, and policies [32,61]. Another consequence of size is greater division of labor. The larger the project, the more specialized and routinized that individual work tasks may become [26,32]. The scale of large collaborations and the nature of the work involved have given rise to publications with many tens, hundreds, or even thousands of co-authors [54,76,120]. In those subfields the reputations of individual researchers depends less on authorship metrics than on other roles in the community [109,110,111,112].

We are investigating how data and data practices differ across the scale of research, whether measured in terms of facilities, size of teams, size of data, funding, or other factors. Large projects tend to be more routinized and to produce large volumes of homogenous data. As projects increase in scale, the conditions under which data are

collected, stored, managed, curated, and made accessible tend to become more standardized and routinized. Smaller projects often are more adaptable to local conditions. Technology requirements for data production also appear to vary by size of project, diversity of data, distribution of the workforce, and other factors [9,13,22,61].

Data: Open and Closed

Digital libraries may support internal research use, particularly at earlier stages of the research process, with or without the intention to make data externally available later. Digital libraries also may support open access to data, especially at the points in a research project where datasets are released to repositories. Open access to data is not necessarily equivalent to open data or to open science. These concepts are either hotly debated or taken as givens in the practice of research. Scholars care deeply about the long-term availability of their publications; few are willing to make comparable investments in the longevity of their data.

Open access to publications is predicated upon two conditions that do not transfer well to data: authors are copyright holders of their published work, until and unless they choose to transfer control; and scholars write articles to influence their audience, rather than for payment [13,106]. In contrast, the ownership and control of data remains among the intractable problems of eResearch [40], and the incentives and available resources to disseminate data are minimal.

Many stakeholders make decisions about openness in any given research project. Decisions include dissemination strategies of scientists, governance models of universities and research centers, and the policies of funding agencies and journals. Openness is further complicated by the aggregate nature of research objects [6,97]. Datasets have little scientific value if documentation is inadequate or if the associated hardware, software, protocols, and other technologies are proprietary, unavailable, or obsolete.

Despite the pressures from funding agencies and other stakeholders to release data, "open data" is far from the norm, whether due to technical, infrastructural, cultural, or social factors. Thus the questions become what "open" means to scientists, the conditions under which they make their data open, what they expect to gain or lose by openness, and what roles other stakeholders play in how openness is negotiated. As our research reveals, openness is not a specific set of practices but a process that occurs throughout the research life cycle.

Research Methods

The Knowledge Infrastructures project addresses four questions across four sites:

1. What new infrastructures, divisions of labor, knowledge, and expertise are required for data-intensive science?

2. How are the infrastructures of multi-disciplinary, data-intensive scientific endeavors established and how are they dismantled?
3. How do data management, curation, sharing, and reuse practices vary among research areas?
4. What data are most important to curate, from whose perspective, and who decides?

The four research sites vary by scale of the data-intensive research and by stage of the project life cycle, as presented in Figure 1. The two smaller scale sites, each of which produces small amounts of heterogeneous data, are the *Center for Embedded Networked Sensing* (CENS) and the *Center for Dark Energy Biosphere Investigations* (C-DEBI). The two larger scale sites, with large instruments and facilities that produce great volumes of relatively homogeneous data, are the *Sloan Digital Sky Survey* (SDSS) and the *Large Synoptic Survey Telescope* (LSST). The life cycle comparison is between sites in earlier stages (C-DEBI and LSST) that are ramping up infrastructure development and data production, and sites at later stages (CENS and SDSS) that have established their infrastructures and completed data collection.

Figure 1: Sites by scope of research projects and stage of life cycle

	Larger Science	**Smaller Science**
Ramping up data collection	LSST	C-DEBI
Ramping down data collection	SDSS	CENS

These projects vary by collaboration size, duration, cost, research technologies, and organizational complexity. They are conducted in spaces that range from one locality to globally distributed sites. We are studying how the work is divided among teams large and small, how local and global are the practices and priorities, and how smaller units aggregate into large collaborations. We also explore how these combinations of arrangements influence data practices at each stage of project life cycle.

Research questions about the types and degrees of openness of these projects cut across the scale and life cycle dimensions. On the one hand, larger science contexts might lend themselves to greater openness as they often are characterized by more homogeneous data and more standardized data practices (contributing to interoperability and sharing), role specialization (people tasked with data management and curation), bureaucracy (mandates to manage, curate, and share data), and more funding for infrastructures. On the other hand, research at smaller sites may be more open as these types of projects are characterized by greater flexibility in methods, tools, and infrastructure. Researchers may adapt their practices more quickly to new audiences and publishing venues, and can adjust their data life cycles accordingly. Our research questions concern the degree to which variations in openness, data life cycles, data management, and homogeneity of data are associated with scale or with a big science / little science dichotomy.

We have studied these four research sites for different periods of time. Each site is distributed across multiple institutions, some internationally. Our research on CENS began prior to its inception in 2002, followed through its ten years as a National Science Foundation Science and Technology Center (NSF STC), and continues, now considering its legacy [15,18,19,20,118,119]. Research questions about SDSS have evolved through several grant projects since 2009, covering data practices, knowledge transfer, and workforce development [49,102,121,122]. Background research on LSST began in 2009 and fieldwork in 2014. Research on C-DEBI began in 2012, and we plan to continue to study this site through its next stages of development [37].

Research on each site involves a mix of methods, including semi-structured and unstructured interviews, ethnographic fieldwork, document analysis, and historical archival research to explore data practices from multiple perspectives. Our methods enable us to investigate individual, collaborative, and community-wide data practices over a long time frame. Locating the intersecting patterns and anomalies in the records gathered by each method strengthens our analyses**.**

Interviews with individuals capture their point of view about their experience with data practices throughout their career trajectories, working on multiple projects, including their present activities. The locations and times for the interviews are established by mutual agreement between the interviewees and interviewers; most are conducted at the interviewee's worksite. These interviews usually last between 45 and 120 minutes. They are recorded, transcribed, and coded. For each interviewee quoted in this paper, we assign an identifier comprising the acronym for the project with which the interviewee is affiliated and a unique number for the individual (for instance, CENS-1, LSST-2, etc.).

Ethnographic fieldwork is focused on collaborations, communities, and working relationships. Our fieldwork is conducted at each site in situations and events structured by the people under study to investigate discussions and practices among participants working together, engaging in the same project, or participating in the same event. We focus on how participants work together to define, modify, and transmit their knowledge, and how they modify the intellectual and infrastructural ecologies in which they are working. More specifically, we observe their divisions of labor, allocations of scarce resources, and their processes for dispute negotiation, maintenance, and resolution. We take extensive field notes based upon our field observations. Our notes are organized by four frames: ecology (funding, built environment, tools, skills), social organization (divisions of labor, allocation of scare resources, decision and disputing processes), stages of a career/project/knowledge development, and knowledge making (styles of inquiry, practices, conceptual strategies).

Document analysis includes collecting public records about the individuals and projects, including publications, websites, presentations, and ephemera. We also collect materials that are made available by our research participants. These provide background information, evidence of activities, and corroborating resources on various aspects of the study.

Historical archival research methods require the examination of documents, formal and informal, in all media that have been generated by and about the group under study throughout its existence. Those documents might be archived at multiple locations: research and work sites, funding agencies, publication venues, data repositories, and even residences. The documents are in multiple formats, shaped by the needs of the people and organizations that generated and required them for different occasions, such as seminars, workshops, conferences, collaboration meetings, and conducting experiments. Each set of documents is analyzed to identify the shared assumptions and disputes that locate the exchange at any specific time and place. Sets of documents are juxtaposed to examine the distribution and sequence of the frames of reference embedded in them. The documents' organizational context, authors, readers, distribution, and access also are studied. These discursive practices are examined for instances of contestation, adjudication, and consensus building.

The CENS comparisons presented here are drawn from a round of 77 semi-structured interviews collected in 2006-2013, participant-observation in a variety of capacities throughout the lifetime of the Center, and analysis of documents such as publications and annual reports. For SDSS, we draw on 60 interviews conducted with 54 participants, five weeks of ethnographic participant-observation, and analysis of webpages and other documents. For C-DEBI, we draw from a round of 49 interviews, participant-observation that included being embedded in a laboratory for eight months and short-term observation in two other laboratories, and involvement in the development of data management infrastructure. For LSST, we draw on background research, 22 interviews, and 7 weeks of ethnographic participant-observation (Figure 2). Analytical coding of interview transcripts, fieldnotes, and documents were done in NVivo 9, a qualitative analysis software package, and analyzed for emergent themes using grounded theory [57].

Figure 2: Interviews used for findings in this article. Cell column totals reflect overlapping participation in institutions and projects

Sites	Interviews	People	Institutions	Period
CENS	77	72	4	2006 - 2013
SDSS	60	54	14	2012 - 2013
C-DEBI	49	49	16	2012 - 2014
LSST	22	21	4	2014
Total	208	193	33	

In this article, we compare our data practice findings from the four sites to address implications for the design of digital libraries. Specific questions include what factors of data management practices are amenable to digital library solutions and which are larger knowledge infrastructure concerns. Of particular interest is the ability to identify data management issues that vary by domain and those that are common across domains. Questions about knowledge infrastructures and openness cut across all of these dimensions.

Findings

Findings are presented in two parts. First, each of the four sites is analyzed with respect to data management concerns that may or may not be amenable to digital library solutions. Second are pairwise comparisons by the two dimensions of our inquiry: scale of scientific practice and life cycle stage of project. We distinguish between data management issues that are specific to individual domains and those that are common across domains. Discussion of the findings follows, addressing the implications for knowledge infrastructures and for the design of digital libraries.

Research Sites: Digital Libraries and Degrees of Openness

In this section, each of the four research sites is described with respect to their use of digital libraries and their types and degrees of openness. These analyses provide the basis for later comparisons on the dimensions of scale and life cycle.

Center for Embedded Networked Sensing (CENS)

CENS (2002-2012) was a U. S. National Science Foundation Science and Technology Center devoted to developing embedded networked sensing systems for scientific and social applications through collaborations between engineers, computer scientists, and domain researchers. By partnering across disciplinary boundaries, participants had to articulate their research practices, methods, and expectations explicitly. Membership varied from year to year as projects began and ended, and as the rosters of students, faculty, post-doctoral fellows, and staff evolved. At its peak, the Center had about 300 participants from the five partner universities in California and collaborators from other institutions. On average over the life of the Center, about 75 to 80% of CENS participants were concerned with the development and deployment of sensing technologies; the rest were in science, medical, or social application domains. Technology research addressed the development and testing of embedded networked sensing systems. Research in the application domains addressed the new methods and findings made possible by these technologies.

Most CENS research was little science in character, conducted in field deployments that produced heterogeneous types of data. Sensor networks produced far more data than did the hand-sampling methods that dominated CENS research in ecology, environmental sciences, biology, and ocean science. As the volume and velocity of data increased, science teams encountered scaling problems that their existing methods could not accommodate. In the marine biology studies, for example, science teams usually captured water samples three to four times in each 24-hour period. Those observations were correlated as time series. Sensor networks, however, sampled the water at five-minute intervals. Simple correlations and time series analyses did not suffice for these data rates, which led to the adoption of complex modeling techniques [20,21,118]. Seismology was the CENS research area that exhibited the most big science characteristics, such as established data standards, community repositories, large-scale equipment, and large

volumes of homogeneous data. CENS teams laid seismic monitoring equipment along transects across entire countries, including Mexico and Peru [87].

CENS data management problems, particularly in the domains with little science characteristics, were less amenable to digital library solutions than expected. Interest in a common data repository was minimal, for a variety of reasons that are explored here. While CENS was publicly committed to sharing data, the Center was established prior to NSF requirements for data management plans and they were not under pressure to develop data curation mechanisms. The "small science" approaches of CENS research resulted in heterogeneous data that were difficult to aggregate for comparison or reuse. A simple digital library, dubbed "The CENS Deployment Center" was developed and populated with descriptions of sets of equipment and personnel from past deployments. These functions were intended to make deployments more efficient and productive and to provide context about past deployments. The system was moderately successful in serving these functions [21,87,115,117]. Seismology was the exception, as noted above. They had a domain-specific repository to which they were required to contribute their data, and had little need for local solutions [45,119].

In CENS, data were a means to an end, which was to answer science domain questions or to build better technologies to ask those questions. The data from field deployments were dispersed to individual science and technology teams, with little concern for the ability to recombine them later. With the exception of seismology, data rarely were kept for reuse beyond the teams that collected them. The majority of CENS participants were technology researchers whose scholarly products were papers, software, and instruments. Software code was often treated as data, and might be contributed to code-sharing sites. Most researchers maintained their data locally. Different data policies led to different degrees of openness of data. CENS researchers were in principle open to sharing their data, but most data exchanges were between individuals. Few CENS data were released publicly or contributed to archives. We also found considerable confusion and disagreement about who was responsible for different types of data, and that responsibility might vary over stages of the project. Lacking agreement on responsibility, data frequently were neglected [20,21,115,119].

Sloan Digital Sky Survey (SDSS)

Astronomy sky surveys are research projects to capture uniform data about a region of the sky. The Sloan Digital Sky Survey, named after its largest funder, the Alfred P. Sloan Foundation, is notable for its commitment to timely data releases. The SDSS data are available through a sophisticated and large-scale digital library infrastructure. Judging by the number of papers mentioning SDSS, the data continue to be heavily used. For example, a May 2015 search of the SAO/NASA Astrophysics Data System (ADS) yields more than 9,000 papers mentioning "SDSS" in the title or abstract [2]. The actual number of papers using SDSS data is probably much higher, given the common practice of reusing data without citing them in publications [58,96,122].

SDSS planning began in the 1990s; the design of digital library infrastructure was integral to the project. Survey data collection began in 2000, mapping about one-quarter of the night sky with a focus on galaxies, quasars, and stars. A 2.5-meter optical telescope at Apache Point Observatory in New Mexico was designed, built, and deployed for the collection of the SDSS survey data. Multiple instruments on the telescope collected optical and spectroscopic data. The first phase of the SDSS project (SDSS-I) ran from 2000 to 2005; SDSS-II covered 2005 to 2008. Each was funded as an independent project. SDSS-II expanded the scientific goals and broadened participation. Over the series of eight data releases from 2002 to 2009, SDSS captured data at higher rates and better resolution due to new instruments added to the telescope, advances in charge-coupled devices (CCDs) for the cameras, spectroscopy, and improvements in computer speed and capacity. SDSS-III continued with largely new leadership, collaborating institutions, and scientific goals. SDSS-III collected data through summer 2014, when SDSS-IV began [3,50,59,104].

Our research focuses on SDSS-I and II and the disposition of the associated datasets. We began to study these SDSS projects in 2009 as they began their archival phase. In 2008, four Memoranda of Understanding (MOU) established how the datasets would be managed for the subsequent five years, until early 2014. The SDSS investigators chose to migrate the dataset, which is between 100 and 200 terabytes in size (depending upon what files are included), from the national laboratory previously hosting the dataset to two university research libraries. The libraries collaborated with SDSS astronomers to ensure proper management of the data during the MOU period [102].

As SDSS phases I and II have completed data collection, it is now apparent how extensively these data continue to be used by individuals and small teams of scientists [102]. Collaborators with roles in the design of the project benefited from early access to, and knowledge of, the data, as one of the SDSS collaborators explained:

> "By being a member, those institutions not only get to see the data immediately... If they're working hands-on with the reduction of the data, they also have a head start in knowing what's in there even before it's released to the universities. There might be somebody who's helping reduce the data, and they can say, 'Oh, well, that's interesting,' and they can follow it up with the [telescope] if they want to during that proprietary period" (SDSS-3).

SDSS data are open to the world after the proprietary period, in the sense of being available without fee or extensive license restrictions. Investigators take data from the SDSS archives and derive new data, alone or in combination with data from other sources, and are not obligated to release their resulting datasets. Small projects and sole investigators thus have access to high quality data, and may not require much time or external funding to conduct research with these data. Small projects exhibit minimal division of labor, with individual scientists often conducting all stages of the project in their laboratories or offices. Tools and practices may be localized, although the use of

open sources based on standard formats may promote the use of common tools. SDSS data flow from the digital library to individual researchers and teams. These teams, in turn, may derive new data products from SDSS and other sources, often in combination. Some of these derived datasets can be submitted back to SDSS, such as well-curated star catalogs. However, such derived data remain the responsibility of the teams that created them. Curation is typically localized and ad hoc, with significant loss of these datasets over time.

The SDSS dataset and its copies reflect an expansion of users and reuses beyond those expected by the data creators. SDSS data have been reused in multiple scientific communities and have become the basis for citizen science projects such as Galaxy Zoo, which led to the Zooniverse [36,124]. As an astronomy professor explains, "I'm not sure this was widely appreciated, but the SkyServer started as… some sort of interface that will allow high school kids to do interesting things with a big scientific database… But it then evolved into this thing, which is crucially important for professional astronomers" (SDSS-1).

The SDSS-I and II datasets remain available at the original national laboratory site. Copies of these datasets, in whole or in part, are also hosted by several other universities around the world, either as backup or for local research access. An SDSS researcher commented:

> "There are some other unofficial mirrors around the world. … There's still one in [names several locations]; I don't know. But these are kind of unofficial mirror sites. We do send data to them, they download our data, but we don't kind of hold their hand while they're bringing up the mirror site or anything. They just do that on their own. If they need anything from us, they just ask us" (SDSS-2).

The proliferation of SDSS dataset copies suggests yet another kind of openness, plus the fact that no master list of all copies appears to exist. SDSS has allowed users to take the data, process them in a variety of ways, and thus create additional value. SDSS investigators who led the initial data management design have received new funding to re-engineer the system, combining datasets from all the SDSS phases, and building a common platform to support data from multiple scientific disciplines [60,72,107,108,123].

Center for Dark Energy Biosphere Investigations (C-DEBI)

The Center for Dark Energy Biosphere Investigations is a ten-year Science and Technology Center that launched in October 2010 with five years of funding and the possibility of renewal for an additional five years [29]. The Center receives funding from the National Science Foundation, much of which is redistributed to participating scientists. These are short-term grants (typically one to three years in length), given to

individuals and small teams. On a six-month cycle, C-DEBI awards small grants to cohorts of doctoral students, post-doctoral researchers, and senior scientists. The project's policy is to fund new participants in each cycle. As a result, a steady stream of new members joins the C-DEBI community, bringing a variety of multidisciplinary approaches and methods to bear on research questions. These participants, more than 90 to date, are distributed across more than 30 institutions in the USA, Europe, and East Asia, and across multiple physical and life science disciplines.

C-DEBI scientists collect and analyze physical samples (known as "cores") from the ocean floor, such as sediments and portions of the basaltic crust, to describe their microbial communities and physical properties. The data life cycle often begins with ocean drilling cruises to collect cores. The most significant cruises are those conducted under the auspices of the *Integrated Ocean Drilling Program (IODP)*, an international organization established to study the seafloor, later known as the *International Ocean Discovery Program* [68]. Some core samples are processed on board these cruises to generate data about the physical characteristics of the seafloor. Other samples are distributed widely across the investigators, projects, and sites. These are used to generate both physical and biological data that can then be correlated to understand interactions between microbial communities and their environment. Physical samples may be stored in repositories for the long term. IODP makes extensive use of digital libraries for the curation and access of data generated on board their cruises, and for cataloging physical samples stored in repositories.

The launch of C-DEBI afforded opportunities to observe how the work of negotiating, building, and maintaining data management infrastructure unfolds in a new collaborative setting. C-DEBI is in the process of implementing digital library approaches to improve the management of data and to make data more accessible to the community. Making data more open was stated as an aspiration in the project's five-year *Strategic Implementation Plan* 2010-2015, although "technical difficulty" was also mentioned as a barrier to establishing the infrastructure [27:11]. Only in 2012 did building the data management infrastructure become part of C-DEBI's active work in response to new NSF requirements that all STCs must implement data management plans.

Their infrastructure includes an online data registry and repository, plus a policy document, *C-DEBI Data Management and Philosophy* [28]. Their data management plan stresses that the "C-DEBI STC is committed to open access for all information and data gathered during scientific research that is conducted as part of C-DEBI" [28:1]. C-DEBI scientists are mandated to make data available after a moratorium (typically two years), either in a public repository external to C-DEBI (if an appropriate repository exists) or in the C-DEBI repository, once established. Scientists will be required to create an entry in the C-DEBI registry for data that are deposited in an external repository. The task of building this infrastructure is complicated by the heterogeneity of methods, array of domain expertise, domain-specific curation needs, and disparate relationships between physical samples and digital data in the C-DEBI community [37].

Complex relationships between the domain of study and the IODP have shaped how C-DEBI is developing its infrastructure. The IODP is the latest iteration in a series of scientific ocean drilling cruise programs operating since the late 1960s. Research space – for people, equipment, and data collection time – is scarce on scientific ocean drilling cruises. Initially, the cruises were intended only for researchers from physical science disciplines. The first cruise dedicated to microbiology took place in 2002. C-DEBI scientists, many of whom are microbiologists, must negotiate with scientists from other projects and research domains. We are finding that the C-DEBI infrastructure for data management is being designed for parallel purposes, both local needs and access to IODP resources. These social considerations motivate the construction of this infrastructure and the choice of features [37]. In particular, C-DEBI aspires that its digital library eventually becomes interoperable with the IODP database, thus becoming part of the larger knowledge infrastructure of their science.

Large Synoptic Survey Telescope (LSST)

The Large Synoptic Survey Telescope is a massive astronomy collaboration that is building a ground-based optical telescope in Chile [82]. Planned as the next major sky survey, LSST is due to launch a decade-long phase of data collection in 2022, generating up to 20 terabytes of data nightly. It will collect time-resolved optical image data [5,69,84]. LSST is intended to support scientific advancements in four science themes: "probing dark energy and dark matter, taking an inventory of the solar system, exploring the transient optical sky, and mapping the Milky Way" [69], enabling individual and small teams of scientists accessing and using the data to work on their own research questions within these themes.

Initial discussions about LSST began in the early 1990s, and by 2001 LSST was one of seven Prioritized Major Initiatives in National Research Council's Decadal Survey of astronomy, ranked as the most important ground-based facility [33]. The LSST Corporation was formed in 2003 as a non-profit organization. In 2012, the National Science Board approved funding for the final design stage, and in summer 2014, NSF approved funding for LSST to transition from its research and development phase to its construction phase. Although planning for digital library approaches to data curation and access have been embedded in the collaboration since LSST's initial conception, this transition is accompanied by the ramping up of the implementation of infrastructure for data collection, management, and accessibility, and by policy decisions about who can access the data and on what terms.

LSST is headquartered at a Center in Tucson, Arizona, under the auspices of the Associations of Universities for Research in Astronomy, with significant aspects of the data management work based at other sites, including the SLAC National Accelerator Laboratory, the University of Washington, the Infrared Processing and Analysis Center, the National Center for Supercomputing Applications, and Princeton University. Scientists at nineteen national laboratories and universities are involved in building the LSST telescope. Eleven Science Collaborations have been convened to plan for the multiple scientific domains of LSST data that address research questions in astronomy.

A significant amount of data-intensive work already has been accomplished on LSST, including simulations to test the emerging infrastructure. Studying these processes enables the Knowledge Infrastructures team to learn how data management practices are being negotiated, resolved, and incorporated as LSST moves towards its data collection phase. Although LSST appears to exemplify big science, given how much of their scientific work is based in large-scale facilities, we are finding that individuals and small teams are carrying out many elements of LSST development work. One example is that some components of the camera are being tested in short-term projects. Such projects are typically funded separately from the main LSST infrastructure development, with small budgets, employing localized data management and record-keeping practices. Findings from these projects are eventually folded into the development of the larger infrastructure.

The ethos of openness is fundamental to LSST data management principles [16], although subject to negotiation and restrictions. LSST policy requires that the code used to build the LSST data management infrastructure is open source. Potential end-users of LSST data and software are invited to use and adapt LSST source code, which later may become part of the official LSST code. Whereas LSST source code will be available globally, LSST datasets will be openly accessible within the United States but available outside the U.S. only through agreements negotiated with individual countries. Within partner countries, access may or may not be limited to LSST investigators.

Comparing Sites: Dimensions of Diversity

This section provides pairwise comparisons of the four sites by two dimensions: scale and stage of life cycle. The two-by-two research design of the Knowledge Infrastructures project hypothesizes that data management practices will vary along these dimensions. In some cases, distinctions are sharp and in others they vary along a continuum. With a view to informing future decision-making about the support of data management practices in scientific collaboration, we consider which practices are amenable to digital library solutions and which are larger knowledge infrastructure concerns. We also identify data management issues that are specific to individual domains and those that are common across domains. Comparisons between sites are summarized in Figure 3.

Figure 3: Pairwise comparisons between sites

	Similarities	**Differences**
CENS and C-DEBI *Smaller scale science*	• No clear chain of responsibility for data; • Different groups responsible for different types of data; • Interpersonal data sharing, with conditions attached.	• CENS was all smaller-scale science, whereas C-DEBI involves science at diverse scales; • C-DEBI is subject to NSF Data Management Plan requirements, whereas CENS was not.
SDSS and LSST *Larger scale science*	• Primary legacies are large-scale astronomy datasets; • Other legacies include instrumentation and expertise; • Many overlaps in personnel, resulting in transfer of expertise from SDSS to LSST.	• LSST will collect larger volumes of data, at much greater velocity than did SDSS; • LSST data will be made available almost immediately, whereas SDSS data were released annually following a proprietary period.
C-DEBI and LSST *Earlier stages of the life cycle*	• LSST and C-DEBI both involve datasets collected within the context of a large-scale enterprise, but which also are used in the context of smaller projects.	• LSST ramping-up is taking about two decades, whereas C-DEBI ramped-up in a period of months; • In C-DEBI, data producers and users are often the same people; in LSST these are different groups, with some overlap; • LSST involves a narrower range of disciplines than C-DEBI; • C-DEBI has greater data diversity and heterogeneity than LSST; • LSST builds on the extensive infrastructure of astronomy, whereas a goal of C-DEBI is to fill major infrastructural gaps; • LSST has dedicated data management staff, whereas C-DEBI is retraining domain staff.
CENS and SDSS *Later stages of the life cycle*	• CENS and SDSS each have had lasting impacts on their respective domains through their publications; • Both developed new technologies; • Both extended the application of existing technologies to address new scientific questions.	• SDSS produced a comprehensive dataset that remains scientifically valuable; • CENS produced technologies and software; • CENS left data curation to individual projects; no common dataset was created.

Smaller-Scale Science: CENS and C-DEBI

CENS and C-DEBI have many similarities. Both projects are interdisciplinary federations of small teams of technologists and scientists working on projects funded by a mixture of internal and external grants. Both projects involve the generation of a wide range of small-scale, heterogeneous datasets that cover many scientific disciplines.

However, multiple factors within each project have implications for data production and management practices. One difference is that CENS focused on developing technologies to facilitate new forms of scientific work. Their technology research teams worked in concert with teams of domain scientists. C-DEBI, in comparison, does not have separate technology and engineering teams. In both CENS and C-DEBI, technologies, methods, and scientific problems evolve together. Technologies and practices are *co-emergent* in the sense that novel or adapted technologies and methods afford the pursuit of novel scientific problems and, conversely, novel scientific problems drive the development of technologies and methods to address these problems. The CENS leadership team referred to this relationship as "co-innovation," asking, for example, " 'Why does it need a sensor to make it happen?'– it was always that combination of scope and time scale of funding in order to be able to do that iterative co-innovation between application and technology" (CENS-1).

Despite their close collaboration and interdependent research interests, the data practices of technology and science researchers in CENS were largely independent. The sensor data were of mutual interest, but for different purposes. The science researchers focused on the science variables, whereas the technology researchers tended to focus on the accuracy and integrity of their systems, abstracting away the science content in the process [20,118].

In C-DEBI, scientists record their technologies or methods alongside the resulting data, for instance in their laboratory notebooks. Details are published in their journal articles. A disparate array of people may assume responsibility or ownership of interdependent datasets. A fundamental feature of C-DEBI is the collection and production of biological data about microbes and physical data about the environment that these microbes inhabit. These data are often correlated to track the impact of the microbes on their environment and vice versa. However, while many journals in which they publish require that biological datasets supporting the findings of journal articles be deposited or otherwise made available, no such mandate exists for physical data, as explained by one of our interviewees:

> "Nowadays they won't publish your work if it has molecular data and it's not in the database somewhere. The other data is just in a table... There are now databases where you could, I guess, submit this type of data like the geological data. But I haven't started doing that yet" (C-DEBI-2).

Differential requirements lead to differences in the extent to which biological and physical data and associated publications are released. Bioscience journals generally require that associated data be available in biodatabases or repositories, whereas physical

science journals rarely have such data deposit requirements. Conversely, physical science journals will publish articles for which preprints have been circulated in arXiv or similar open access repositories, whereas bioscience journals are less willing to publish articles for which preprints have been posted. The order of authors and the significance of the ordering also vary greatly within and between fields of the biosciences and physical sciences.

Another important difference in data practices between CENS and C-DEBI is that while the entirety of data life cycles in CENS generally unfolded within a single, small team, a significant portion of the C-DEBI data life cycles may also occur in the context of very large-scale infrastructure, namely the Integrated Ocean Drilling Program. 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 *Chikyu*, cost approximately 600 million U.S. dollars to build in 2001-2002 [81]. MOUs were signed between the countries involved to govern the terms of each nation's role, including financial contributions and the number of berths to be allocated on each expedition. Other aspects of the IODP's operation were governed by documents such as a cruise's Sampling Plan. The conduct of work on each expedition involved a high degree of role specialization. IODP curators oversaw core processing and data production, as a curator explained to us:

> "I'm responsible for the core material that we receive from the time that it's handed to us from the rig floor… I'm responsible for coordinating any sampling that is done for research on the core material… I enter all of the information into the database... I'm responsible for overseeing that whole process [of core analysis] and making sure that one particular instrument isn't holding up the process... I'm just always kind of monitoring that everyone is properly handling the core material" (IODP-1).

Data and other knowledge product management practices in IODP cruises were subjected to a wide range of analyses that were consistent across all IODP expeditions, and conducted according to standardized procedures. As one of our interviewees explained: "It took them decades to come up with the system... standard protocols, standard procedures, standard storage" (C-DEBI-3). This division of labor enabled high-throughput processing of cores and data collection; as explained by the above IODP curator, "it's a much more concentrated environment so we're able to produce a tremendous amount of work… a lot gets done in an expedition for a 60-day period" (IODP-1). Resultant data, accompanied by rich metadata, have been stored in the publicly accessible IODP database, which is still in operation.

The data life cycle in our C-DEBI case study unfolds across several scales of practice, and these different scales are involved in complex relationships that shape and constrain each other. One example relates to ongoing efforts to standardize procedures for biological analysis of cores. Currently, scientists apply a variety of methods, even for basic analyses. This heterogeneity means that biological analyses are not conducted as part of the standard package of analyses on board IODP cruises. As a result,

microbiologists must spend a great deal of time in their onshore laboratories after a cruise conducting these basic analyses, thus reducing the amount of time and money available for more advanced analyses:

> "The difference between what we can use that money for, and say, what a sedimentologist can use that money for, is grossly different. Because a sedimentologist, the geochemist, the petrologist, the paleo-mag guys, all of them pretty much have all the data. And so, they're looking at the $15,000 as seed money to maybe do some analysis that they maybe pay for a grad student, maybe pay for a technician, maybe pay for somebody's time, to analyze it, to maybe take it another direction, but then work up that data and submit it for your grant. For the biologist, we have $15,000 to now process all of our samples, do all the sequence analysis, do the bulk labor of all of our work on the equipment that we already have to have in our lab versus what everybody else is using on the ship" (C-DEBI-4).

A second result is that microbiologists are often unwilling to travel on the time-consuming IODP cruises as little microbiological work is carried out, meaning many cruises do not have any microbiologists on board, as explained by the following quotation from one of our interviewees:

> "If you're really honestly doing anything microbiological on a ship, you're probably doing some culture-based analysis, which is important…but it is also extremely limited, extremely slow… So the motivation to sail is not that high… 10 weeks out of a semester, you may lose teaching, you lose a publication, you lose a meeting, and when you're trying to get tenure, it may be more valuable to stay here. And a lot of our colleagues have chosen to stay home because they lose a lot by going to sea" (C-DEBI-4).

The absence of a microbiologist can mean nobody advocating for microbiology during negotiations around the allocation of cores to cruise participants, resulting in fewer and poorer quality cores allocated to microbiology:

> "Due to limitations and then lack of lobbying because I was not out at sea and there was no microbiologist out at sea, I only got about 50 samples of my 150. And then half of those were from the top and bottom sections of the core, which means most of those are probably not useful to me. And there were no contamination checks that were done on any of the cores, so the integrity of each one of my samples is then questioned" (C-DEBI-4).

A group of leading deep subseafloor biosphere researchers has identified "a dramatic need to standardize how we do our science, but then also there is a need for us to standardize how the samples are handled from the ship" (C-DEBI-4). Their efforts include workshops at international conferences and high-profile articles in journals [93]. To reconfigure the community infrastructure, C-DEBI microbiologists are reconfiguring their individual and small team scientific practices in their own onshore laboratories.

Similarities between C-DEBI and CENS reflect the conditions under which researchers in little science domains share data. Both of these research sites have difficulty managing or curating their datasets due to the heterogeneity of expertise and competing research goals of collaborators.

Most data exchange in CENS was between individuals. CENS researchers reported that they were generally willing to share data. The frequency of data release was low, and accomplished by a variety of means. Some released data upon request, as this scientist explained:

> "If such a request [to share data] arrives, then we immediately grant it because we would like to actually share all of our information and all of our systems to the wider academic community. And so such a request arrived before, and [we] provided the data immediately, and we not only share the data but also the platforms we developed. We would like other researchers to adopt them and use them in their research as well and there are many cases when we did that as well" (CENS-3).

Other researchers, however, attached conditions to data release, whether due to a sense of ownership or the amount of work that had gone into collecting them [18,119]; for instance:

> "If you walk out into a swamp... out in this wacky eel grass and marsh, along with your hip-waders and [are] attacked by alligators... and then you do it again and again and again... I don't [want to] share that right away. I [want to] analyze it because I feel like it's mine" (CENS-4).

Although the life cycle of CENS began eight years earlier than C-DEBI, with interim technological advances affording greater opportunities to share data, similar approaches to interpersonal exchanges of datasets are observed within C-DEBI. However, one critical difference is that C-DEBI is subject to NSF requirements to formulate a data management plan, whereas CENS ended before these requirements came into effect. As a result, C-DEBI is now mandating open access to data that support their publications, and implementing that policy in their data management technology.

Larger Scale Science: SDSS and LSST

The Sloan Digital Sky Survey and the Large Synoptic Survey Telescope are large-scale projects to generate massive datasets in astronomy. In both of these projects, new instrumentation, the establishment of collaborations and sharing of expertise, and, above all, the resultant datasets, will be the scientific legacy products. However, SDSS and LSST also offer several important contrasts. Advances in technology enable LSST to collect data at greater volumes and velocity than SDSS, which influences the means by which they release data to the community. Differences in scientific goals also contribute to choices of instruments, areas of the sky to survey, types of data to collect, and patterns and rates of data collection.

LSST will collect data to address a broader set of scientific research goals than did SDSS. These scientific drivers require more complex negotiations among the collaboration members than was evident in SDSS. While confident that they can address all of their science goals with a single dataset [69:7], eleven distinct Science Collaborations [83], and the LSST Science Advisory Council are the current community input mechanisms where negotiations are taking place. LSST claims that, "no other project matches this diversity and LSST's potential impact on society in general" [69:33]. On the other hand, SDSS was more highly focused on mapping galaxies, quasars, and stars, for which it was not necessary to ensure compatibility of resources.

SDSS provided project data annually in the form of data releases, which were available to the entire world. These finely processed, public data releases were seen as an innovative approach to scientific dissemination. One astronomy professor involved in SDSS stated:

> "Overall, I would say that the creation of the SDSS archives was one of the major achievements of SDSS; as opposed to the data, the fact you can get at the data and… it's freely accessible with no [restrictions]... The major success of the whole project is the fact that thousands of people have been using this data" (SDSS-1).

LSST is in the midst of planning a three-pronged approach to distributing data [69:22–23]. It will provide annual data releases similar to SDSS; however, these releases will not be made public globally, but instead will be made available to participating countries and institutions. In addition to the data releases, LSST's plans for data release include a strategy to provide nearly immediate publicly available alerts of transient objects and events that may require follow-up investigation. Finally, LSST also plans to provide access to the processing and storage capabilities for end-user analyses. While SDSS was innovative in terms of globally dispensing annual data releases, LSST expects to innovate with nearly immediate turnaround of the notification of objects for follow-up and by enabling end-users to execute "customized codes at the LSST data centers" [69:23].

A significant number of personnel from SDSS now occupy senior leadership positions within LSST, and we are studying the types and degrees of knowledge that are transferred from SDSS to LSST. This continuity of expertise is expected to improve LSST as it faces similar challenges in collecting, managing, and analyzing data at unprecedented scales [69:11]. A team member of both SDSS and LSST explained data access requirements as a responsibility associated with public funding:

> "The reason why the LSST was ranked number one in the Decadal Survey of Astronomy… I believe it's primarily due to the fact that it will be [a] survey for everyone…. We have [a] team of [a] few hundred people who work on the project, but they all understand that we're all going to be surveying not for

> ourselves but survey for everyone in [the] US and hopefully around the world. Data will be public and we are all okay with that" (LSST-2).

Earlier Stages of the Life Cycle: C-DEBI and LSST

Many comparisons between C-DEBI and LSST can be made in terms of the scale of data, stage of development, diversity of expertise, organization, and scope of infrastructures. An additional comparison is temporal scale: the ramping-up of data collection at the early stages of the C-DEBI life cycle is relatively brief compared with the two decades from initial conception of LSST to the anticipated commencement of data collection. C-DEBI data collection is designed and often performed by the researchers who will use the data themselves in the near future. In contrast, LSST must be designed in anticipation of research questions and technologies many years hence, including the hope that "new and unanticipated phenomena will be discovered" [84:14].

The range of scientific disciplines in LSST is narrower than in C-DEBI, but broader than most astronomy projects. We are studying closely how these disciplinary differences shape collaborative practices. C-DEBI scientists come from a wide range of scientific disciplines, which contributes to greater diversity of data practices along three dimensions. One dimension is the types of data produced, with one of our interviewees reporting that they generated and used datasets covering "metagenomics... genomics... isotope geochemistry... bioinformatics... mineralogy, proteomics, geochemistry, geology" (C-DEBI-5). A second dimension is the methods used to produce similar types of datasets: individual scientists have been observed making disparate choices at almost every stage of the data life cycle, such as the choice of methods to extract DNA. As explained by one of our interviewees, "the DNA extraction, or RNA extraction procedure is just all over the place. I mean, we all use different types of DNA extraction procedures" (C-DEBI-6). The third dimension is the recordkeeping practices about the methods used to generate datasets; as another interviewee explains, "people were keeping lab notebooks…But it was kind of more all over the place" (C-DEBI-7). This mix of practices makes data management in C-DEBI particularly challenging.

The infrastructure available to these respective communities to manage, curate, and access data also differs considerably. Astronomy has a relatively sophisticated infrastructure, with more data becoming accessible through investments in stewardship and in knowledge infrastructures [13]. Conversely, C-DEBI was established to address the "lack [of] the infrastructural coordination mechanisms to guide and support the research" in the domain of deep subseafloor biosphere research [42:1]. LSST can build upon more existing practices and infrastructures than can C-DEBI. For instance, LSST is using expertise from prior sky surveys, reusing code from SDSS, and using data from the Dark Energy Survey to test the software and infrastructure [51]. However, existing astronomy practices and infrastructures also serve as constraints on LSST innovation, whereas C-DEBI partners have more flexibility. They also have a lower threshold for success, as reflected in this discussion of techniques to extract DNA: "My philosophy is pick one that's not terrible, accept the fact that they're all gonna be bad and just move forward with what you got" (C-DEBI-6).

LSST has staff dedicated to developing the data management infrastructure, which will take a decade or more to complete. Scientists affiliated with C-DEBI, in contrast, do not have data management teams to support their research (although C-DEBI has recently appointed a postdoctoral researcher to focus on data issues). Their immediate scientific results are of primary importance; data play instrumental roles in the production of these results. Scholarly credit accrues for scientific publications; little or no credit accrues for the collection of data per se, as one of our interviewees explains:

> "There's a community of scientists who are…thinking about the process of data analysis, and thinking about how to get rewarded, making sure that scientists get rewarded for doing a good job with sharing their data…I would say that that community is definitely a minority" (C-DEBI-8).

By contrast, in the current design and construction phases of LSST, credit accrues for developing data management software and for using simulated data to guide the future operation of the telescope, its components, and data collection. The interests of LSST collaboration members are well served by developing and deploying digital libraries. Both C-DEBI and LSST have stated clearly their aspirations for wide accessibility and circulation of their data [42,84].

Later Stages of the Life Cycle: CENS and SDSS

Operational funding has ceased for both CENS and the first two phases of SDSS, providing opportunities to assess data management strategies at the end of project life cycles. These sites are very similar in some respects and dramatically different in others. Both projects developed new instrumentation, enabled new research questions to be asked, and produced new kinds of data for their domains. They made important contributions in their publications, fostered new collaborations, and graduated students with new expertise. They are similar organizationally, being loose confederations of researchers from participating institutions. Where they differ is in the mix of expertise, purposes for collaboration, forms of data, and relative value of their research products.

CENS was established as a multi-disciplinary center focused on developing new technologies for scientific, medical, social, and educational domains. Technology researchers benefited from access to real world problems to solve. Science and other application domain researchers benefited from new technologies to collect, analyze, and interpret their data. CENS was composed of many small projects, with data products that are small in size, large in number, heterogeneous, and complex. It also included teams that were large in size and widely distributed, with many locations for data collection, and substantial funding. While CENS had a sizable amount of funding, on a scale with some "big science" projects, the scientific activities were highly distributed and data practices were largely "little science" in character.

In CENS, data were shared mainly within research teams [119]. SensorBase, the most successful of several efforts to share data internal to CENS, was used only by a few

teams and was not sustained beyond the end of the Center. The website, which had public and private areas, was the means by which slides, images, and some datasets were shared. One of the staff reflected on how the website might have been used more effectively for sharing data:

> "This is more of an output thing than a sharing one, but it's kind of a sharing one in that, the website was grotesquely out of date throughout the entirety of CENS. And it's not that that information was not available. … in hindsight, there should have been some sort of mechanism for that sharing data to happen. … The other one that was really difficult that I see other centers being more successful with is photo repositories… Especially looking back now, because it's going to be impossible virtually to get, and ask somebody to go back to 2002 and look for a photo of XYZ is going to be virtually impossible… I mean, it's really up to the way CENS ran, it was really up to the area leader or PI to take care of that. There could've been a little bit more systematic approach at updating that stuff" (CENS-5).

SDSS is a large project in that it was planned and executed over the course of decades, with the collaboration of hundreds of individuals across 25 institutions [1], and a unified outcome, that of a 2.5-meter telescope installation in New Mexico yielding data of high velocity and volume. The legacy of CENS mostly resides in publications, collaborations, technologies (including software code), and educational outreach. While the legacy of SDSS also includes those elements, the main feature of this legacy is the SDSS datasets.

The ramping down of both projects was a time for reflection and an opportunity for partners to reassess their research directions. CENS and the first two phases of SDSS each ended officially when their project funding finished, but their research continued in other ways. Faculty at CENS were members of academic departments and of the Center, so when CENS ended they remained in their respective departments. Other than faculty members, most CENS participants were employed on grants. The few people employed by the Center were administrative staff and a few research staff; these positions terminated with the end of funding. Many of them secured positions in other departments or other CENS partner institutions. CENS students graduated, carrying their expertise and institutional memory to other academic institutions and to industry. Some of the CENS research projects continued under other funding. A number of CENS alumni started new large projects such as Mobilize and Nexleaf which came out of CENS and use similar technologies [47,91,113].

Like CENS, SDSS faculty researchers also were members of academic departments. The cohort of administrative staff in SDSS was more stable than in CENS, as many were part of the administrative structure of the larger astronomy community. Staff employed by SDSS grants to conduct research or to work on instruments, technology, and software continued on to SDSS-III, to other astronomy projects, to other domains, or entered into industry. Students funded by SDSS often graduated to SDSS partner institutions where they could continue their research. SDSS-I and II were so successful that further funding

was sought, and many partners continued to collaborate on SDSS-III and SDSS-IV. These subsequent projects add data to the existing dataset and address new research goals. The human expertise, technologies, and collaborator relationships developed in the early phases of SDSS contributed to the success of later projects and to LSST, which employs some former SDSS team members.

The disposition of data was the most pronounced difference between the projects. At CENS, the stewardship of data resources fell to individual investigators and teams rather than being an institutional priority. Researchers were not prepared to deal with their data. Datasets in CENS were seen as a means to publications, but not as products with their own value. To the question of what will happen to their data, a researcher replied, "Well, our project is going to continue for a couple of years so we haven't considered it. But yeah, I don't know what will happen with all the data that's on the servers" (CENS-6).

One reason for the minimal data release was the lack of repositories to which data might be contributed. Seismology was the only domain for which a community repository existed. Some genomic data were contributed to biological databases. Software code was sometimes deposited for public use. No comparable resources existed for ecology, environmental sciences, or most of the other CENS domain areas [119]. Publications are the primary research assets that remain available from CENS. Largely through the efforts of the CENS Data Practices team, which was the predecessor to the Knowledge Infrastructures team, the CENS publication repository was created within the University of California's eScholarship system [20,88,117]. The team also developed a data registry as part of the annual reporting system to NSF. The CENS data registry was minimally populated by CENS researchers and contains only metadata records. It was later developed into a university data registry by the UCLA library [85]. Administrative and research staff are adding metadata records for CENS datasets to this registry. A registry documents the existence of data and provides contact information for access; it does not hold datasets per se.

SDSS-I and II, in contrast, executed formal plans for the uniform acquisition, processing, and short-term delivery of their data, and started planning for long-term stewardship early in the project. As one researcher remembers:

> "It didn't take us very long to figure this out that the value of the data we were collecting was enormously great, and we needed it to... find some path to preserve it into the indefinite future… We started to say, 'Okay, it's time that we started to get a plan.' But I was actually looking through some of my documents, and it was clear that when we transitioned from SDSS-I to SDSS-II, we were already well under way of thinking about what we were going to do… You think about what do you want. You want stability. So university libraries are gonna be around for a long time" (SDSS-1).

Discussion

Managing research data is a knowledge infrastructure problem that cannot be addressed by individual researchers or projects alone, regardless of project scale or stage of life cycle. A wide range of expertise is required, as are new forms of collaborations. Standing by to accept research data at the end of a project is but one role for digital libraries and librarians. The real challenges lie in designing digital libraries to assist in the capture, management, interpretation, use, reuse, and stewardship of research data. Opportunities and challenges for the digital library community are plentiful.

The Knowledge Infrastructures project is the first study of data practices and infrastructures conducted at this scale. The four sites studied exhibit a wide array of characteristics that influence the requirements for digital libraries, especially the types and degrees of openness within these communities. The larger projects made more explicit plans for data release and invested more heavily in digital libraries systems. Datasets, as sustained in SDSS digital library systems, are their primary scientific legacy. Data resources are similarly central to the research goals of LSST. These large sites have to negotiate with multiple stakeholders to make their data accessible. In contrast, investments in digital libraries for C-DEBI began in earnest several years into the research project and have yet to be fully implemented. CENS, which was established well before funding agency requirements for data management plans were promulgated, made minimal investments in digital library systems or services. CENS researchers were willing to share their data, but had few mechanisms, incentives, or resources to do so. Negotiations about data access were more often internal to CENS or between CENS researchers and external parties who requested access to their data [119].

Our pairwise comparisons between sites reflect the problematic nature of both of our critical dimensions: scale of research and stage of life cycle. Previous studies of scientific data practices and life cycles typically either characterize the domain of study as big science [55,114] or small science [19,35,63]. Instead, our research findings suggest that projects often combine characteristics of big and small science. Data may circulate at various scales, sometimes aggregating into larger datasets and sometimes dispersing into smaller units [39].

The role of digital libraries may depend not only on the scale of data for a research project, but also on its scientific goals. The astronomy projects built digital library services into their research goals to ensure that the datasets are a legacy product. Astronomy data are reused for many years after they are collected. Much effort is devoted to the design of systems and the curation of data. In contrast, smaller science research conducted at CENS and C-DEBI is more concerned with immediate scientific breakthroughs than with the data that lead to those findings. The data are a means to an end, which is the publication of findings in scientific papers. Digital libraries may serve more transient purposes for current access to research resources in these smaller science projects.

By comparing the earlier and later stages of two astronomy surveys, LSST and SDSS respectively, we see that large telescope projects may not fully exemplify big science, as they are usually understood [89,105]. Both projects have little science characteristics that shape the big science context, and vice versa. The work to build LSST infrastructure depends upon small teams that test and evaluate various components.

The opposite is true of C-DEBI. Despite having the surface characteristics of little science, the research depends on big science facilities, namely the Integrated Ocean Drilling Program. The IODP, which is a primary source of data for C-DEBI, shares many of the hallmarks of big science, in terms of the cost [77], large-scale facilities [55], international collaboration [76], organization of the work on board expeditions [32,61], and in data and other knowledge product management practices [9,13,61].

The data life cycle in our C-DEBI case study unfolds across multiple scales, and these contexts have complex relationships to each other. The priorities and practices in one context shape, influence, enable, constrain, and mandate practices in the other. The scarce resources of IODP cruises (ship space, cores, and data) and resultant negotiations about the distribution of these resources both enable and limit the progress that can be made in individual scientists' onshore laboratories. In response, the scientists try to reconfigure their laboratory practices to bring about desired changes in the operation of the large-scale infrastructure of IODP cruises.

Another comparison between these cases relevant to digital libraries is the temporal scale of system design and data collection. In the smaller scale projects comprising CENS and C-DEBI, data management tools are selected, designed, and used by the same individuals. Technologies can be readily adapted to the problem at hand. Conversely, in the multi-decade time scale of developing larger research instruments and facilities in astronomy, data management technologies, policies, and practices are designed for anticipated future uses and users. Those developing the digital libraries may be different individuals, with different expertise, than those who curate the data. That is certainly the case with SDSS, where astronomers, computer scientists, software engineers, and other technologists designed the instruments and data collection mechanisms, then handed off the dataset to research library staff about 20 years later.

Conclusions

The Knowledge Infrastructures project is studying the transfer of people, technology, data, and knowledge within and between four distributed research sites. The particular combination of these factors in each site influences requirements for their digital libraries and for the expertise necessary to manage their data. Each of our conclusions has implications for knowledge infrastructures for science generally, and for the future directions of our own research.

Knowledge Infrastructures at Scale

The most consistent finding throughout our analyses is the influence of scale factors on data practices. By juxtaposing sites with canonical characteristics of little science and big science, many dimensions of scale come into view. These include not only the scale of research facilities and numbers of people involved, but also the numbers of teams, countries, and locales; the amount of money invested; the variety of disciplines and research domains; the volume, variety, and velocity of data; duration in time; and the scale of infrastructures required to conduct the research enterprise. The two larger science projects, the Sloan Digital Sky Survey and the Large Synoptic Survey Telescope, require large investments in facilities, partnerships across many locations and countries, and a long time frame to conduct their research. Upon closer inspection, small teams play important roles within the project work. Substantial portions of that work, especially in the early stages of the life cycle, converge into the larger knowledge infrastructure. Other parts of the work stay local, especially in the latter parts of the cycle when data can be reused in combination with other data to derive new findings. Whereas most of the work in the Center for Embedded Networked Sensing and the Center for Dark Energy Biosphere Investigations could be conducted locally, C-DEBI depends upon the large facilities of the IODP cruises for much of its data.

We found different combinations of project scale and mix of disciplines. While SDSS and LSST include participants with many specialties of astronomy, physics, and computing, project expertise tends to be within the bounds of the physical sciences and engineering. CENS and C-DEBI include participants from a broad array of physical and life sciences; the majority of CENS partners were from computer science and engineering. Whereas both SDSS and LSST have common goals to construct the technology (telescopes and instruments) and supporting methods and infrastructure for a massive and long-term sky survey, CENS and C-DEBI have more divergent goals. The latter are large endeavors, consisting of hundreds of participants over a period of five to ten years, with the goal of bringing together diverse expertise to advance the science and methods of their research domains. They explore many methods and technologies toward these ends. C-DEBI would like to have a common data repository, whereas CENS had different concerns. The broad array of disciplines in the smaller science sites leads to minimal role specialization, with each investigator and team exploring particular questions with particular methods and technologies. The self-contained nature of individual teams within CENS and C-DEBI makes it difficult for each center to converge on common goals such as data management. The narrower array of specialties and the common scientific goals of the larger projects lead to greater role specialization. To construct a telescope, convergence of methods, technologies, and data is essential.

Types and degrees of openness vary along these dimensions of scale, interacting with many other factors. Data release is central to the scientific goals of SDSS and LSST. Data repositories are part of the initial project goals and design, which leads to standardized methods of data collection and management. However, decisions about what can be released, when, and to whom vary between SDSS and LSST, and between observational data and software code. Other stakeholders, including funding agencies, may be the final

arbiters of openness. The two smaller science projects also vary on types and degrees of openness. CENS aspired to releasing more of their data, but had great difficulty finding the means to do so. C-DEBI has similar aspirations, and is developing the means to manage, share, and reuse their data. In all four of these sites, releasing software code appears easier to accomplish than is releasing research data. Again, these factors vary considerably by local circumstance.

Knowledge Infrastructures in Rhythm

Investments in knowledge infrastructures vary greatly over the life cycles of the four sites studied. Consistent with others' findings, patterns of infrastructure building are associated more with rhythms of collaboration than with life cycles per se [70,71]. In these four projects, rhythms include stages of the project and of collaborative partnerships; the maturity of tools, standards, practices, methods, and protocols; data production; careers; and funding. In SDSS and LSST, a decade or two of work precedes data collection. Astronomers may devote large portions of their careers to developing the facilities necessary to acquire the data essential for their science. Once acquired, those data must be cleaned and processed through a pipeline. Most large astronomy projects release those processed data on cycles of a year or so.

CENS and C-DEBI also adapt their practices to the time frames of their data collection technologies. Turnaround time for designing or adapting sensor technologies may be a matter of months, although some technologies took longer to develop and deploy. However, the IODP cruises on which C-DEBI depends for data require long-term commitments, which include the time to apply for ship space, participate in cruises, and process the resulting data.

The time frames of these projects also influence their types and degrees of openness. For example, SDSS and LSST investigators may have proprietary access to the data prior to their public release. LSST is proposing a multi-pronged approach to data release, which includes making unprocessed data available immediately. CENS launched long before the current pressures from funding agencies to release data at the time of publishing journal articles. Two research domains within CENS, seismology and genomics, had discipline-specific data release requirements, and these obligations were met. C-DEBI launched prior to current NSF data management plan requirements. They implemented these requirements retroactively, and are incorporating data management practices in the mid-stages of project life cycles.

Future Research Directions

The research reported here has raised as many new questions about scale, life cycle, and openness as it has answered. We continue to pursue our initial questions, along with new ones, and are exploring new sites for comparison. Although our analyses are broad in scope and rich in detail, our methods do not yield statistical comparisons. Our findings can be compared to other domains through similar methods. Most of the identified issues vary by local conditions, so they would be difficult to identify through survey methods.

Further document and network analyses are in progress. Knowledge infrastructures are complex ecologies, adapting continuously to local and global conditions and to changes in technology, policy, and stakeholders. Long-term, multi-method studies such as those reported here are necessary to understand the roles of digital libraries and digital library workforces in science.

Acknowledgements

The research reported in this paper is supported by Alfred P. Sloan Foundation Award #20113194, *The Transformation of Knowledge, Culture and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective*. We are grateful to our program officer, Joshua Greenberg, and to our external advisory board – Alyssa Goodman, George Djorgovski, and Alex Szalay – for their guidance and support. We also acknowledge the contributions of Laura A. Wynholds and David S. Fearon, Jr. for conducting early interviews; and Elaine Levia for technical, bibliographic, and administrative support.

References

[1] Abazajian, K.N., Adelman-McCarthy, J.K., Agüeros, M.A., et al. The Seventh Data Release of the Sloan Digital Sky Survey. *The Astrophysical Journal Supplement Series 182*, 2 (2009), 543–558.

[2] ADS. The SAO/NASA Astrophysics Data System. 2015. http://www.adsabs.harvard.edu/.

[3] Ahn, C.P., Alexandroff, R., Allende Prieto, C., et al. The Ninth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Baryon Oscillation Spectroscopic Survey. *The Astrophysical Journal Supplement Series 203*, 2 (2012), 21.

[4] Arzberger, P., Schroeder, P., Beaulieu, A., et al. An International Framework to Promote Access to Data. *Science 303*, 5665 (2004), 1777–1778.

[5] Astronomy and Astrophysics Survey Committee. *Astronomy and Astrophysics in the New Millennium.* National Academy of Sciences, Washington, DC, 2001.

[6] Bechhofer, S., Ainsworth, J., Bhagat, J., et al. Why Linked Data is Not Enough for Scientists. *2010 IEEE Sixth International Conference on e-Science (e-Science)*, (2010), 300–307.

[7] Bell, G., Hey, T., and Szalay, A.S. Beyond the Data Deluge (Computer Science). *Science 323*, 5919 (2009), 1297–1298.

[8] Berman, F. and Cerf, V.G. Who Will Pay for Public Access to Research Data? *Science 341*, 6146 (2013), 616–617.

[9] Bicarregui, J., Gray, N., Henderson, R., Jones, R., Lambert, S., and Matthews, B. Data Management and Preservation Planning for Big Science. *International Journal of Digital Curation 8*, 1 (2013), 29–41.

[10] Blocker, A.W. and Meng, X.-L. The potential and perils of preprocessing: Building new foundations. *Bernoulli 19*, 4 (2013), 1176–1211.

[11] Borgman, C.L. What Are Digital Libraries? Competing Visions. *Information Processing & Management 35*, 3 (1999), 227–243.

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

[13] Borgman, C.L. *Big Data, Little Data, No Data: Scholarship in the Networked World.* The MIT Press, Cambridge, Massachusetts, 2015.

[14] Borgman, C.L., Bates, M., Cloonan, M., et al. *Social Aspects of Digital Libraries. Final Report to the National Science Foundation.* 1996.

[15] Borgman, C.L., Bowker, G.C., Finholt, T.A., and Wallis, J.C. Towards a Virtual Organization for Data Cyberinfrastructure. *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM (2009), 353–356.

[16] Borgman, C.L., Darch, P.T., Sands, A.E., Wallis, J.C., and Traweek, S. The Ups and Downs of Knowledge Infrastructures in Science: Implications for Data Management. *2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, IEEE Computer Society (2014), 257–266.

[17] Borgman, C.L. and Traweek, S. The Transformation of Knowledge, Culture, and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective. 2012. http://knowledgeinfrastructures.gseis.ucla.edu/?page_id=50.

[18] Borgman, C.L., Wallis, J.C., and Enyedy, N.D. Building Digital Libraries for Scientific Data: An Exploratory Study of Data Practices in Habitat Ecology. *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries*, Springer Berlin Heidelberg (2006), 170–183.

[19] Borgman, C.L., Wallis, J.C., and Enyedy, N.D. Little Science Confronts the Data Deluge: Habitat Ecology, Embedded Sensor Networks, and Digital Libraries. *International Journal on Digital Libraries 7*, 1-2 (2007), 17–30.

[20] Borgman, C.L., Wallis, J.C., and Mayernik, M.S. Who's Got the Data? Interdependencies in Science and Technology Collaborations. *Computer Supported Cooperative Work 21*, 6 (2012), 485–523.

[21] Borgman, C.L., Wallis, J.C., Mayernik, M.S., and Pepe, A. Drowning in Data: Digital library architecture to support scientific use of embedded sensor networks. *Joint Conference on Digital Libraries*, Association for Computing Machinery (2007), 269–277.

[22] Borne, K.D. Virtual Observatories, Data Mining, and Astroinformatics. In T.D. Oswalt and H.E. Bond, eds., *Planets, Stars and Stellar Systems*. Springer Netherlands, 2013, 403–443.

[23] Boulton, G., Campbell, P., Collins, B., et al. *Science as an open enterprise.* The Royal Society, London, UK, 2012.

[24] Bowker, G.C. *Memory Practices in the Sciences.* MIT Press, Cambridge, Mass., 2005.

[25] Brunsmann, J., Wilkes, W., Schlageter, G., and Hemmje, M. State-of-the-art of long-term preservation in product lifecycle management. *International Journal on Digital Libraries 12*, 1 (2012), 27–39.

[26] Capshew, J.H. and Rader, K.A. Big Science: Price to the Present. *Osiris 7*, (1992), 2–25.

[27] Center for Dark Energy Biosphere Investigations. *C-DEBI Strategic Implementation Plan, 2010-2015.* 2010.

[28] Center for Dark Energy Biosphere Investigations. *C-DEBI Data Management Philosophy and Policy.* 2012.

[29] Center for Dark Energy Biosphere Investigations. C-DEBI. 2014. http://www.darkenergybiosphere.org/.

[30] Chompalov, I. 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*. Springer Netherlands, 2014, 167–184.

[31] CODATA-ICSTI Task Group on Data Citation Standards and Practices. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal 12*, (2013), 1–75.

[32] Collins, H.M. LIGO Becomes Big Science. *Historical Studies in the Physical and Biological Sciences 33*, 2 (2003), 261–297.

[33] Committee for a Decadal Survey of Astronomy and Astrophysics; National Research Council. *New Worlds, New Horizons in Astronomy and Astrophysics.* The National Academies Press, Washington, D.C., 2010.

[34] Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age.* National Academy Press, Washington, D.C., 2009.

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

[36] Darch, P.T. When Scientists Meet the Public: An Investigation into Citizen Cyberscience. 2011.

[37] Darch, P.T. and Borgman, C.L. Ship Space to Database: Motivations to Manage Research Data for the Deep Subseafloor Biosphere. *Proceedings of the 77th Annual Meeting of the Association for Information Science and Technology*, (2014).

[38] Darch, P.T., Borgman, C.L., Traweek, S., Cummings, R.L., Wallis, J.C., and Sands, A.E. What lies beneath?: Knowledge infrastructures in the subseafloor biosphere and beyond. *International Journal on Digital Libraries 16*, 1 (2015), 61–77.

[39] Darch, P.T. and Sands, A.E. Beyond Big or Little Science: Understanding Data Lifecycles in Astronomy and the Deep Subseafloor Biosphere. (2015).

[40] David, P.A. The economic logic of 'Open Science' and the balance between private property rights and the public domain in scientific data and information: A primer. In *The Role of the Public Domain in Scientific Data and Information*. National Academy Press, Washington, D.C., 2003, 19–34.

[41] Digital Curation Centre. What is digital curation? 2014. http://www.dcc.ac.uk/digital-curation/what-digital-curation.

[42] Edwards, K. *Center for Dark Energy Biosphere Investigations (C-DEBI): A Center for Resolving the Extent, Function, Dynamics and Implications of the Subseafloor Biosphere.* 2009.

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

[44] Edwards, P.N., Jackson, S.J., Chalmers, M.K., et al. *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges.* University of Michigan, Ann Arbor, MI, 2013.

[45] Edwards, P.N., Mayernik, M.S., Batcheller, A.L., Bowker, G.C., and Borgman, C.L. Science Friction: Data, Metadata, and Collaboration. *Social Studies of Science 41*, 5 (2011), 667–690.

[46] European Commission High Level Expert Group on Scientific Data. *Riding the wave: How Europe can gain from the rising tide of scientific data.* European Union, 2010.

[47] Exploring Computer Science. Mobilize: Mobilizing for Innovative Computer Science Teaching and Learning. 2014. http://www.exploringcs.org/about/related-grants/mobilize.

[48] Faniel, I.M. and Jacobsen, T.E. Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. *Journal of Computer Supported Cooperative Work 19*, 3-4 (2010), 355–375.

[49] Fearon Jr., D.S., Borgman, C.L., Traweek, S., and Wynholds, L.A. Curators to the Stars (Poster). *Annual Meeting of the American Society for Information Science & Technology*, (2010).

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

[51] Frieman, J. Dark Energy Survey. *Bulletin of the American Astronomical Society*, (2011), 20501.

[52] Furner, J. Little Book, Big Book: Before and After Little Science, Big Science: A Review Article, Part I. *Journal of Librarianship and Information Science 35*, 2 (2003), 115–125.

[53] Furner, J. Little Book, Big Book: Before and After Little Science, Big Science: A Review Article, Part II. *Journal of Librarianship and Information Science 35*, 3 (2003), 189–201.

[54] Galison, P. The Collective Author. *Scientific authorship: Credit and intellectual property in science*, (2003), 325–355.

[55] Galison, P. and Hevly, B.W. *Big Science: The Growth of Large-Scale Research.* Stanford University Press, Stanford, Calif., 1992.

[56] Gitelman, L., ed. *"Raw Data" Is an Oxymoron.* The MIT Press, Cambridge, Massachusetts, 2013.

[57] Glaser, B.G. and Strauss, A.L. *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Aldine Pub. Co., Chicago, 1967.

[58] Goodman, A.A., Pepe, A., Blocker, A.W., et al. Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Computational Biology 10*, 4 (2014), e1003542.

[59] Gray, J., Liu, D.T., Nieto-Santisteban, M., Szalay, A.S., DeWitt, D.J., and Heber, G. Scientific Data Management in the Coming Decade. *SIGMOD Rec. 34*, 4 (2005), 34–41.

[60] Gray, J., Slutz, D., Szalay, A.S., et al. *Data Mining the SDSS SkyServer Database.* 2002.

[61] Gray, N., Carozzi, T.D., and Woan, G. Managing Research Data in Big Science. *arXiv:1207.3923*, (2012).

[62] Greenberg, J. 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 (2009), 380–402.

[63] Heidorn, P.B. Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends 57*, 2 (2008), 280–299.
[64] Hey, A.J.G. and Trefethen, A.E. Cyberinfrastructure for e-Science. *Science 308*, 5723 (2005), 817–821.
[65] Higgins, S. The DCC Curation Lifecycle Model. *International Journal of Digital Curation 3*, 1 (2008), 134–140.
[66] Higgins, S. The lifecycle of data management. In *Managing Research Data*. Facet Publishing; 1 edition (January 31, 2012), 2012, 224.
[67] Humphrey, C. *e-Science and the Life Cycle of Research.* 2008.
[68] IODP. International Ocean Discovery Program. 2014. http://iodp.org/.
[69] Ivezic, Z., Tyson, J.A., Abel, B., et al. LSST: from Science Drivers to Reference Design and Anticipated Data Products (Version 4.0). 2014. http://arxiv.org/abs/0805.2366.
[70] Jackson, S.J. and Buyuktur, A. Who Killed WATERS? Mess, Method, and Forensic Explanation in the Making and Unmaking of Large-scale Science Networks. *Science, Technology & Human Values 39*, 2 (2014), 285–308.
[71] Jackson, S.J., Ribes, D., Buyuktur, A., and Bowker, G.C. Collaborative Rhythm: Temporal Dissonance and Alignment in Collaborative Scientific Work. *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, ACM (2011), 245–254.
[72] Johns Hopkins University. Krieger Astronomer Awarded $9.5 Million to Create "Virtual Telescope. 2013. http://krieger.jhu.edu/blog/2013/11/04/krieger-astronomer-awarded-9-5-million-to-create-virtual-telescope/.
[73] Karasti, H. and Baker, K.S. Digital Data Practices and the Long Term Ecological Research Program Growing Global. *International Journal of Digital Curation 3*, 2 (2008), 42–58.
[74] Karasti, H., Baker, K.S., and Halkola, E. 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 (2006), 321–358.
[75] Karasti, H., Baker, K.S., and Millerand, F. Infrastructure Time: Long-term Matters in Collaborative Development. *Computer Supported Cooperative Work (CSCW) 19*, 3-4 (2010), 377–415.
[76] Knorr-Cetina, K. *Epistemic Cultures: How the Sciences Make Knowledge.* Harvard University Press, Cambridge Mass., 1999.
[77] Lambright, W.H. Government and Science: A Troubled, Critical Relationship and What Can Be Done about It. *Public Administration Review 68*, 1 (2008), 5–18.
[78] Laney, D. *3D Data Management: Controlling Data Volume, Velocity and Variety".* META Group (Gartner), 2001.
[79] Latour, B. and Woolgar, S. *Laboratory Life: The Construction of Scientific Facts.* Princeton University Press, Princeton, N.J., 1986.
[80] Lenoir, T. and Hays, M. The Manhattan project for biomedicine. *Controlling Our Destinies. Historical, Philosophical, Ethical, and Theological Perspectives on the Human Genome Project*, (2000), 29–62.
[81] Liu, X., Wang, Q., and Zhou, Z. IODP in Japan. *Advance in Earth Sciences 4*, (2004), 10.

[82] LSST. Large Synoptic Survey Telescope: Timeline. 2013. http://www.lsst.org/lsst/science/timeline.

[83] LSST Collaboration. Community Science Input and Participation. *Large Synoptic Survey Telescope*, 2013. http://www.lsst.org/lsst/science/participate.

[84] LSST Science Collaboration, Abell, P.A., Allison, J., et al. *LSST Science Book, Version 2.0.* 2009.

[85] Mandell, R.A. *Researchers' Attitudes towards Data Discovery: Implications for a UCLA Data Registry.* Social Science Research Network, Rochester, NY, 2012.

[86] Maurer, B.A. Models of Scientific Inquiry and Statistical Practice: Implications for the structure of scientific knowledge. In *The Nature of Scientific Evidence: Statistical, philosophical, and empirical considerations*. The University of Chicago Press, Chicago, 2004, 17–50.

[87] Mayernik, M.S. Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators. 2011. http://dx.doi.org/10.2139/ssrn.2042653.

[88] Mayernik, M.S., Wallis, J.C., and Borgman, C.L. Unearthing the Infrastructure: Humans and Sensors in Field-Based Research. *Computer Supported Cooperative Work 22*, 1 (2013), 65–101.

[89] McCray, W.P. *Giant Telescopes: Astronomical Ambition and the Promise of Technology.* Harvard University Press, Cambridge, MA, 2004.

[90] Meyer, E.T. and Schroeder, R. *Knowledge Machines: Digital Transformations of the Sciences and Humanities.* MIT Press, Cambridge, MA, 2015.

[91] Nexleaf. 2013. http://nexleaf.org/about-us-0.

[92] Onsrud, H. and Campbell, J. Big Opportunities in Access to "Small Science" Data. *Data Science Journal 6*, (2007), OD58–OD66.

[93] Orcutt, B.N., LaRowe, D.E., Biddle, J.F., et al. Microbial Activity in the Marine Deep Biosphere: Progress and Prospects. *Extreme Microbiology 4*, (2013), 189.

[94] Palmer, C.L., Cragin, M.H., Heidorn, P.B., and Smith, L.C. Data Curation for the Long Tail of Science: The Case of Environmental Sciences. (2007).

[95] Parsons, M.A. and Fox, P.A. Is Data Publication the Right Metaphor? *Data Science Journal 12*, (2013), WDS32–WDS46.

[96] Pepe, A., Goodman, A., Muench, A., Crosas, M., and Erdmann, C. How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. *PLoS ONE 9*, 8 (2014), e104798.

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

[98] Price, D.J. de S. *Little Science, Big Science.* Columbia University Press, New York, NY, USA, 1963.

[99] Ray, J.M., ed. *Research Data Management: Practical Strategies for Information Professionals.* Purdue University Press, West Lafayette, Ind, 2014.

[100] Renear, A.H., Sacchi, S., and Wickett, K.M. Definitions of Dataset in the Scientific and Technical Literature. *Proceedings of the American Society for Information Science and Technology 47*, 1 (2010), 1–4.

[101] Ribes, D. and Jackson, S.J. Data Bite Man: The Work of Sustaining a Long-Term Study. In L. Gitelman, ed., *"Raw Data" Is an Oxymoron*. The MIT Press, Cambridge, MA, 2013, 147–166.

[102] Sands, A.E., Borgman, C.L., Traweek, S., and Wynholds, L.A. We're Working On It: Transferring the Sloan Digital Sky Survey from Laboratory to Library. *International Journal of Digital Curation 9*, 2 (2014), 98–110.

[103] Schofield, P., Eppig, J., Huala, E., et al. Sustaining the data and bioresource commons. *Science 330*, (2010), 592–593.

[104] SDSS. Sloan Digital Sky Survey. 2015. http://www.sdss.org/.

[105] Smith, R.W. The Biggest Kind of Big Science: Astronomers and the Space Telescope. In P. Galison and B.W. Hevly, eds., *Big science: the growth of large-scale research*. Stanford University Press, Stanford, Calif., 1992, 184–211.

[106] Suber, P. *Open Access.* MIT Press, Cambridge, Mass, 2012.

[107] Szalay, A.S. Jim Gray, astronomer. *Communications of the ACM 51*, (2008), 59–65.

[108] Thakar, A.R., Szalay, A.S., Fekete, G., and Gray, J. The Catalog Archive Server Database Management System. *Computing in Science & Engineering 10*, 1 (2008), 30–37.

[109] Traweek, S. *Beamtimes and Lifetimes: The World of High Energy Physicists.* Harvard University Press, Cambridge, Mass., 1988.

[110] Traweek, S. Big Science as Colonialist Discourse: Regional Differences in Japanese High Energy Physics. In P. Galison and Bruce William Hevly, eds., *Big Science: The Growth of Large-scale Research*. Stanford University Press, Stanford, Calif., 1992, 100–128.

[111] Traweek, S. Border Crossings: Narrative Strategies in Science Studies and Among High Energy Physicists at Tsukuba Science City, Japan. In *Science as Practice and Culture*. University of Chicago Press, Chicago, 1992, 429–465.

[112] Traweek, S. Generating High-Energy Physics in Japan: Moral Imperatives of a Future Pluperfect. In D. Kaiser, ed., *Pedagogy and the Practice of Science: Historical and Contemporary Perspectives*. The MIT Press, Cambridge, MA, 2005, 357–392.

[113] UCLA OIP. Social Entrepreneurship: Nexleaf Takes it to the Next Level. *The Inventor: Intellectual Property News 2*, 3 (2010).

[114] Vermeulen, N. *Supersizing Science: On Building Large-Scale Research Projects in Biology*. Universal Publishers, Boca Raton, FL, 2010.

[115] Wallis, J.C. The Distribution of Data Management Responsibility within Scientific Research Groups. 2012. http://search.proquest.com/docview/1029942726/abstract?accountid=14512.

[116] Wallis, J.C. and Borgman, C.L. Who is Responsible for Data? An Exploratory Study of Data Authorship, Ownership, and Responsibility. *Annual Meeting of the American Society for Information Science & Technology*, Information Today (2011), 1–10.

[117] Wallis, J.C., Borgman, C.L., Mayernik, M.S., and Pepe, A. 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 (2008), 114–126.

[118] Wallis, J.C., Borgman, C.L., Mayernik, M.S., Pepe, A., Ramanathan, N., and Hansen, M.A. Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries. *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries*, Berlin: Springer (2007), 380–391.

[119] Wallis, J.C., Rolando, E., and 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 (2013), e67332.

[120] Wray, K.B. Scientific authorship in the age of collaborative research. *Studies in History and Philosophy of Science Part A 37*, 3 (2006), 505–514.

[121] Wynholds, L.A., Fearon, D.S., Borgman, C.L., and Traweek, S. When Use Cases Are Not Useful: Data Practices, Astronomy, and Digital Libraries. *Proceedings of the 11th Annual Joint Conference on Digital Libraries*, ACM (2011), 383–386.

[122] Wynholds, L.A., Wallis, J.C., Borgman, C.L., Sands, A.E., and Traweek, S. Data, Data Use, and Scientific Inquiry: Two Case Studies of Data Practices. *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, Association for Computing Machinery (2012), 19–22.

[123] York, D.G., Adelman, J., Anderson, J.E., et al. The Sloan Digital Sky Survey: Technical Summary. *Astronomical Journal 120*, (2000), 1579–1587.

[124] Zooniverse. Galaxy Zoo. 2014. http://www.galaxyzoo.org/.