September, 2014

The Ups and Downs of Knowledge Infrastructures in Science: Implications for Data Management

Christine L Borgman, *University of California, Los Angeles*
Peter T Darch, *University of California, Los Angeles*
Ashley E Sands, *University of California, Los Angeles*
Jillian C Wallis, *University of California, Los Angeles*
Sharon Traweek, *University of California, Los Angeles*

The Ups and Downs of Knowledge Infrastructures in Science: Implications for Data Management

Christine L. Borgman, Peter T. Darch, Ashley E. Sands, Jillian C. Wallis, Sharon Traweek
Knowledge Infrastructures Project
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
borgman@gseis.ucla.edu, petertdarch@gmail.com, ashleysa@ucla.edu, jwallisi@ucla.edu, traweek@history.ucla.edu

ABSTRACT
The promise of technology-enabled, data-intensive scholarship is predicated upon access to knowledge infrastructures that are not yet in place. Scientific data management requires expertise in the scientific domain and in organizing and retrieving complex research objects. The Knowledge Infrastructures project compares data management activities of four large, distributed, multidisciplinary scientific endeavors as they ramp their activities up or down; two are big science and two are small science. Research questions address digital library solutions, knowledge infrastructure concerns, issues specific to individual domains, and common problems across domains. Findings are based on interviews (n=113 to date), ethnography, and other analyses of these four cases, studied since 2002. Based on initial comparisons, we conclude that the roles of digital libraries in scientific data management often depend upon the scale of data, the scientific goals, and the temporal scale of the research projects being supported. Digital libraries serve immediate data management purposes in some projects and long-term stewardship in others. In small science projects, data management tools are selected, designed, and used by the same individuals. In the multi-decade time scale of some big science research, data management technologies, policies, and practices are designed for anticipated future uses and users. The need for library, archival, and digital library expertise is apparent throughout all four of these cases. Managing research data is a knowledge infrastructure problem beyond the scope of individual researchers or projects. The real challenges lie in designing digital libraries to assist in the capture, management, interpretation, use, reuse, and stewardship of research data.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital libraries – *collection, dissemination, standards, user issues.*

2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL), September 8-12, 2014, London, UK, Copyright 2014 ACM 1-58113-000-0/00/0010 …$15.00.

General Terms
Management, Documentation, Design, Economics, Reliability, Human Factors, Standardization.

Keywords
Big data, big science, little science, small science, digital libraries, knowledge infrastructures, astronomy, biology, sensor networks, data management

1. INTRODUCTION
Modern sensor networks, satellites, telescopes, and laboratory instruments can collect vastly more data, at far faster rates and far greater variety, than ever before. Scientific methods and the organization of collaborative work must adapt to the volumes and diversity of data being generated. As data are combined from multiple sources and are mined for new interpretations, the challenges of designing effective data management strategies multiply.

Scientific data management requires deep expertise in scientific theory, method, instrumentation, interpretation, and knowledge organization. The relevant expertise is complex and divided differently within each field and specialty. Each step in data handling requires knowledge and judgment of the steps that went before. Necessary details of data provenance often go undocumented, leaving researchers in the position of making inferences with insufficient information [1]. Minute differences in calibration, miniscule artifacts in a data stream, and other perturbations may be spotted by those closest to the research design – but these factors decrease in visibility the farther the interpreter lies from the source of the data.

Requirements of funding agencies and journals to release research data highlight the complexity of modern science: not only the contested notion of data, but competing views of research, innovation, and scholarship, disparate incentives for collecting and releasing data, the economics and intellectual property of research products, and public policy. The promise of technology-enabled, data-intensive scholarship is predicated upon available systems, services, tools, content, policies, practices, and human resources to discover, mine, and use research products. Not only is this infrastructure not yet in place, it is not yet clear what should be built or how to build it. Digital libraries are a small but important part of the solution [2]–[4]. We take a broad view of digital libraries, spanning the range from local systems for managing research data to large-scale public data repositories.

Digital libraries can be deployed from the initial stages of data collection through archiving and preservation.

Big science, such as astronomy and the biosciences, is characterized by international, collaborative efforts that produce vast amounts of data. These data often are big in volume and velocity, but may be homogenous in form and structure. Small science, sometimes called little science, is typified by heterogeneous methods, diverse forms of data, and by local control and analysis. Data management concerns and practices appear to differ greatly between big and small science [5]–[7].

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 cases studied are big science and two are small science. Two are in the process of ramping down their data collection and active research and the other two are ramping up their research activities.

Here we frame the Knowledge Infrastructures project, explain the research questions, outline the research methods, and present initial comparisons of the four case studies. More detailed analyses of the individual cases are presented elsewhere and others are in progress.

2. Why and How to Study Knowledge Infrastructures in Science

Managing research data is difficult. Making research data useful to unknown others, for unanticipated purposes, is far harder. As researchers reach the limits of available tools and resources to collect, interpret, and manage their data, they hit the scaling problem. Having more data requires different tools and different questions. 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.

The term *knowledge infrastructures* builds upon earlier developments in information, infrastructure, and the Internet. 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. Paul Edwards [8, p. 17] defined *knowledge infrastructures* as "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds." This scope has been extended to include technology, intellectual activities, learning, collaboration, and distributed access to human expertise and to documented information [4].

2.1 Motivation for the Knowledge Infrastructures Project

Infrastructure for research data is much more than disseminating resources; it must support data collection, analysis, use, and reuse for new scientific methods and also improve access to information. Knowledge infrastructures are expensive to construct and maintain. The value proposition and burden of costs are much debated [9]. Their design rests on the ability to explicate the socio-technical structures that are embodied in the data, the practices, technical arrangements, and policies. These interdependencies present significant risks to adoption and implementation of effective infrastructure. Among the digital library challenges in managing research data are granularity, provenance, structures, identity, identifiers, and functions of data [10]–[12].

While countless policy reports call for the building of information infrastructure and capacity for research data, only a handful of researchers consider how knowledge of 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 [2], [8], [13]–[18].

2.2 Digital Libraries, Data, and Knowledge Infrastructures

Digital libraries, whether for data or documents, typically serve as repositories for content no longer in active use by its creators. That narrow view limits the application of digital libraries for scientific data management. They can be part of the solution when conceived as systems that encompass the entire information life cycle [19]. Digital libraries originated with textual content and expanded quickly to include multi-media resources and research data. Design requirements vary content type, user community, and other factors. Managing data requires a much different architecture than systems for publications or other textual 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 [2], [4], [20].

2.3 Research Questions

The Knowledge Infrastructures project addresses four questions across the four research sites:

- What new infrastructures, divisions of labor, knowledge, and expertise are required for data-intensive science?
- How are the infrastructures of multi-disciplinary, data-intensive scientific endeavors established and how are they dismantled?
- How do data management, curation, sharing, and reuse practices vary among research areas?
- What data are most important to curate, from whose perspective, and who decides?

For this paper, we focus on initial comparisons of the four cases and the implications for the design of digital library systems and services. 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 are specific to individual domains and those that are common across domains.

2.4 Research methods

The four cases vary by stage of project and scale of the data-intensive research, as presented in Figure 1. The two research projects that produce large volumes of relatively homogeneous data are the *Sloan Digital Sky Survey* (SDSS) and the *Large Synoptic Survey Telescope* (LSST). The two projects that produce smaller amounts of relatively heterogeneous data are the *Center for Embedded Network Sensing* (CENS) and the *Center for Dark Energy Biosphere Investigations* (C-DEBI). The other comparison is between projects in earlier stages of their life cycles (C-DEBI and LSST) that are ramping up data production and projects at

later stages of their life cycles (CENS and SDSS) that have completed data collection. Research on the two ramping down projects, CENS and SDSS, laid the foundation for comparisons. We began research on CENS in 2002 [5], [21]–[25] and on SDSS in 2009 [26]–[28] with grants from the National Science Foundation and other sources. The Knowledge Infrastructures Project launched in January, 2012, with funding from the Alfred P. Sloan Foundation. Interviews and ethnographic work on C-DEBI began in 2012 and on LSST in 2013.

These comparisons are used to assess knowledge infrastructure requirements for a broad spectrum of scientific research and practice. We are studying knowledge transfer within and between projects, among scientists, and between scientists and information professionals. Findings about effective and ineffective strategies for data management will contribute to recommendations for digital library technologies, practices, and policies. Prior studies of scientific practices and infrastructure, while many in number, tend to focus on specific cases. The Knowledge Infrastructures project is the first study of data practices and infrastructures conducted at this scale, spanning more sites, more research subjects, and a longer time frame.

Figure 1: Cases by scope of data and stage of life cycle

	Big Data	**Small Data**
Ramping up data collection	LSST	C-DEBI
Ramping down data collection	SDSS	CENS

Research on each case has been performed with a mix of methods, including semi-structured and unstructured interviews, ethnographic participant-observation, and document analysis. The CENS comparisons presented here are drawn from a round of 34 semi-structured interviews collected in 2012-2013, 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 38 interviews conducted with 35 participants, five weeks of ethnographic participant-observation at one of the SDSS data management sites, and analysis of publicly available webpages and memoranda. For C-DEBI, we draw from a round of 47 interviews and participation in the development of data management infrastructure. For LSST, we draw on background research and initial observations. Analytical coding of interview transcripts, fieldnotes, and documents was done in NVIVO 9, a qualitative analysis software package, and analyzed for emergent themes using grounded theory [29].

3. Findings

Findings are presented in two parts. First, each of the four cases are analyzed with respect to data management practices that may or may not be amenable to digital library solutions. Second, we make comparisons between the four projects to examine the implications for knowledge infrastructure requirements. We distinguish between data management issues that are specific to individual domains and those that are common across the domains.

3.1 Case studies

The four case studies are in different stages of development. We have studied CENS throughout its decade-long lifespan (2002-2012) and continue to study its legacy. Research questions about the Sloan Digital Sky Survey address data practices, knowledge transfer, and workforce development. These questions have evolved through several grant projects since 2009. Background research on the Large Survey Synoptic Telescope began in 2009 and fieldwork in 2013. Research on C-DEBI began in 2012 and data collection is largely complete. The short descriptions of individual cases provide the framing necessary to make the digital library comparisons.

3.1.1 Center for Embedded Networked Sensing (CENS)

CENS (2002-2012) was a 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-docs, 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 75-80% 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.

Scientific research in CENS, as conducted in field deployments, were heterogeneous in character. Sensor networks produced far more data than did the hand-sampling methods that dominated these domains. As the volume and velocity of data increased, science teams encountered scaling problems that their current 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 [23], [24], [30].

CENS data management problems were less amenable to digital library solutions than expected. Interest in a common data repository was minimal due to the diversity of data and lack of need to pool data 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 [18], [30]–[32].

In CENS, data was 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 no intent to recombine them later. Rarely were data kept for reuse beyond the teams that collected them. The majority of participants were technology researchers whose scholarly products were papers, instruments, and software. Most researchers maintained their data locally. Relatively little CENS data was shared outside these collaborations. 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 [23], [25], [30]

3.1.2 Sloan Digital Sky Survey (SDSS)

Astronomy sky surveys are research projects to capture large and detailed amounts of data about a region of the sky. The Sloan Digital Sky Survey, named for its largest funder, the Alfred P. Sloan Foundation, is notable for its commitment to timely data releases to the public. The SDSS website lists more than five hundred papers published up to mid-2009 that use SDSS data or are about SDSS. The actual number of papers using their data to date is probably several thousand, given the many citations to SDSS data papers and the common practice of reusing public data without citing them in publications [28], [33].

SDSS planning began in the 1990s. Survey data collection began in 2000, mapping about one-quarter of the night sky with a focus on galaxies. Data were collected by a 2.5 meter optical telescope at Apache Point Observatory in New Mexico. The first phase of the SDSS project (SDSS I) ran from 2000-2005 and the second (SDSS II) from 2005-2008. Each was funded as an independent project; SDSS II expanded the scientific goals and broadened participation. In a 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 continues with largely new leadership, collaborating institutions, and scientific goals. SDSS-III is collecting data through summer 2014; initial funding for SDSS IV was acquired in 2012 [34]–[37].

Our research examines SDSS-I/II and the dataset that resulted from initial funding phases. The SDSS-I/II project is now in its archival phase. In 2008, four Memoranda of Understanding (MOU) established how the collection would be managed for the subsequent five years, until early 2014. The SDSS investigators chose to migrate the dataset, which is about 130 terabytes in size, from the national laboratory previously hosting the data to two university research libraries. The libraries collaborated with SDSS astronomers to ensure proper management of the data during the MOU period [38].

3.1.3 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 September 2010 [39]. 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. C-DEBI is massively distributed across 40 or so institutions in the USA and Europe and is highly interdisciplinary. It serves as an exemplar of the complexity of data-intensive small science.

C-DEBI scientists collect and analyze physical samples from beneath 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, the most significant of which 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 [40]. Data are processed on board ship, in laboratories, and in other field sites. Samples are distributed widely across the investigators, projects, and sites. Physical samples may be stored in repositories for the long term.

Complex relationships between the domain of study and the IODP have shaped the data management challenges facing C-DEBI and their responses to these challenges. The IODP is the latest iteration in a series of scientific ocean drilling cruise programs operating since the late 1960s. Initially, only researchers from the physical science disciplines could participate in these cruises. Only in the latter 1990s did microbiologists gain access to IODP cruises. Research space – for people, equipment, and data collection time – is scarce on scientific ocean drilling cruises. C-DEBI scientists, many of whom are microbiologists, must compete against other projects and disciplines. We are finding that the C-DEBI infrastructure for data management is being designed for access to IODP resources in addition to advancing their own scientific work per se. These social considerations motivate the very construction of this infrastructure and the choice of features [41].

The launch of C-DEBI afforded opportunities to observe how the work of negotiating, building, and maintaining data management practices unfolds in a new collaborative setting. C-DEBI is developing and implementing a data portal that will be a repository for datasets produced by their scientists. Members of the Knowledge Infrastructures team have been involved in this development process, enabling us to understand how C-DEBI partners negotiate what data are to be managed, who decides, and who is responsible for each part of management and curation.

3.1.4 Large Synoptic Survey Telescope (LSST)

The Large Synoptic Survey Telescope is a massive astronomy project that is building a ground-based telescope in Chile [42]. Planned as the next major sky survey, LSST is due to launch a decade-long phase of data collection in 2023, generating 30 terabytes of data nightly. Unlike previous sky surveys, the LSST aims to combine images and data about multiple domains of the universe, including galaxies, the Milky Way, and the Solar System, into a single dataset [43]–[45].

Initial discussions about the LSST began in the early 1990s, and by 2001 the LSST was one of seven Prioritized Major Initiatives in National Research Council's decadal survey of astronomy. The LSST Corporation was formed in 2003. Google joined in 2007 to assist in making the data publicly accessible. In 2012, the National Science Board approved funding for the final design stage .

The LSST is headquartered at the University of Arizona in Tucson, with significant aspects of the work based at other sites, including the Stanford Linear Accelerator Center, the University of Washington, and the University of California, Davis.. Scientists at nineteen national laboratories and universities are involved in building the LSST telescope. Eleven scientific projects are under way to plan the analysis of LSST data that address research questions in multiple domains of astronomy.

A significant amount of data-intensive work has already been accomplished on the LSST project, including simulations to test the infrastructure. Studying these processes enables the Knowledge Infrastructures project to understand how data management practices are being negotiated, resolved, and embedded as the LSST moves towards its data collection phase.

3.2 Comparisons of cases

Here we make pairwise comparisons of the four cases, focusing first on the stages of the projects and then on the scale of the data handled. 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. We consider which data management 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.

3.2.1 Ramping Down: CENS and SDSS

Operational funding has ceased for both CENS and SDSS I/II. These projects 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, fostering collaborations, and graduating students with new expertise. They were similar organizationally, as both were loose confederations of researchers from participating institutions. Where they differed was in the mix of expertise, purposes for collaboration, forms of data, and relative value of their research products.

CENS was 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 embodies small science, with data products that are small in size, large in number, heterogeneous, and complex. SDSS, in contrast, is an astronomical survey devoted to collecting the highest quality data possible, using instrumentation that continued to improve over the course of the research. Participants were largely from astronomy and astrophysics, but also included collaborators from computer science, statistics, and other domains. Collaborators benefited from early access to data and roles in the design of the project. SDSS is big science, yielding data of high velocity and volume. The legacy of CENS resides in publications, people, and technologies, whereas the legacy of SDSS resides in these and in the data, which continue to be used well beyond the initial research project.

The ramping down of both projects was a time for reflection and an opportunity for partners to reassess their research directions. CENS and SDSS-I/II 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. Administrative staff members were the only people employed by the Center, so their positions terminated. 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. One large project, Mobilize, carried forward with a substantial number of CENS alumni. SDSS I/II was so successful that many collaborations continued under SDSS III, which continues to add data to the existing dataset while addressing new research goals.

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, or to other domains. Students funded by SDSS often graduated to SDSS partner institutions where they could continue their research.

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. Data release and sharing consisted largely of private exchanges between collaborators, outside of the few domain areas where repositories existed. Software code was sometimes deposited for public use [25]. 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. 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. Administrative and research staff are adding metadata records for CENS datasets to this registry.

SDSS I/II, in contrast, executed formal plans for stewardship of their data. Data from SDSS-I/II were transferred to two academic libraries for longer term curation. Several astronomy departments also have backup copies. These data management groups brought distinctive kinds of expertise to the long-term stewardship of the data resources. By implementing multiple, complementary methods of storing, curating, and accessing the SDSS data, the SDSS investigators are ensuring that data remain scientifically useful for as long as possible.

3.2.2 Ramping Up: 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 in C-DEBI is relatively brief compared with the two decades from initial conception of the LSST to the anticipated commencement of data production. 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.

Another difference is the heterogeneity of expertise in the two projects. C-DEBI scientists come from a wide range of scientific disciplines, which contributes to greater diversity of data practices along three dimensions: the types of datasets produced; the methods used in producing similar types of datasets; and recordkeeping practices about the methods used to generate datasets. This mix of practices makes data management in C-DEBI particularly challenging. 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.

A third difference is the infrastructure of these communities. Astronomy has the most sophisticated and coordinated infrastructure of any we have seen. Their data management practices also are more mature and standardized across the field than in the fields comprising C-DEBI. As a result, LSST partners are more able to draw on pre-existing practices and infrastructures. Conversely, however, these pre-existing astronomy practices and infrastructures also act as constraints on LSST research practices, whereas C-DEBI partners have more flexibility to design solutions that meet their needs. Comparing C-DEBI and LSST allows us to see ways in which infrastructures

and data practices of the respective domain sciences are resources for individual investigators, projects, and research sites.

Another critical difference is the significance of the collected data for the research fields as a whole. For SDSS and LSST, the data are the primary justification for the scale of the projects, although the scientific questions those data are expected to address in the long term are important, too. For C-DEBI, the immediate scientific results are the legacy, and data play instrumental roles in the production of these results. This distinction influences relationships between researchers and their data. In C-DEBI, credit accrues for scientific publications. In the current design phase of the LSST, scientific credit is tied to successful data management and simulation of the future operation of the telescope, its components, and the data collection process. LSST members are motivated to manage the data effectively for future users.

While C-DEBI is usually regarded as small science, the Integrated Ocean Drilling Program – which plays a critical role in the first stages of the data life cycle – shares many of the hallmarks of big science: it provides large-scale infrastructure and employs standardized practices for the collection and curation of data and samples. Conversely, the LSST data are likely to be used by small teams who manipulate subsets of data to produce new datasets that they will handle in ways characteristic of small science. Therefore, both projects appear to be a mix of big science and small science, requiring more sophisticated ways of thinking about data-intensive collaboration in science.

3.2.3 Small 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 are small science, involving the generation of varied, small-scale datasets.

They also have important differences, the most significant of which is the mix of small and big science. Almost all the data produced and used by CENS researchers is characteristic of small science, with exceptions such as genomic data on marine life and seismic data, both of which are contributed to repositories. Most C-DEBI research is conducted in a small science manner, but some of the research depends on data of big science origins, namely the Integrated Ocean Drilling Program (IODP). The data generated on IODP expeditions about the physical properties of samples (seawater and seafloor sediments and basalts) are highly structured, professionally curated to stringent standards, and archived in public databases. We are examining interactions between the IODP standards and the day-to-day data practices of C-DEBI researchers. By using IODP data, more stakeholders, with a greater array of research interests, are involved in data management.

Another significant difference is that CENS focused on developing emergent technologies to support scientific work, whereas C-DEBI focuses on emergent scientific problems. In CENS, most researchers were more concerned with the operation of the sensors than with the data they generated. Conversely, few C-DEBI scientists are interested specifically in the tools that support their research. Information about these tools will be important for the subsequent interpretation and reuse of scientific data generated by C-DEBI, but it is difficult to see whose interests are served by the collection, storage, and curation of such information. The same was true in CENS with respect to datasets that resulted from joint field research. CENS technology researchers were interested in data that documented sensor operations, whereas the scientists were more interested in the scientific data generated by the sensors. Consequently, the data practices of technology and science researchers in CENS were independent, whereas their research interests were interdependent [23].

C-DEBI produces biological data about microbes and physical data about the environment which these microbes inhabit. These data are often correlated to track the impact of the microbes on their environment and vice versa. However, where different types of data are produced by different scientists and managed in different contexts, interoperability can be problematic and can influence subsequent reproduction and verification of scientific analyses. Practices also diverge between digital data and physical samples. Whereas physical samples, such as cores from the ocean floor, may be kept indefinitely, digital records of analyses that lead to papers may be kept only for short periods of time.

Similarities between C-DEBI and CENS allow us to examine more closely the conditions under which researchers in small science domains share data. CENS researchers were generally willing to share data, but most exchange was between individuals, and conditions often were attached to data release [22], [25]. Findings on C-DEBI data sharing are not yet available. We are studying the interplay of technology, infrastructure, and social factors that influence data release and sharing.

Our engagement in CENS and C-DEBI introduces another type of comparison between these small science projects. We have been involved in the design of data management infrastructure for both projects. The embeddedness of our research team in CENS has resulted in many insights about tools and infrastructure for data management in distributed, multidisciplinary, small science projects [30], [46], [47]. Further comparisons between CENS and C-DEBI will be reported in subsequent papers.

3.2.4 Big Science: SDSS and LSST

The Sloan Digital Sky Survey and Large Synoptic Survey Telescope are large-scale infrastructure projects to generate massive datasets in astronomy. In both of these projects, new instrumentation and data will be 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. Differences in scientific goals also contribute to choices of instruments, areas of the sky to scan, types of data to collect, and rates of collection.

As we are in the early stages of studying LSST, comparisons to other cases are the most speculative. We are studying the types and degrees of knowledge that are transferred from the SDSS to LSST projects. They have many personnel in common and are facing some similar challenges in terms of data collection, curation, and analytical methods at unprecedented scales. Differences in funding, scientific goals, technologies, collaboration, and other factors are likely to influence the infrastructure required for LSST. Their broader scientific goals may require even more complex negotiations for collaboration and data management than was evident in SDSS.

4. Discussion and Conclusions

The data management challenges of these four projects have a range of implications for the design, use, and sustainability of digital libraries. The role of digital libraries may depend not only on the scale of data for a research project, but on its scientific goals. Big science projects in astronomy build digital library services into the goals of the research to ensure that the data are central to their legacy. Astronomy data are reused by the research

community for many years after they are collected. Much effort is devoted to the design of systems and the curation of data. In contrast, small science research such as CENS and C-DEBI is more concerned with scientific breakthroughs than with the data that lead to those findings. The data are a means to an end, which is the resulting scientific papers, and not an end in themselves. Relatively little of CENS or C-DEBI digital data are kept for long-term reuse by parties outside of these research centers. Digital libraries may serve more transient purposes for current access to research resources in these small science projects.

Another comparison between these cases that is relevant to digital libraries is the temporal scale of system design and data collection. In the small science projects of 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 big science research 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.

Across these cases, we identified the transfer of people, technology, data, and knowledge. The particular combination of these will influence the design of individual digital libraries and the necessary staffing. To obtain funding, investigators must pitch radically new, rather than incremental, scientific goals. As a result, they may risk reinvention rather than learning from the strengths and weaknesses of projects that went before – digital library and otherwise.

These four cases can be compared along many dimensions. In this short paper, we have focused on comparisons between big and small science and the state of their infrastructure, whether ramping up or down. Future papers will address differences in governance, scientific goals, communities of practice, and other factors. Although our case studies are broad in scope and rich in detail, our sample is not large enough to draw statistical comparisons. Some of our methods and findings could be transferred to other domains through case studies or large surveys. However, since many of the issues identified depend heavily upon local conditions, such as those relating to exchange of datasets, generic surveys risk missing important context. Other quantitative methods such as network analysis can complement case studies [48]. We are currently analyzing the email archives of CENS and SDSS, to determine how topical conversations evolved over the course of projects and to identify communication patterns among collaborators.

The need for library, archival, and digital library expertise is apparent throughout all four of these projects. Concepts as fundamental as the distinction between backup and curation remain opaque to many scientists. Researchers bring expertise in their science domain and in their methods. Rarely do they bring expertise in knowledge organization and data management to their collaborations. Some are willing and able to bring digital library expertise to their projects. Others may focus on short term technical solutions to data management challenges that really are long-term problems. Research libraries have become more proactive in acquiring datasets large and small. They, too, are learning the strengths and limitations of their digital library expertise.

Managing research data is a knowledge infrastructure problem and not one that can be addressed by individual researchers or projects alone, whether big science or small. 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. 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.

5. Acknowledgements

The research reported in this paper is supported by 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; Milena Golshan for additional data analysis; and Elaine Levia for technical, bibliographic, and administrative support.

6. REFERENCES

[1] A. W. Blocker and X.-L. Meng, "The potential and perils of preprocessing: Building new foundations," *Bernoulli*, vol. 19, no. 4, pp. 1176–1211, 2013.

[2] C. L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge MA: MIT Press, Forthcoming.

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

[4] P. N. Edwards, S. J. Jackson, M. K. Chalmers, G. C. Bowker, C. L. Borgman, D. Ribes, M. Burton, and S. Calvert, "Knowledge Infrastructures: Intellectual Frameworks and Research Challenges," University of Michigan, Ann Arbor, MI, May 2013.

[5] C. L. Borgman, J. C. Wallis, and N. D. Enyedy, "Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries," *Int J Digit Libr*, vol. 7, no. 1–2, pp. 17–30, Oct. 2007.

[6] M. H. Cragin, C. L. Palmer, J. R. Carlson, and M. Witt, "Data sharing, small science and institutional repositories," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 368, pp. 4023–4038, 2010.

[7] M. L. Taper and S. R. Lele, Eds., "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*, Chicago: University of Chicago Press, 2004, pp. 17–50.

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

[9] F. Berman and V. G. Cerf, "Who Will Pay for Public Access to Research Data?," *Science*, vol. 341, no. 6146, pp. 616–617, Aug. 2013.

[10] G. Bell, T. Hey, and A. Szalay, "Beyond the Data Deluge (Computer Science)," *Science*, vol. 323, no. 5919, pp. 1297–1298, Mar. 2009.

[11] 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*, vol. 12, pp. 1–75, 2013.

[12] A. H. Renear, S. Sacchi, and K. M. Wickett, "Definitions of Dataset in the Scientific and Technical Literature," *Proc. Am. Soc. Info. Sci. Tech.*, vol. 47, no. 1, pp. 1–4, Nov. 2010.

[13] I. M. Faniel and T. E. Jacobsen, "Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data," *Comput Supported Coop Work*, vol. 19, no. 3–4, pp. 355–375, Sep. 2010.

[14] H. Karasti, K. S. Baker, and F. Millerand, "Infrastructure Time: Long-term Matters in Collaborative Development," *Computer Supported Cooperative Work (CSCW)*, vol. 19, no. 3–4, pp. 377–415, Aug. 2010.

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

[16] B. Latour and S. Woolgar, *Laboratory Life: The Construction of Scientific Facts*, 2nd ed. Princeton, N.J.: Princeton University Press, 1986.

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

[18] J. C. Wallis, C. L. Borgman, M. S. Mayernik, and A. Pepe, "Moving archival practices upstream: An exploration of the life cycle of ecological sensing data in collaborative field research," *IJDC*, vol. 3, no. 1, pp. 114–126, 2008.

[19] C. L. Borgman, M. Bates,, M. Cloonan, E. Efthimiadis, A. Gilliland-Swetland, Yasmin B. Kafai, Gregory H. Leazer, and Anthony B. Maddox, "Social Aspects of Digital Libraries. Final Report to the National Science Foundation," Background paper for UCLA - National Science Foundation Workshop, 1996.

[20] M. A. Parsons and P. A. Fox, "Is Data Publication the Right Metaphor?," *Data Science Journal*, vol. 12, pp. WDS32–WDS46, 2013.

[21] C. L. Borgman, G. C. Bowker, T. A. Finholt, and J. C. Wallis, "Towards a Virtual Organization for Data Cyberinfrastructure," in *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York, NY, USA, 2009, pp. 353–356.

[22] C. L. Borgman, J. C. Wallis, and N. D. Enyedy, "Building Digital Libraries for Scientific Data: An Exploratory Study of Data Practices in Habitat Ecology," in *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries*, Alicante, Spain, 2006, vol. LINCS 4172, pp. 170–183.

[23] C. L. Borgman, J. C. Wallis, and M. S. Mayernik, "Who's Got the Data? Interdependencies in Science and Technology Collaborations," *Computer Supported Cooperative Work*, vol. 21, no. 6, pp. 485–523, Dec. 2012.

[24] J. C. Wallis, C. L. Borgman, M. S. Mayernik, A. Pepe, N. Ramanathan, and M. A. Hansen, "Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries," in *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries*, Budapest, Hungary, 2007, vol. LINCS 4675, pp. 380–391.

[25] J. C. Wallis, E. Rolando, and C. L. Borgman, "If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology," *PLoS ONE*, vol. 8, no. 7, p. e67332, Jul. 2013.

[26] D. S. Fearon Jr., C. L. Borgman, S. Traweek, and L. A. Wynholds, "Curators to the Stars (Poster)," in *Proceedings of the American Society for Information Science and Technology*, Pittsburgh, PA, 2010, vol. 47.

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

[28] L. A. Wynholds, J. C. Wallis, C. L. Borgman, A. E. Sands, and S. Traweek, "Data, Data Use, and Scientific Inquiry: Two Case Studies of Data Practices," in *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York, NY, USA, 2012, pp. 19–22.

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

[30] C. L. Borgman, J. C. Wallis, M. S. Mayernik, and A. Pepe, "Drowning in Data: Digital library architecture to support scientific use of embedded sensor networks," in *Joint Conference on Digital Libraries*, Vancouver, British Columbia, Canada, 2007, pp. 269–277.

[31] M. S. Mayernik, "Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators," PhD Dissertation, UCLA, Los Angeles, CA, 2011.

[32] J. C. Wallis, "The Distribution of Data Management Responsibility within Scientific Research Groups," Ph.D., University of California, Los Angeles, United States -- California, 2012.

[33] A. Pepe, A. A. Goodman, A. Muench, M. Crosas, and C. Erdmann, "Sharing, archiving, and citing data in astronomy," *PLoS ONE / Authorea*, forthcoming.

[34] C. P. Ahn, R. Alexandroff, C. A. Prieto, S. F. Anderson, T. Anderton, B. H. Andrews, É. Aubourg, S. Bailey, E. Balbinot, R. Barnes, J. Bautista, T. C. Beers, A. Beifiori, A. A. Berlind, V. Bhardwaj, D. Bizyaev, C. H. Blake, M. R. Blanton, M. Blomqvist, J. J. Bochanski, A. S. Bolton, A. Borde, J. Bovy, W. N. Brandt, J. Brinkmann, P. J. Brown, J. R. Brownstein, K. Bundy, N. G. Busca, W. Carithers, A. R. Carnero, M. A. Carr, D. I. Casetti-Dinescu, Y. Chen, C. Chiappini, J. Comparat, N. Connolly, J. R. Crepp, S. Cristiani, R. A. C. Croft, A. J. Cuesta, L. N. da Costa, J. R. A. Davenport, K. S. Dawson, R. de Putter, N. D. Lee, T. Delubac, S. Dhital, A. Ealet, G. L. Ebelke, E. M. Edmondson, D. J. Eisenstein, S. Escoffier, M. Esposito, M. L. Evans, X. Fan, B. F. Castellá, E. F. Alvar, L. D. Ferreira, N. F. Ak, H. Finley, S. W. Fleming, A. Font-Ribera, P. M. Frinchaboy, D. A. García-Hernández, A. E. G. Pérez, J. Ge, R. Génova-Santos, B. A. Gillespie, L. Girardi, J. I. G. Hernández, E. K. Grebel, J. E. Gunn, H. Guo, D. Haggard, J.-C. Hamilton, D. W. Harris, S. L. Hawley, F. R. Hearty, S. Ho, D. W. Hogg, J. A. Holtzman, K. Honscheid, J. Huehnerhoff, I. I. Ivans, Ž. Ivezić, H. R. Jacobson, L. Jiang, J. Johansson, J. A. Johnson, G. Kauffmann, D. Kirkby, J. A. Kirkpatrick, M. A. Klaene, G. R. Knapp, J.-P. Kneib, J.-M. L. Goff, A. Leauthaud, K.-G. Lee, Y. S. Lee, D. C. Long, C. P. Loomis, S. Lucatello, B. Lundgren, R. H. Lupton, B. Ma, Z. Ma, N. MacDonald, C. E. Mack, S. Mahadevan, M. A. G. Maia, S. R. Majewski, M. Makler, E. Malanushenko, V. Malanushenko, A. Manchado, R. Mandelbaum, M. Manera, C. Maraston, D. Margala, S. L. Martell, C. K. McBride, I. D. McGreer, R. G. McMahon, B. Ménard, S. Meszaros, J. Miralda-Escudé, A. D. Montero-Dorta, F. Montesano, H. L. Morrison, D.

Muna, J. A. Munn, H. Murayama, A. D. Myers, A. F. Neto, D. C. Nguyen, R. C. Nichol, D. L. Nidever, P. Noterdaeme, S. E. Nuza, R. L. C. Ogando, M. D. Olmstead, D. J. Oravetz, R. Owen, N. Padmanabhan, N. Palanque-Delabrouille, K. Pan, J. K. Parejko, P. Parihar, I. Pâris, P. Pattarakijwanich, J. Pepper, W. J. Percival, I. Pérez-Fournon, I. Pérez-Ràfols, P. Petitjean, J. Pforr, M. M. Pieri, M. H. Pinsonneault, G. F. P. de Mello, F. Prada, A. M. Price-Whelan, M. J. Raddick, R. Rebolo, J. Rich, G. T. Richards, A. C. Robin, H. J. Rocha-Pinto, C. M. Rockosi, N. A. Roe, A. J. Ross, N. P. Ross, G. Rossi, J. A. Rubiño-Martin, L. Samushia, J. S. Almeida, A. G. Sánchez, B. Santiago, C. Sayres, D. J. Schlegel, K. J. Schlesinger, S. J. Schmidt, D. P. Schneider, M. Schultheis, A. D. Schwope, C. G. Scóccola, U. Seljak, E. Sheldon, Y. Shen, Y. Shu, J. Simmerer, A. E. Simmons, R. A. Skibba, M. F. Skrutskie, A. Slosar, F. Sobreira, J. S. Sobeck, K. G. Stassun, O. Steele, M. Steinmetz, M. A. Strauss, A. Streblyanska, N. Suzuki, M. E. C. Swanson, T. Tal, A. R. Thakar, D. Thomas, B. A. Thompson, J. L. Tinker, R. Tojeiro, C. A. Tremonti, M. V. Magaña, L. Verde, M. Viel, S. K. Vikas, N. P. Vogt, D. A. Wake, J. Wang, B. A. Weaver, D. H. Weinberg, B. J. Weiner, A. A. West, M. White, J. C. Wilson, J. P. Wisniewski, W. M. Wood-Vasey, B. Yanny, C. Yèche, D. G. York, O. Zamora, G. Zasowski, I. Zehavi, G.-B. Zhao, Z. Zheng, G. Zhu, and J. C. Zinn, "The Ninth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Baryon Oscillation Spectroscopic Survey," *ApJS*, vol. 203, no. 2, p. 21, Dec. 2012.

[35] "Sloan Digital Sky Survey (SDSS): Home," May-2014. [Online]. Available: http://www.sdss.org/. [Accessed: 07-Apr-2013].

[36] A. K. Finkbeiner, *A Grand and Bold Thing: the extraordinary new map of the universe ushering in a new era of discovery*. New York: Free Press, 2010.

[37] J. Gray, D. T. Liu, M. Nieto-Santisteban, A. Szalay, D. J. DeWitt, and G. Heber, "Scientific Data Management in the Coming Decade," *SIGMOD Rec.*, vol. 34, no. 4, pp. 34–41, Dec. 2005.

[38] A. E. Sands, C. L. Borgman, L. A. Wynholds, and S. Traweek, "'We're Working on It:' Transferring the Sloan Digital Sky Survey from Laboratory to Library," presented at the 9th International Digital Curation Conference, San Francisco, CA, 24-Feb-2014.

[39] "Center for Dark Energy Biosphere Investigations," 2014. [Online]. Available: http://www.darkenergybiosphere.org/. [Accessed: 05-Apr-2013].

[40] "International Ocean Discovery Program," 2014. [Online]. Available: http://iodp.org/. [Accessed: 13-Jun-2014].

[41] P. T. Darch and C. L. Borgman, "Ship Space to Database: Scientific and Social Motivations for a Database to Support Deep Subseafloor Biosphere Research," in *Proceedings of the 77th Annual Meeting of the Association for Information Science and Technology*, Seattle, WA, 2014.

[42] "Large Synoptic Survey Telescope: Timeline," *Large Synoptic Survey Telescope*, 2013. [Online]. Available: http://www.lsst.org/lsst/science/timeline. [Accessed: 03-Jun-2014].

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

[44] Z. Ivezic, J. A. Tyson, E. Acosta, R. Allsman, S. F. Anderson, J. Andrew, R. Angel, T. Axelrod, J. D. Barr, A. C. Becker, J. Becla, C. Beldica, R. D. Blandford, J. S. Bloom, K. Borne, W. N. Brandt, M. E. Brown, J. S. Bullock, D. L. Burke, S. Chandrasekharan, S. Chesley, C. F. Claver, A. Connolly, K. H. Cook, A. Cooray, K. R. Covey, C. Cribbs, R. Cutri, G. Daues, F. Delgado, H. Ferguson, E. Gawiser, J. C. Geary, P. Gee, M. Geha, R. R. Gibson, D. K. Gilmore, W. J. Gressler, C. Hogan, M. E. Huffer, S. H. Jacoby, B. Jain, J. G. Jernigan, R. L. Jones, M. Juric, S. M. Kahn, J. S. Kalirai, J. P. Kantor, R. Kessler, D. Kirkby, L. Knox, V. L. Krabbendam, S. Krughoff, S. Kulkarni, R. Lambert, D. Levine, M. Liang, K-T. Lim, R. H. Lupton, P. Marshall, S. Marshall, M. May, M. Miller, D. J. Mills, D. G. Monet, D. R. Neill, M. Nordby, P. O'Connor, J. Oliver, S. S. Olivier, K. Olsen, R. E. Owen, J. R. Peterson, C. E. Petry, F. Pierfederici, S. Pietrowicz, R. Pike, P. A. Pinto, R. Plante, V. Radeka, A. Rasmussen, S. T. Ridgway, W. Rosing, A. Saha, T. L. Schalk, R. H. Schindler, D. P. Schneider, G. Schumacher, J. Sebag, L. G. Seppala, I. Shipsey, N. Silvestri, J. A. Smith, R. C. Smith, M. A. Strauss, C. W. Stubbs, D. Sweeney, A. Szalay, J. J. Thaler, D. Vanden Berk, L. Walkowicz, M. Warner, B. Willman, D. Wittman, S. C. Wolff, W. M. Wood-Vasey, P. Yoachim, H. Zhan, and for the LSST Collaboration, *LSST: from Science Drivers to Reference Design and Anticipated Data Products*. 2011.

[45] LSST Science Collaboration, P. A. Abell, J. Allison, S. F. Anderson, J. R. Andrew, J. R. P. Angel, L. Armus, D. Arnett, S. J. Asztalos, T. S. Axelrod, S. Bailey, D. R. Ballantyne, J. R. Bankert, W. A. Barkhouse, J. D. Barr, L. F. Barrientos, A. J. Barth, J. G. Bartlett, A. C. Becker, J. Becla, T. C. Beers, J. P. Bernstein, R. Biswas, M. R. Blanton, J. S. Bloom, J. J. Bochanski, P. Boeshaar, K. D. Borne, M. Bradac, W. N. Brandt, C. R. Bridge, M. E. Brown, R. J. Brunner, J. S. Bullock, A. J. Burgasser, J. H. Burge, D. L. Burke, P. A. Cargile, S. Chandrasekharan, G. Chartas, S. R. Chesley, Y.-H. Chu, D. Cinabro, M. W. Claire, C. F. Claver, D. Clowe, A. J. Connolly, K. H. Cook, J. Cooke, A. Cooray, K. R. Covey, C. S. Culliton, R. de Jong, W. H. de Vries, V. P. Debattista, F. Delgado, I. P. Dell'Antonio, S. Dhital, R. Di Stefano, M. Dickinson, B. Dilday, S. G. Djorgovski, G. Dobler, C. Donalek, G. Dubois-Felsmann, J. Durech, A. Eliasdottir, M. Eracleous, L. Eyer, E. E. Falco, X. Fan, C. D. Fassnacht, H. C. Ferguson, Y. R. Fernandez, B. D. Fields, D. Finkbeiner, E. E. Figueroa, D. B. Fox, H. Francke, J. S. Frank, J. Frieman, S. Fromenteau, M. Furqan, G. Galaz, A. Gal-Yam, P. Garnavich, E. Gawiser, J. Geary, P. Gee, R. R. Gibson, K. Gilmore, E. A. Grace, R. F. Green, W. J. Gressler, C. J. Grillmair, S. Habib, J. S. Haggerty, M. Hamuy, A. W. Harris, S. L. Hawley, A. F. Heavens, L. Hebb, T. J. Henry, E. Hileman, E. J. Hilton, K. Hoadley, J. B. Holberg, M. J. Holman, S. B. Howell, L. Infante, Z. Ivezic, S. H. Jacoby, B. Jain, R, Jedicke, M. J. Jee, J. G. Jernigan, S. W. Jha, K. V. Johnston, R. L. Jones, M. Juric, M. Kaasalainen, Styliani, Kafka, S. M. Kahn, N. A. Kaib, J. Kalirai, J. Kantor, M. M. Kasliwal, C. R. Keeton, R. Kessler, Z. Knezevic, A. Kowalski, V. L. Krabbendam, K. S. Krughoff, S. Kulkarni, S. Kuhlman, M. Lacy, S. Lepine, M. Liang, A. Lien, P. Lira, K. S. Long, S. Lorenz, J. M. Lotz, R. H. Lupton, J. Lutz, L. M. Macri, A. A. Mahabal, R. Mandelbaum, P. Marshall, M. May, P. M. McGehee, B. T. Meadows, A. Meert, A. Milani, C. J. Miller, M. Miller, D. Mills, D. Minniti, D. Monet, A. S. Mukadam, E. Nakar, D. R. Neill, J. A. Newman, S. Nikolaev, M. Nordby, P.

O'Connor, M. Oguri, J. Oliver, S. S. Olivier, J. K. Olsen, K. Olsen, E. W. Olszewski, H. Oluseyi, N. D. Padilla, A. Parker, J. Pepper, J. R. Peterson, C. Petry, P. A. Pinto, J. L. Pizagno, B. Popescu, A. Prsa, V. Radcka, M. J. Raddick, A. Rasmussen, A. Rau, J. Rho, J. E. Rhoads, G. T. Richards, S. T. Ridgway, B. E. Robertson, R. Roskar, A. Saha, A. Sarajedini, E. Scannapieco, T. Schalk, R. Schindler, S. Schmidt, S. Schmidt, D. P. Schneider, G. Schumacher, R. Scranton, J. Sebag, L. G. Seppala, O. Shemmer, J. D. Simon, M. Sivertz, H. A. Smith, J. A. Smith, N. Smith, A. H. Spitz, A. Stanford, K. G. Stassun, J. Strader, M. A. Strauss, C. W. Stubbs, D. W. Sweeney, A. Szalay, P. Szkody, M. Takada, P. Thorman, D. E. Trilling, V. Trimble, A. Tyson, R. Van Berg, D. V. Berk, J. VanderPlas, L. Verde, B. Vrsnak, L. M. Walkowicz, B. D. Wandelt, S. Wang, Y. Wang, M. Warner, R. H. Wechsler, A. A. West, O. Wiecha, B. F. Williams, B. Willman, D. Wittman, S. C. Wolff, W. M. Wood-Vasey, P. Wozniak, P. Young, A. Zentner, and H. Zhan, "LSST Science Book, Version 2.0," arXiv e-print, Nov. 2009.

[46] M. S. Mayernik, J. C. Wallis, and C. L. Borgman, "Unearthing the infrastructure: Humans and sensors in field-based research," *Computer Supported Cooperative Work*, vol. 22, no. 1, pp. 65–101, 2013.

[47] J. C. Wallis, M. S. Mayernik, C. L. Borgman, and A. Pepe, "Digital libraries for scientific data discovery and reuse," in *Proceedings of the 10th annual joint conference on Digital libraries - JCDL '10*, 2010, p. 333.

[48] P. R. Monge and N. S. Contractor, *Theories of communication networks*. Oxford, U.K.: Oxford University Press, 2003.