Spring May 7, 2016

Open Data in Scientific Settings: From Policy to Practice

Irene V Pasquetto, *University of California, Los Angeles*
Ashley E. Sands, *University of California - Los Angeles*
Peter T Darch, *University of California, Los Angeles*
Christine L Borgman

Open Data in Scientific Settings: From Policy to Practice

Irene V. Pasquetto[1] **Ashley E. Sands**[1] **Peter T. Darch**[2] **Christine L. Borgman**[1]
[1]Department of Information Studies, University of California, Los Angeles (US)
[2]Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign (US)
irenepasquetto@ucla.edu, ashleysa@ucla.edu, ptdarch@illinois.edu, christine.borgman@ucla.edu

ABSTRACT
Open access to data is commonly required by funding agencies, journals, and public policy, despite the lack of agreement on the concept of "open data." We present findings from two longitudinal case studies of major scientific collaborations, the Sloan Digital Sky Survey in astronomy and the Center for Dark Energy Biosphere Investigations in deep subseafloor biosphere studies. These sites offer comparisons in rationales and policy interpretations of open data, which are shaped by their differing scientific objectives. While policy rationales and implementations shape infrastructures for scientific data, these rationales also are shaped by pre-existing infrastructure. Meanings of the term "open data" are contingent on project objectives and on the infrastructures to which they have access.

Author Keywords
Open Data; Science Policy; Computational Infrastructure; Human Infrastructure; Data Practice.

INTRODUCTION
Open data is a prevalent notion in scientific research and policy. Most scientific stakeholders, which include policy makers, funding agencies, publishers and digital librarians, believe open data provides many benefits to science, for example making science more efficient and trustworthy [14]. Citing these benefits, stakeholders undertake initiatives with the aim of making scientific data more open. Some approaches involve the design, building, and implementation of computational infrastructure with the intention of facilitating the international circulation of scientific data. Other initiatives involve policies mandating scientists to make data open; and the *National Science Foundation* (*NSF*) requires funding applications to include *Data Management Plans* [47]. Often, policies and infrastructures operate in conjunction with each other. One such example is that microbiology journals require scientists to deposit genetic sequence data in a publicly accessible database prior to article publication [9].

Despite increasing provisions for computational infrastructures and enforcement of open data policies, open data largely remains an unrealized ambition across most scientific domains [12]. Existing efforts to open scientific data often take definitions of, and rationales for, open data for granted. Our analysis of recent literature and policy reports shows that open data is described in multiple and contradictory ways [53]. A deeper understanding of relationships between rationales, policies, and computational infrastructures for open data is required to clarify whether, how, and when open data can indeed be beneficial for science.

In this paper, we explore the following research questions:

1. What rationales, definitions, and infrastructures are provided in support of scientific open data?
2. What are the relationships between these rationales, definitions, and infrastructures?

We draw on longitudinal, qualitative case studies of two, large scientific collaborations (one in the domain of astronomy, and the other in the domain of the deep subseafloor biosphere, which studies interactions between seafloor microbial communities and the environments they inhabit) to show that not only do rationales and policies help shape infrastructures, but the affordances and constraints of pre-existing infrastructures also profoundly shape rationales and policies.

LITERATURE REVIEW
While open data has received much attention in the HCI community [33,73], the primary emphasis is on studying definitions and barriers to open data in government and industry. However, open data in science has received far

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858543

less attention in this community. The forms and uses of scientific data differ from those of government and industry in at least three ways. First is the difference in goals. The benefits of open government data include increasing the efficiency of the bureaucratic machine, the transparency of government practices, and citizen participation. Making scientific data open promotes reproducibility and fosters the reuse of public-funded assets. Second is the difference in stakeholders. In government open data these are bureaucrats, industry, and the public. In science, stakeholders include policy makers, funding agencies, publishers, libraries, scientists, and the public. Third is the difference in who does the work to make data open. Government agencies are expected to invest the resources necessary to document, format, and release their data for use by the public. In contrast, the work of open data generally falls upon individual scientists, who may be ill equipped to curate data in ways that those data are useful to others, are discoverable, and are sustainable over the long term. To be trustworthy and interpretable, scientific data must be released in specific formats, along with necessary metadata, provenance documentation, and software. The forms of open data release vary widely by scientific domain, thus policies and practices must be adapted to a diverse array of infrastructures and environments.

Given the assumption that open data benefits science and society, many stakeholders support policies and infrastructures to enable openness. However, the perceptions of what open data means varies widely amongst stakeholders [53]. Here we examine further what it means for scientific data to be open, and the reasons why openness benefits science. We then discuss the computational infrastructures necessary to enable scientific openness, and draw attention to the complexities between these infrastructures, policies, and designs.

Definitions of Open Data

Most science stakeholders define open data as "research data collected using public funds" [52], as distinguished from other forms of data such as government statistics or business records [50]. Beyond this general definition, in the scientific community open data is understood in different ways. For example, there is no agreement on the intended audiences for open data. While some policy organizations focus on the idea that data should be open mainly for scientists [24,52,54], other stakeholders include "the public" among the potential recipients of open data, as is the case of the Open Knowledge Foundation [14,42,50,63].

Policymakers' definitions of what "openness" means converge on two factors: legal and technical availability [27,68]. However, policy definitions rarely specify the extent to which open data need to be technically and legally open. Rather, they offer generic expressions such as "fewest restrictions" and "lowest possible costs." [52:15,65:42] As a consequence, differences in conditions around how and when data can be reused are negotiated from time to time depending on the scientific community involved in the policy. Often, a moratorium is established between the data collection period and the day the data are publicly released.

Rationales for Open Data

Borgman [11:208] identifies four rationales for research data sharing: 1) to reproduce research; 2) to make public assets available to the public; 3) to allow others (scientists and non-scientists alike) to use extant data to answer new questions; and 4) to advance research and innovation. These rationales relate to making data open either to researchers (rationales 1 and 4), to the public at large (rationale 2), or to a mixture of both (rationale 3).

These rationales are echoed elsewhere [5,37,39,44,45]. The most frequently reported motivations to make data open are economic and/or quality-related. The economic benefit of open data consists in the idea that scientific data, once collected and cleaned, should be shared and reused by scientists from all over the world. In doing so, the scientific enterprise can avoid investing resources to harvest data that had been already collected and, consequently, allocate funds more efficiently. The quality argument refers to the fact that openly available datasets can be easily verified and used in reproducing scientific studies. In this sense, open data activates a mechanism of quality control, which can also lead to enhanced trust among peers.

Others also stress the benefits of opening access to research data beyond the scientific community [42,50,64]. Some examples include educational tools for K-12 students and the general public [70] shared common resources to promote capacity building in developing countries [63], and the ability for crowd-sourced and citizen science projects to promote scientific public outreach and engagement [13,18].

Computational Infrastructure for Open Data

There are many ways to disseminate data, such as depositing datasets in digital archives or repositories, packaging data as supplemental materials with journal articles, contributing to domain-specific collections, depositing in university library special collections, posting on personal or laboratory websites, and through private exchange between individuals [71].

Examples of computational infrastructure that aim to facilitate data openness include repositories and archives such as *GenBank* [9], federated data networks such as the *Long-Term Ecological Research Network* [69], and international standardization missions such as the *International Virtual Observatory Alliance* [77]. However, availability of these infrastructures varies widely by domain, data type, and country. Another obstacle to data reuse is the fact that many of these infrastructures have only short-term funding [65]. Commercial services for data management, storage and access are appearing, as are data journals in which datasets can be contributed as citable publications (for instance, *Dryad Digital Repository* [75])

Addressing interoperability of infrastructures for data and for scholarly communication motivates further conceptual and technical work. One line of research investigates how to model relationships between datasets, such as strategies for identifying, retrieving and linking datasets. These include *Digital Object Identifiers* [21], *Linked Open Data*, based on WC3 standards [10], Object Reuse and Exchange [8], *Resource Sync* [66], Scholarly Research Objects [7] and *Linked Open Science*, which supports "executable papers" [36]. Computational strategies for opening access to data are evolving rapidly.

Computational Infrastructure, Policy and Design

Often, initiatives for improving the accessibility and circulation of research data rest on an assumption that the definition of open data is unproblematic and that rationales for open data shape policies, which in turn shape the computational infrastructure [37,52,68].

However, studies of scientific infrastructure suggest that the relationship between computational infrastructure and policies is complex [23,35]. Computational infrastructure has been described, "as much the child of science policy as it is of technology per se" [32]. Values and standards are embedded in infrastructure as it is built and configured [31,34]. Conversely, the configurations of infrastructure can shape the values of scientific researchers [30]. Indeed, Jackson et al. [32] regard "policy, practice and design" as interdependent parts of the same complex system. They describe this three-way relationship as similar to a tangled knot: it is not possible to establish clear cause and effect. Thus, more attention should be paid to the relationships between definitions of open data in policies, rationales for open data, and the computational infrastructure that is built to support the accessibility and circulation of research data.

CASE STUDIES

To address our research questions, we present findings from two longitudinal, qualitative case studies of large, distributed, multidisciplinary scientific collaborations that provide important contrasts in type of scientific research, project scale, types of data collected, and data management practices. These communities afford rich opportunities for answering our research questions, enabling us to explore the relationships between open data policies and infrastructures, and how and why scientists engage in building, configuring, and negotiating these infrastructures and policies. Here, we introduce our case studies and methods.

Sloan Digital Sky Server

The *Sloan Digital Sky Survey* (*SDSS*) is a large telescope project built and operated by a consortium of hundreds of astronomers, software engineers, instrument builders, and managers [78]. The first phase of SDSS, *SDSS-I*, was in operation from 2000-2005, the second, *SDSS-II*, from 2005-2008, and subsequent SDSS projects continue today. Our case study focuses on SDSS-I & II, which included 25 member organizations and hundreds of researchers internationally. SDSS received tens of millions of US dollars from multiple sources, including core funding from the *Alfred P. Sloan Foundation*. The astronomy survey, originally intended to provide quantitative data for the study of galaxies, has proven beneficial to nearly every subfield of astronomy.

Center for Dark Energy Biosphere Investigations

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

Researchers are geographically distributed, with the Principal Investigator (PI) and four co-PIs based at five US universities distributed coast to coast. C-DEBI funds short-term research projects conducted by teams across 50 institutions in the USA, Europe, and Asia [16]. C-DEBI scientists generate, analyze and correlate data about rock samples' microbial communities and the physical properties of the samples themselves. Rock samples, also called cores, are typically collected on ocean drilling cruises conducted by the *Integrated Ocean Drilling Program* (*IODP*), which ran from 2003–2013, and its successor, the *International Ocean Discovery Program* (*IODP2*, 2013-present) [29].

METHODS

We employed qualitative research methods including ethnographic observations, semi-structured interviews, and document analysis. Qualitative methods have been widely and successfully employed to study scientific work [41,62], including distributed and multidisciplinary collaborations [26,49]. Conducting case studies of two different domains sharpened our focus on each by enabling comparisons and contrasts [38]. The distributed nature and scale of each case study posed particular challenges, which we addressed with a combination of local and general investigations [55].

Observational work

A key feature of both case studies is long-term ethnographic observation [25]. For our C-DEBI case study, one of the authors was embedded for eight months in a laboratory headed by a leading figure in C-DEBI at a large US research university. This author also conducted weeklong observational work in two other participating laboratories in the US and joined researchers on a three-day field research expedition. Another author conducted observational work of SDSS-I & II collaboration members and data users at seven SDSS Participating Institutions (primarily university Astronomy departments), for a total period of nine weeks.

We recorded extensive notes about what we observed, including the physical layout of offices and laboratories, tools and methods used, patterns of collaboration, as well as

what our informants told us about their backgrounds, aspirations, and experiences in their workplaces.

SDSS-I & II and C-DEBI are distributed across multiple institutions and countries, which poses issues of scalability for the ethnographic researcher [57]. The work of these organizations spans more sites than a small team of researchers can visit, much less to meet face-to-face with all personnel. One way to address this issue was to focus on the techniques and technologies – the "*scalar devices*"– employed by our research subjects to themselves come to understand the collaborations in which they are involved [55:158].

One such device that we observed was the C-DEBI All-Hands Meeting and several other workshops. Another was the American Geophysical Union Fall Meeting 2013 in San Francisco, a major conference for C-DEBI-affiliated scientists, and where an author presented findings from our research. We also attended and presented findings at two American Astronomical Society meetings. These events enabled our research subjects to take stock of the scale of the communities and infrastructures in which they are embedded, in terms of the people involved, organizational hierarchies and policies, and the range of scientific work conducted.

The distributed nature of C-DEBI, IODP, and SDSS-I & II also means that work in these organizations often takes place between non-collocated people through multiple communications media. By using multiple forms of media, we could establish "co-presence" when "co-location" was not possible [6]. Co-presence involves the researcher witnessing how the work of scientific collaborations is conducted even when they are not physically (nor necessarily temporally) collocated with the subjects of research.

For instance, it is not possible to observe practices on board an IODP cruise, given the expense and limited places available. Furthermore, not all work in relation to the IODP is conducted on cruses. We attended online meetings and seminars where participation and data collection were planned. Other online observations included workshops, meetings where key C-DEBI personnel planned how to build and implement centralized infrastructure to coordinate data management across the project, and websites of organizations and people.

Interviews

Our interview sample for this article consists of 49 people from C-DEBI and IODP, and 134 people from SDSS-I & II. Interviews ranged in length from 30 minutes to three hours, with the majority between one and two hours long. With the consent of the interviewees, interviews were recorded and professionally transcribed.

C-DEBI interviewees were initially recruited from those scientists being observed in the laboratory, and were typically interviewed after an extended period of observation. Other C-DEBI interviewees have been recruited from those who had been awarded C-DEBI-funded grants, with these interviews typically taking place over Skype. We have interviewed undergraduate and graduate students, postdoctoral researchers, faculty members, and other senior staff involved in administering and operating C-DEBI. IODP interviewees were identified and approached through a range of methods, including personal introductions from C-DEBI-affiliated scientists and other IODP personnel, and from public websites.

SDSS-I & II interviewees were chosen to cover a broad array of the kinds of expertise necessary to the collaboration. First, interviewees were chosen to reflect both those who built or maintain the project and those who have used SDSS data for their personal research. Often, interviewees can speak to both relationships with the SDSS data. Interviews were conducted at multiple university astronomy departments, national laboratories, data centers, and research institutes, primarily located in the US. Interviewees covered a range of career stages (including graduate students, faculty, staff, and retirees) and types of expertise (including astronomers, computer scientists, engineers, administrators). Interviewees were identified through ethnographic work at the SDSS Participating Institutions, and by identifying authors of journal articles using SDSS data.

Our interviews cover a range of topics, including interviewees' backgrounds and career trajectories. We ask scientists and technical staff detailed questions about the scientific work they are undertaking, and the importance and role of data in their work. Where relevant, we ask stakeholders about their role in formulating and implementing policies and infrastructure within their collaborations.

Document analysis

We have also assembled a corpus of documents for analysis. Documents such as instruction manuals for laboratory equipment and documentation for software, help explain the work conducted by C-DEBI-affiliated scientists and users of SDSS-I & II data in their laboratories and offices. Other documents help us to interpret contexts in which C-DEBI and SDSS-I & II personnel operate, and often function as scalar devices as well, providing details and metrics about activities, plans, and available infrastructural resources. Such documents include both informal and official documents such as funding proposals, and Annual Reports, operating documents, and Memoranda of Understanding (MOUs).

Data analysis

Our initial data analysis involved close reading of our ethnographic notes, interview transcripts, and documents. We identified emerging themes, based on our understandings of the relational, complex, and dynamic nature of knowledge infrastructures, and coded our data accordingly. In particular, we focused on themes relating to:

how those we interviewed described their own work (scientific, organizational, building infrastructure); how they identified and defined what they consider to be data in their own work and, specifically, what the term *open data* means to them; what resources, both currently and anticipated in the future, they identify as necessary to their own work and to realizing their community's aspirations for data openness; what they consider as infrastructure; and how they and their community negotiate, access, and build infrastructure. We refined our coding scheme iteratively, going back and forth between our scheme and the data. Using a range of sources enables us to triangulate, cross-checking our data to validate our findings [48].

For both cases, we began data analysis mid-way through our data collection. We have thus been able to strike a balance between, on the one hand, ensuring our observations have not been biased by preconceived ideas and, on the other, being able to assess our emerging findings and tentative hypotheses against further observations. We have also presented our emerging findings to domain scientists at major scientific meetings for feedback and clarification.

RESULTS

We present results from both case studies, organized by case and presented in thematically parallel sections. First, we begin our results by describing what scientific data are in each setting. Second, we describe the motivations that guided the release of open data. Then, we describe how open data are discussed and conceptualized in the collaborations' documents regulating scientific practices. Finally, we conclude with an overview of the computational infrastructure for data built by the collaborations.

Open Data and SDSS

Here we discuss what the SDSS data are, the motivations for and written policies about the data, and the computational infrastructures that enabled the data to be open.

What are the SDSS data?

SDSS-I & II dataset is a large, complex aggregation of information about astronomical objects, including galaxies and quasars. This dataset comprises images, spectra, and catalogs of the scientific parameters gathered through the image and spectra collection [60,61]. Other complementary information includes data processing software, metadata, and documentation. In total, the SDSS-I & II archive forms a collection of information of between 100 and 200 terabytes.

SDSS data are handled through a software pipeline to prepare the pixels from the detectors for scientific analysis; as the data move through the processing pipeline, different levels of data products are created. For example, the direct data stream from the telescopes and detectors is referred to as primary data or raw data [74,76]. Data that are processed through complicated pipelines are then vetted and verified by the collaboration, and finally made available to the world through "data releases" [28,61].

Once the data have been released, astronomers around the world use the data for their scientific objectives, which may necessitate further refinement and processing. Such data products, derived from work conducted outside of the SDSS collaboration can include catalogs that combine SDSS data with other sources of data. The resulting derived data products have been locally processed by individuals and small groups and tend to be stored on university computer networks or personal computers, with archival and sharing practices local and ad hoc. SDSS project documents did not specify preservation and access for derived and hybrid data products produced by end-user astronomers and therefore do not follow a standardized openness, sharing, or preservation policy [4,20].

Motivations for data openness in SDSS

We identified four primary motivations for opening up SDSS data. First, the collaboration mentioned benefits that we describe as improving the efficiency of the science [61:3]. As with many kinds of science, making SDSS images and spectra available means that the data do not need to be collected again for most kinds of research, until a new wave of telescope or imaging capabilities occurs. Telescope time saved on repetitive observations can be used to increase the importance and usefulness of the scientific information collected.

A second kind of motivation is what we refer to as quality-related [74,76]. For example, open dissemination of the SDSS data is useful to the project as it increases the number of astronomers working with the data and software and thus increases the amount and diversity of helpful feedback provided to the collaboration in terms of ways to improve the dataset. Opening the SDSS data thus helped ensure the amount and quality of feedback the team received.

A third motivation for data openness, which we learned from our interviewees, was that of ensuring continued funding from the NSF. In particular, in order to ensure distribution of the public funds, the SDSS team released the Early Data Release (EDR) [58] as an act of good faith to the NSF.

Finally, the SDSS community identified some benefits of making data open to the public for educational and research purposes [3]. Amateur involvement in astronomy has an extremely rich history and has been critical for many new discoveries of objects [17], much more so than for the majority of other scientific disciplines [67]. A sophisticated infrastructure has emerged over the decades to support and integrate amateur observations into the body of astronomy knowledge [43]. SDSS very much regarded itself as part of this tradition, and also anticipated that members of the public might be able to contribute to astronomy through the use of SDSS data.

Data openness in SDSS policies

SDSS was founded on principles of open data including public distribution and long-term access. Since the earliest periods of development of the sky survey, SDSS leaders agreed to opening the data and ensuring its public availability. In the first Principles of Operation (PoO) in 1989, it was stated that "...a reliable and easily utilized data base… will be made available to the public...."[1:Preamble C]. The SDSS data were thus made available not only to astronomers across the globe, but also to the general public [46].

The emphasis on enabling data access to not only astronomers, but also the general public, only grew over time. The amount and kinds of the SDSS data that should be made available also increased, as evidenced in project documentation. The processed data, often in the form of official data releases, is the level of data to which the openness documentation generally refers. However, by 1997, the collaboration expanded by saying, "The data will be available in its entirety, in both raw and various reduced forms, to the collaboration and, ultimately, to the entire educational, astronomical and public communities" [2:14.1.2]. By 2000, the collaboration was clear that the raw data, processing pipeline, and other distinct levels of processed data products were all important for data release and sharing. Eleven years later, the 2000 PoO explained, "The data should be retained as a full dataset of all pixels on the sky as well as in reduced datasets for later analysis and distribution" [3].

SDSS was also characterized by a strong commitment to *long-term* data access. Early on, SDSS team members thought that, "This public archive is expected to remain the standard reference catalog for the next several decades" [61:3]. The collaboration has turned out to be correct and the SDSS is remains a primary resource for data calibration for other instruments as well as continued scientific investigations.

SDSS computational infrastructure

SDSS policies mandated that the data were made "available through public data release" [3]. The SDSS Principles of Operation committed to public, scientifically accurate, and technically usable, data releases: "Consistent with plans to maintain the integrity and usability of the Science Archive, and as mandated by the funding agencies, the SDSS-II will construct periodic public releases of its contents" [3]. Each data release was announced through a journal article and made available online. The data are accessible in two forms: a flat file format, to enable use by a range of levels of astronomy data expertise, and an organized database, which allows precise search and retrieval.

The SDSS not only released the processed data publicly, but also provided tools to enable scientific use of the data. The data and documentation are available online: "Object catalogs, imaging data, and spectra are all available through the SDSS web site <http://www.sdss.org>, along with detailed documentation and powerful search tools" [40:2]. SkyServer is a SQL database that can be queried by anyone around the world via the website. The SkyServer is a user interface that enables effective search of the database[56,72]. In operation since June 2001, it "supports both professional astronomers and education access" [59]. The SkyServer interface provides different levels of discovery, based on the technical capability of the users. SDSS team members overall tout the success of the SkyServer interface.

Open Data and C-DEBI

The domain of the deep subseafloor biosphere is characterized by a scarcity of data and resources. Although it began in 2010, C-DEBI only developed a plan for data openness in 2012 [15]. In this section, we discuss the motivations behind data openness in C-DEBI, and how C-DEBI is leveraging extant, and building new, computational infrastructure to realize its plans for openness. First, however, we briefly outline what are the relevant data.

What are the data in C-DEBI?

To answer their research questions, C-DEBI scientists use multiple sources of data. Here, we focus on the most common and critical sources. One source of data is the results of analyses of the physical composition of cores. These analyses are conducted on board all IODP cruises, according to standardized procedures. These data are then made available via an online database.

Other sources of data come from analyses of cores from IODP cruises conducted by C-DEBI scientists in their onshore laboratories. Some of these data result from analyses of the physical composition of cores. These analyses are more specialized than IODP analyses and tailored to the particular needs of that scientist's research project.

A second type of laboratory-generated data is data about the composition of the microbiological communities in core samples. Initially, scientists extract DNA from core samples. Following some further processing steps, DNA samples are sent to external sequencing facilities (usually either companies or other university laboratories) that, for a fee, generate DNA sequences. These sequences are then sent back to the scientists, who use computational tools to clean and analyze the sequences.

Motivations for data openness in C-DEBI

For the purposes of this paper, and given constraints of space, we focus on openness in relation to the physical science and microbiological data generated in the scientists' onshore laboratory. Data openness emerged as an official aspiration and policy of C-DEBI in 2012, once it became increasingly apparent that promoting openness was in the interests of C-DEBI as an entity, and of the deep subseafloor biosphere as a whole.

One way in which data openness serves the interests of C-DEBI is that it has played a critical role in the successful

renewal of National Science Foundation (NSF) funding for C-DEBI in 2015. After C-DEBI was launched in 2010, the NSF introduced a requirement for recipients of NSF funding to implement a data management plan (National Science Foundation, 2010). The *C-DEBI Data Management Philosophy and Policy* document (henceforth referred to as the *DMPP*), was developed in response, in time for submission of renewal proposal (Center for Dark Energy Biosphere Investigations, 2012).

Further impetus to encourage data openness has resulted from the experiences of scientists since C-DEBI was launched. During the first 18 months of C-DEBI, three major microbiology-focused IODP expeditions took place, providing the C-DEBI Principal Investigator (PI) and two of C–DEBI's co-PIs with their first experience of leading IODP expeditions. Furthermore, C-DEBI has brought dozens of scientists into the domain of the deep subseafloor biosphere. Combined, these activities have served to make the C-DEBI community aware of the potential benefits of greater data openness to the domain, in a number of ways.

One way is that greater openness is expected to promote more efficient exploitation of scarce resources. The deep subeafloor biosphere is a very new scientific domain, and very little relevant data was collected before the early 2000s. Further, IODP cruises are infrequent, and costly. Thus, data about the subseafloor biosphere is very scarce, and greater openness is associated with more opportunities to reuse data.

A final anticipated benefit of greater openness of data is addressing the challenges of the extensive methodological heterogeneity across the domain, particularly relating to methods of conducting microbiological analyses in onshore laboratories. We have observed many disparate methods and tools used by scientists - even those on adjacent benches in the same laboratory – to accomplish the same task (for more details, see Darch et al. [19]). Some methods may produce biased results, whilst others may be more efficient than others, producing greater volumes of data from the same quantity of core samples. Greater data openness is anticipated to enable meta-analyses by allowing scientists to compare datasets produced different methods in order to identify the most, and least, reliable and efficient methods [51].

Data openness in C-DEBI policies

The official C-DEBI policies relating to data openness are to be found in the C-DEBI Data Management Philosophy and Policy document [15]. Although C-DEBI was launched in October 2010, the DMPP was the first C-DEBI policy document addressing the issue of data openness, as well as being the first policy released by C-DEBI to explicitly address the issue of data management and curation.

DMPP states 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" [15:1]. In particular, they stress that access to data is for other members of the deep subseafloor biosphere community, making no mention of other possible audiences (such as members of the public, or researchers in other scientific domains). However, DMPP also emphasizes that they wish to protect the professional interests of researchers who have spent much time, effort, and funding in collecting their own data. Consequently, the DMPP "strives to strike and equitable balance between open access and protection of intellectual capital" [15:1].

This commitment translates to a number of concrete policy requirements. The policy applies to data produced by C-DEBI-funded researchers during the course of C-DEBI-funded research projects. Researchers are required to make these data, and other information, available "as soon as possible following data collection and analysis" [15:1]. They are allowed a moratorium of up to two years after data collection.

Microbiological and physical science data must be uploaded to relevant openly accessible, publicly funded scientific databases. For instance, genetic data should be archived in databases operated by the *National Institute of Health* (www.ncbi.nlm.nih.gov), while physical science data should be "made available through publication and to all appropriate geochemical databases (e.g., EarthChem - www.earthchem.org, Pangaea - www.pangaea.de, or VentDB - www.ventdb.org)" [15:2].

Consequently, the policies for what data are eligible to be uploaded to these extant databases effectively become policies for data openness in C-DEBI. For instance, for a genetics dataset to be eligible for inclusion in an NIH database, such a dataset must support the conclusions of a scientific article [9]. In other words, genetics data that do not get used for publications (for instance, data that is produced during lines of inquiry that ultimately prove to be dead-ends), do not fall under the purview of the DMPP. C-DEBI-funded researchers will also be required to register data they upload to these databases in an online C-DEBI Data Portal that is currently under development.

In the context of C-DEBI, openness of data is thus subject to a number of limitations: data openness has not applied in since C-DEBI's inception, but only since 2012; the data covered by the DMPP does not include all data produced during the course of C-DEBI-funded research; data does not have to be released immediately upon collection; and the intended audience for C-DEBI data is other domain researchers only.

C-DEBI computational infrastructure

C-DEBI's approach to building infrastructure for data primarily involves using and tying together pre-existing infrastructure comprising a range of publicly accessible scientific databases (such as GenBank and Pangaea, as discussed above) and building some limited computational infrastructure of its own. C-DEBI is leveraging this extant

infrastructure due to limited resources for building its own infrastructure *de novo*.

The infrastructure that C-DEBI is building itself is intended to function as a data registry, with entries for each datasets deposited in the disciplinary databases. The entry for each dataset includes a number of categories, including a link to the dataset in the database, the publication that the dataset supports, and information about which cruises provided the physical samples and data.

DISCUSSION

Both C-DEBI and SDSS collaborations addressed questions about why they should make data produced by their collaboration members open, how to define data openness, and how to leverage extant, and build more, infrastructures to realize their aspirations around data openness. Here, we discuss how rationales for open data, and the definitions of open data in their official policies, differ between the two collaborations. Then, we relate these rationales and definitions to the data infrastructures.

Definitions of Open Data

In their respective policy documents, C-DEBI and SDSS define "open data" differently. Two particularly important components of how the collaborations define open data relate to the intended audience(s) for the data, and what data are included in these definitions.

Audiences of open data are often conceptualized differently between scientific communities: these differences are echoed in our case studies. SDSS intended for its data to be openly available to professional astronomers and members of the public (including amateur astronomers and students) alike, whereas C-DEBI's policies focus on making data openly available to only other deep subseafloor biosphere researchers. This difference echoes a common, and well-established divide between stakeholders, with some focusing on making data open to scientists only [24,52,54], and others also concerned with data accessibility by members of the public [14,42,50,63].

Secondly, we saw that neither SDSS nor C-DEBI conceptualized openness as relating to all data produced by collaboration members; instead, only specific types of data fall under the purview of the projects' open data policies. For example, in the case of SDSS, openness was primarily intended for processed data rather than raw, intermediate, and derived or hybrid data. The coverage of C-DEBI's policy was also restricted, for example limiting to datasets that supported publication in the case of genetics data.

Rationales for Open Data

A variety of rationales motivating open data policies were found in both C-DEBI and SDSS, some relating to opening data to scientists and some to opening data to members of the public, echoing Borgman [11]. Rationales advanced elsewhere by advocates of scientific open data often focus on the benefit to science as a whole, including benefits of more efficient exploitation of extant data, improving the quality of science, and benefits to the public [5,12,37,39,44,45].

While many of these rationales are echoed in our SDSS and C-DEBI case, we found that these rationales were also often closely tied to the specific objectives and interests of the scientific projects or domains themselves. For instance, the quality-related motivations for opening SDSS data served the interests of the project by providing critical feedback to the SDSS team for improving the project's output. Social motivations for opening up SDSS data to the public can be understood in light of the desire to leverage pre-existing infrastructure and traditions of public involvement in scientific discoveries.

At C-DEBI, the rationales of more efficient exploitation of extant data can be seen as a response to scarce resources. Rationales related to the quality of deep subseafloor biosphere science are focused on enabling comparisons of methods, a particular concern in a context of high methodological heterogeneity.

The differences between the C-DEBI and SDSS open data rationales can thus be seen in light of the different challenges and opportunities facing each project. Furthermore, these rationale differences also shape the emphases in each projects' definitions of open data, such as the intended audiences.

Finally, beyond the rationales advanced elsewhere for open data, both SDSS and CDEBI were motivated to open their data by funding concerns, again a rationale related to the projects' own interests. SDSS publicly released their data more quickly than planned in order to prove their commitment to openness to the NSF. Likewise, C-DEBI's development of a plan for data openness was developed in light of the project's impending funding renewal application in 2015.

As our findings suggest, different scientific communities need to make their data open for a variety of purposes. However, open data policies are often standardized and not responsive to the idiosyncratic needs of specific scientific communities.

Infrastructure, Rationales, and Policies: Mutual Shaping

Both SDSS and C-DEBI have built, or are in the process of building, computational infrastructure to realize their specific open data policies. SDSS infrastructure was configured to enable access to data by both professional researchers and members of the public, whereas C-DEBI infrastructure is being designed to tie together deep subseafloor biosphere data deposited in various extant disciplinary databases.

However, our results suggest a more complex relationship between infrastructures, rationales, and policies: while policy definitions for open data do shape scientific infrastructure, extant configurations of available infrastructure also shape open data policies in terms of what

specific types of data are covered by the policies, and how these data are to be made available., to whom, and under what conditions. Scientists do not operate in a vacuum, but in relation to infrastructures and practices. We thus confirm in our case studies that infrastructures are emergent, impact and are impacted by, policy, design, and practice [23,35].

For instance, the inclusion of the public in the intended audiences for SDSS open data can be accounted for in terms of the desire of collaboration members both to leverage, and to continue the tradition of, the sophisticated social and material infrastructure that has integrated amateur astronomers and observations into the body of accepted astronomy knowledge for many decades [18,43]. As C-DEBI relies on external database infrastructures for data deposit, the existing policies of these databases, about what data should be made open to whom, shapes C-DEBI's policies.

CONCLUSIONS

Open data is a term widely used by scientific stakeholders, yet its meaning varies across contexts. This variability inhibits the development of policies and infrastructures that successfully promote the circulation and accessibility of scientific data. New understandings of the relationships between rationales for, definitions of, and infrastructure to support, open data are required.

Our findings demonstrate that rationales and definitions of open data differ between communities. We explored these relationships through the case studies of two major scientific projects, and found them to be very complex, challenging the idea of a linear relationship that sees rationales shaping policies, and then policies shaping infrastructure. Instead, we found these relationships to be much more complex. Certainly, differences in definitions between the two projects are shaped by differences in rationales, and in turn shape differences in the infrastructure developed by both projects. However, rationales and policies are also shaped both by the specific interests of, and extant infrastructure available to, each project.

Our case study of C-DEBI is ongoing, and we are also conducting a case study of the Large Synoptic Survey Telescope, a major data-intensive telescope project currently under development [79]. In the cases of both projects, infrastructure continues to develop, the circulation of data is changing, and project objectives are being modified over time. We will be able to further explore the complexity of relationships between open data rationales, policies, and infrastructure, and the implications of this complexity for the many initiatives that promote open data.

ACKNOWLEDGEMENTS

The work in this paper has been supported by Alfred P. Sloan Foundation Award #20113194, The Transformation of Knowledge, Culture and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective. Thank you to Sharon Traweek, Milena Golshan, and Bernadette Randles for commenting on earlier drafts of this paper.

REFERENCES

1. Astrophysical Research Consortium. 1989. *Principles of Operation of the Sky Survey Project*.
2. Astrophysical Research Consortium. 1997. *A Digital Sky Survey of the Northern Galactic Cap*. Astrophysical Research Consortium. Retrieved from https://catalog.lib.uchicago.edu/vufind/Record/8393880
3. Astrophysical Research Consortium. 2000. *Principles of operation for the Sloan Digital Sky Survey*. Retrieved from http://classic.sdss.org/policies/sdss_poo.html
4. Astrophysical Research Consortium. 2005. *Principles of operation for the Sloan Digital Sky Survey II (PoO-II)*. Retrieved from http://classic.sdss.org/surveyops/docs/PoO-II-10a.pdf
5. Australian National Data Service. 2014. ANDS: Australian National Data Service. Retrieved January 24, 2014 from http://www.ands.org.au/
6. Anne Beaulieu. 2010. Research Note: From co-location to co-presence: Shifts in the use of ethnography for the study of knowledge. *Social Studies of Science* 40, 3: 453–470. Retrieved October 13, 2014 from http://sss.sagepub.com/content/40/3/453.short
7. Sean Bechhofer, Iain Buchan, David De Roure, et al. 2013. Why Linked Data is Not Enough for Scientists. *Future Generation Computer Systems* 29, 2: 599–611. http://doi.org/10.1016/j.future.2011.08.004
8. Sean Bechhofer, David De Roure, Matthew Gamble, Carole Goble, and Iain Buchan. 2010. Research objects: Towards exchange and reuse of digital knowledge. *Nature Precedings*, 713. http://doi.org/10.1038/npre.2010.4626.1
9. Dennis A. Benson, Mark Cavanaugh, Karen Clark, et al. 2013. GenBank. *Nucleic Acids Research* 41, Database issue: D36–D42. http://doi.org/10.1093/nar/gks1195
10. Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked Data - The story so far. *International Journal on Semantic Web and Information Systems* 5, 3: 1–22. http://doi.org/10.4018/jswis.2009081901
11. Christine L. Borgman. 2012. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* 63, 6: 1059–1078. http://doi.org/10.1002/asi.22634
12. Christine L. Borgman. 2015. *Big data, little data, no data: Scholarship in the networked world*. The MIT Press, Cambridge, MA. Retrieved from http://mitpress.mit.edu/big-data
13. Maged N. Kamel Boulos, Bernd Resch, David N. Crowley, et al. 2011. Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application

examples. *International Journal of Health Geographics* 10. http://doi.org/10.1186/1476-072X-10-67
14. Geoffrey Boulton, Michael Rawlins, Patrick Vallance, and Mark Walport. 2011. Science as a public enterprise: The case for open data. *The Lancet* 377, 9778: 1633–1635. http://doi.org/10.1016/S0140-6736(11)60647-8
15. Center for Dark Energy Biosphere Investigations. 2012. *C-DEBI Data Management Philosophy and Policy*. Retrieved from http://www.darkenergybiosphere.org/internal/docs/C-DEBIDataManagementPlan_2012draft.pdf
16. Center for Dark Energy Biosphere Investigations. 2014. *Center for Dark Energy Biosphere Investigations STC Annual Report 2013*. Retrieved from http://www.darkenergybiosphere.org/internal/docs/C-DEBI-Annual-Report-2013.pdf
17. Allan Chapman. 1998. *The Victorian amateur astronomer: independent astronomical research in Britain, 1820-1920*. Wiley published in association with Praxis Publishing, Chichester, Chichester; New York.
18. Carol Christian, Chris Lintott, Arfon Smith, Lucy Fortson, and Steven Bamford. 2012. Citizen Science: Contributions to Astronomy Research. *arXiv:1202.2577*. Retrieved April 6, 2012 from http://arxiv.org/abs/1202.2577
19. Peter T. Darch, Christine L. Borgman, Sharon Traweek, Rebekah L. Cummings, Jillian C. Wallis, and Ashley E. Sands. 2015. What lies beneath?: Knowledge infrastructures in the subseafloor biosphere and beyond. *International Journal on Digital Libraries* 16, 1: 61–77. http://doi.org/10.1007/s00799-015-0137-3
20. Peter T. Darch and Ashley E. Sands. 2015. Beyond big or little science: Understanding data lifecycles in astronomy and the deep subseafloor biosphere. *iConference 2015 Proceedings*, iSchools. Retrieved from https://www.ideals.illinois.edu/handle/2142/73655
21. DataCite. 2014. About DataCite. Retrieved September 22, 2015 from https://www.datacite.org/about-datacite
22. Katrina Edwards. 2009. *Center for Dark Energy Biosphere Investigations (C-DEBI): A Center for Resolving the Extent, Function, Dynamics and Implications of the Subseafloor Biosphere*. Retrieved July 15, 2014 from http://www.darkenergybiosphere.org/internal/docs/2009C-DEBI_FullProposal.pdf
23. Paul N. Edwards, Steven J. Jackson, Melissa K. Chalmers, et al. 2013. *Knowledge infrastructures: Intellectual frameworks and research challenges*. University of Michigan, Ann Arbor, MI. Retrieved July 15, 2013 from http://deepblue.lib.umich.edu/handle/2027.42/97552
24. European Commission High Level Expert Group on Scientific Data. 2010. *Riding the wave: How Europe can gain from the rising tide of scientific data*. European Union. Retrieved from http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204
25. Martyn Hammersley and Paul Atkinson. 2007. *Ethnography: Principles in Practice*. Routledge, London. Retrieved from https://www.routledge.com/products/9780415396059
26. Christine Hine. 2007. Connective Ethnography for the Exploration of e-Science. *Journal of Computer-Mediated Communication* 12, 2: 618–634. http://doi.org/10.1111/j.1083-6101.2007.00341.x
27. Scott Hissam, Barbara Russo, and Fabio Kon. 2011. *Open Source Systems: Grounding Research: 7th IFIP 2.13 International Conference, OSS 2011, Salvador, Brazil, October 6-7, 2011, Proceedings*. Springer Science & Business Media.
28. C. H. Huang, J. Munn, B. Yanny, et al. 1995. *Object-oriented modeling and design for Sloan Digital Sky Survey retained data*. Fermi National Accelerator Lab., Batavia, IL (United States). Retrieved May 22, 2014 from http://www.osti.gov/scitech/biblio/220453
29. IODP. 2014. International Ocean Discovery Program. Retrieved June 13, 2014 from http://iodp.org/
30. Steven J. Jackson and Sarah Barbrow. 2013. Infrastructure and vocation: Field, calling and computation in ecology. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2873–2882. http://doi.org/10.1145/2470654.2481397
31. Steven J. Jackson and Sarah Barbrow. 2015. Standards and/as innovation: Protocols, creativity, and interactive systems development in ecology. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM, 1769–1778. http://doi.org/10.1145/2702123.2702564
32. Steven J. Jackson, Stephanie B. Steinhardt, and Ayse Buyuktur. 2013. Why CSCW needs science policy (and vice versa). *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, ACM, 1113–1124. http://doi.org/10.1145/2441776.2441902
33. Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. 2012. Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management* 29, 4: 258–268. http://doi.org/10.1080/10580530.2012.716740
34. Marina Jirotka, Rob Procter, Mark Hartswood, et al. 2005. Collaboration and trust in healthcare innovation: The eDiaMoND case study. *Computer Supported Cooperative Work (CSCW)* 14, 4: 369–398. http://doi.org/10.1007/s10606-005-9001-0

35. Marina Jirotka, Rob Procter, Tom Rodden, and Geoffrey C. Bowker. 2006. Special Issue: Collaboration in e-Research. *Computer Supported Cooperative Work (CSCW)* 15, 4: 251–255. http://doi.org/10.1007/s10606-006-9028-x
36. Tomi Kauppinen and Giovana Mira de Espindola. 2011. Linked open science-communicating, sharing and evaluating data, methods and results for executable papers. *Procedia Computer Science*, 726–731. http://doi.org/10.1016/j.procs.2011.04.076
37. Stefanie Kethers, Xiaobin Shen, Andrew E. Treloar, and Ross G. Wilkinson. 2010. Discovering Australia's research data. *Proceedings of the 10th annual joint conference on Digital libraries*, ACM, 345–348. http://doi.org/10.1145/1816123.1816175
38. Karin Knorr-Cetina. 1999. *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press, Cambridge Mass.
39. Stacy Kowalczyk and Kalpana Shankar. 2011. Data sharing in the sciences. *Annual Review of Information Science and Technology* 45, 1: 247–294. http://doi.org/10.1002/aris.2011.1440450113
40. Richard G. Kron, James E. Gunn, David H. Weinberg, William N. Boroski, and Michael L. Evans. 2008. *Final Report to the Alfred P. Sloan Foundation*. Retrieved from http://classic.sdss.org/surveyops/annual_reports/SDSSII_final_rep.pdf
41. Bruno Latour and Steve Woolgar. 1979. *Laboratory Life: The Social Construction of Scientific Facts*. Sage Publications, Beverly Hills.
42. Sabina Leonelli. 2013. Why the current insistence on open access to scientific data? Big data, knowledge production, and the political economy of contemporary biology. *Bulletin of Science, Technology & Society* 33, 1-2: 6–11. http://doi.org/10.1177/0270467613496768
43. Philip J. Marshall, Chris J. Lintott, and Leigh N. Fletcher. 2015. Ideas for Citizen Science in Astronomy. *Annual Review of Astronomy and Astrophysics* 53, 1: 247–278. http://doi.org/10.1146/annurev-astro-081913-035959
44. Fiona Murphy. 2014. Data and scholarly publishing: The transforming landscape. *Learned Publishing* 27, 5: 3–7. http://doi.org/10.1087/20140502
45. Peter Murray-Rust. 2008. Open Data in science. *Serials Review* 34, 1: 52–64. http://doi.org/10.1080/00987913.2008.10765152
46. National Science Foundation. 2002. Award#0225645 - SkySurvey: Using SDSS Data in the Classroom. Retrieved September 19, 2015 from http://nsf.gov/awardsearch/showAward?AWD_ID=0225645&HistoricalAwards=false
47. National Science Foundation. 2010. *NSF Data Management Plans*. National Science Foundation, Washington, D.C. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp
48. Tom O'Donoghue and Keith Punch. 2004. *Qualitative Educational Research in Action: Doing and Reflecting*. Routledge.
49. Gary M Olson, Ann S. Zimmerman, and Nathan Bos (eds.). 2008. *Scientific Collaboration on the Internet*. MIT Press, Cambridge, Mass. Retrieved from http://ieeexplore.ieee.org/xpl/bkabstractplus.jsp?bkn=6267421
50. Open Knowledge Foundation. 2014. Open Knowledge: What is Open? Retrieved October 19, 2014 from https://okfn.org/opendata/
51. Beth N. Orcutt, Douglas E. LaRowe, Jennifer F. Biddle, et al. 2013. Microbial Activity in the Marine Deep Biosphere: Progress and Prospects. *Frontiers in Microbiology* 4: 189. http://doi.org/10.3389/fmicb.2013.00189
52. Organisation for Economic Co-operation and Development. 2007. *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Organisation for Economic Co-Operation and Development, Paris. Retrieved from http://www.oecd.org/dataoecd/9/61/38500813.pdf
53. Irene V. Pasquetto, Ashley E. Sands, and Christine L. Borgman. 2015. Exploring Openness in Data and Science: What is "Open," to Whom, When, and Why? *Proceedings of the 78th ASIS&T Annual Meeting*. Retrieved from https://www.asist.org/files/meetings/am15/proceedings/submissions/posters/296poster.pdf
54. RDA Europe. 2014. The Data Harvest: How sharing research data can yield knowledge, jobs and growth. Retrieved December 4, 2014 from http://www.europe.rd-alliance.org/documents/publications-reports/data-harvest-how-sharing-research-data-can-yield-knowledge-jobs-and
55. David Ribes. 2014. Ethnography of scaling, or, how to a fit a national research infrastructure in the room. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, ACM, 158–170. Retrieved October 10, 2014 from http://dl.acm.org/citation.cfm?id=2531624
56. Sloan Digital Sky Survey. 2014. SDSS SkyServer DR12. *Sloan Digital Sky Survey*. Retrieved July 7, 2015 from http://skyserver.sdss.org/dr12/en/credits/creditshome.aspx
57. Susan Leigh Star. 1999. The ethnography of infrastructure. *American Behavioral Scientist* 43, 3: 377–391. Retrieved January 11, 2014 from http://abs.sagepub.com/content/43/3/377.short
58. Chris Stoughton, Robert H. Lupton, Mariangela Bernardi, et al. 2002. Sloan Digital Sky Survey: Early data release. *The Astronomical Journal* 123, 1: 485–548. http://doi.org/10.1086/324741

59. Alexander S. Szalay, Jim Gray, Aniruddha R. Thakar, et al. 2002. *The SDSS SkyServer: Public access to the Sloan Digital Sky Server data*. Microsoft Research. Retrieved December 3, 2009 from http://arxiv.org/abs/cs/0202013
60. Alexander S. Szalay, Peter Kunszt, Aniruddha R. Thakar, and Jim Gray. 1999. *Designing and mining multi-terabyte astronomy archives: The Sloan Digital Sky Survey (Original)*. Retrieved December 3, 2009 from http://adsabs.harvard.edu/abs/1999cs.....7009S
61. Alexander S. Szalay, Peter Z. Kunszt, Aniruddha R. Thakar, Jim Gray, and Don Slutz. 2000. The Sloan Digital Sky Survey and its archive. *Astronomical Data Analysis Software and Systems IX*, Astronomical Society of the Pacific, 405. Retrieved November 14, 2009 from http://www.adass.org/adass/proceedings/adass99/O1-02/
62. Sharon Traweek. 1988. *Beamtimes and Lifetimes: The World of High Energy Physicists*. Harvard University Press, Cambridge, Mass.
63. Paul F. Uhlir. 2007. Open data for global science: A review of recent developments in national and international scientific data policies and related proposals. *Data Science Journal* 6: OD1–OD3. http://doi.org/10.2481/dsj.6.OD1
64. Paul F. Uhlir. 2012. *The Future of Scientific Knowledge Discovery in Open Networked Environments: Summary of a Workshop*. The National Academies Press, Washington, D.C.
65. Paul F. Uhlir and Peter Schröder. 2007. Open Data for Global Science. *Data Science Journal* 6: OD36–OD53. http://doi.org/10.2481/dsj.6.OD36
66. Herbert Van de Sompel, Robert Sanderson, Martin Klein, et al. 2012. A perspective on resource synchronization. *D-Lib Magazine* 18, 9/10. http://doi.org/10.1045/september2012-vandesompel
67. Jeremy Vetter. 2011. Introduction: Lay Participation in the History of Scientific Observation. *Science in Context* 24, 02: 127–141. http://doi.org/10.1017/S0269889711000032
68. Robert Viseur and Nicolas Devos. 2015. How openness can change scientific practice. *ERCIM News 100*, 37. Retrieved April 17, 2015 from http://ercim-news.ercim.eu/en100/special/how-openness-can-change-scientific-practice
69. Robert B. Waide and McOwiti O. Thomas. 2012. Long-Term Ecological Research Network. In *Encyclopedia of Sustainability Science and Technology*, Robert A. Meyers (ed.). Springer New York, New York, NY, 6216–6240. Retrieved December 27, 2015 from DOI:10.1007/978-1-4419-0851-3_749
70. Jillian C. Wallis, S. Milojevic, Christine L. Borgman, and William A. Sandoval. 2006. The Special Case of Scientific Data Sharing with Education. *Annual Meeting of the American Society for Information Science & Technology*, Information Today, Inc., 1–13. http://doi.org/10.1002/meet.14504301169
71. Jillian C. Wallis, Elizabeth Rolando, and Christine L. Borgman. 2013. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE* 8, 7: e67332. http://doi.org/10.1371/journal.pone.0067332
72. Jian Zhang and Chaomei Chen. 2010. Collaboration in an open data eScience: A case study of Sloan Digital Sky Survey. *iConference 2010*. Retrieved November 9, 2015 from https://www.ideals.illinois.edu/handle/2142/14948
73. Anneke Zuiderwijk, Marijn Janssen, Sunil Choenni, Ronald Meijer, and R. Sheikh Alibaks. 2012. Socio-technical impediments of open data. *Electronic Journal of e-Government* 10, 2: 156–172. Retrieved December 27, 2015 from http://www.ejeg.com/issue/download.html?idArticle=255
74. 2003. SDSS scientific and technical publication policy. Retrieved February 2, 2010 from http://classic.sdss.org/policies/pub_policy.html
75. 2010. Dryad. Retrieved from http://datadryad.org/
76. 2014. Publication policy | SDSS. Retrieved July 25, 2015 from http://www.sdss.org/collaboration/publication-policy/
77. 2015. International Virtual Observatory Alliance. Retrieved from http://www.ivoa.net/
78. 2015. Sloan Digital Sky Survey | Home. Retrieved November 8, 2015 from http://www.sdss.org/
79. 2015. Large Synoptic Survey Telescope: Home. Retrieved November 8, 2015 from http://www.lsst.org/lsst