Collaborative Qualitative Research at Scale: Reflections on 20 years of Acquiring Global Data and Making Data Global

Christine L. Borgman, Christine.Borgman@UCLA.edu, Corresponding Author
Morgan F. Wofford, mfwofford@ucla.edu
Milena S. Golshan, milenagolshan@ucla.edu
Peter T. Darch, ptdarch@illinois.edu
Center for Knowledge Infrastructures, UCLA

Authors' Final Accepted Version. In press. Accepted November 29, 2020. To appear as:

Borgman, C. L., Wofford, M. F., Golshan, M. S., & Darch, P. T. (2021). Collaborative Qualitative Research at Scale: Reflections on 20 years of Acquiring Global Data and Making Data Global. *Journal of the Association for Information Science & Technology*. https://doi.org/10.1002/ASI.24439

Abstract

A 5-year project to study scientific data uses in geography, starting in 1999, evolved into 20 years of research on data practices in sensor networks, environmental sciences, biology, seismology, undersea science, biomedicine, astronomy, and other fields. By emulating the 'team science' approaches of the scientists studied, the UCLA Center for Knowledge Infrastructures accumulated a comprehensive collection of qualitative data about how scientists generate, manage, use, and reuse data across domains. Building upon Paul N. Edwards's model of 'making global data' – collecting signals via consistent methods, technologies, and policies – to 'make data global' – comparing and integrating those data, the research team has managed and exploited these data as a collaborative resource. This article reflects on the social, technical, organizational, economic, and policy challenges the team has encountered in creating new knowledge from data old and new. We reflect on continuity over generations of students and staff, transitions between grants, transfer of legacy data between software tools, research methods, and the role of professional data managers in the social sciences.

Collaborative Qualitative Research

Collaborative research using quantitative methods is common throughout the physical and life sciences, engineering and computer sciences, medical and health sciences, and social sciences such as economics, psychology, and sociology. To work together successfully, collaborators need to agree on common methods to acquire, clean, calibrate, reduce, and document their data. Even when done well, challenges arise in integrating, comparing, and exploiting data that originate in different research groups. The larger the collaboration in terms of number of participants, research sites, volume of data, technologies, and variety of data, the greater the challenges become. Similarly, the longer the duration of a collaboration, the more resources that need to be devoted to curation and stewardship of those data (Borgman, 2015; Borgman et al., 2012; Bos et al., 2008; Cummings et al., 2008; Cummings & Kiesler, 2004; Gorman, 2010; Jirotka et al., 2013; Olson et al., 2008; Ribes & Finholt, 2009).

Collaborative research using qualitative methods data is less common, smaller in scale, and typically shorter in duration than scientific projects that deploy quantitative methods. Many reasons exist for these differences, such as less research funding available to qualitative researchers in the social sciences and humanities, fewer rewards for managing large projects, and the idiographic nature of problems that are suited to qualitative methods. Ethnographies, open-ended interviews, document analyses, focus groups, and other forms of qualitative methods support inductive "deep dives" into a problem. Grounded theory approaches to data analysis enable researchers to develop and test hypotheses iteratively. Qualitative methods can be fruitful approaches to case studies and to exploratory research questions. They work well in projects where one or a few researchers collect and compare data. However, these methods rarely scale to the dozens, hundreds, or thousands of sites of data collection possible with quantitative methods (Beaulieu, 2010; Bowker et al., 2009; Ribes, 2014).

Despite these different circumstances, the challenges of data management and coordination are similar with qualitative and quantitative methods. Sharing data between collaborators is easier in fields that have established standards, common data formats, community repositories, and tools to integrate data from multiple sources. Funding agencies, both public and private, require grantees to provide access to data produced with those funds. Journals may require authors to provide access to data associated with their articles, whether or not the research required extramural funding. Managing data in ways to make them openly available can be a complex and expensive process.

Research data are scientific assets that can be mined, combined, and bartered, but they also are liabilities. Maintaining and servicing data are continuing challenges for scientists and social scientists alike. The payoff for investing in data management is the ability to integrate data across projects to address larger research questions.

UCLA Center for Knowledge Infrastructures

Our team, under one PI (Borgman), has conducted qualitative studies of scientific practices since the late 1990s, accumulating a rich trove of interviews, ethnographic notes, documents, and publications. As our focus evolves from active data collection to consolidating our findings, we write this article to reflect on our research methods, in theory and in practice, to offer lessons learned and guidance for others who may embark on similar journeys. We write in the first person, using "the royal we," to represent the many members of the research team who have conducted this body of research over a 20-year period. The authors of this article are current or very recent members of the UCLA Center for Knowledge Infrastructures. Earlier members of the Center and its predecessor teams, and our collaborators at UCLA and other universities, are represented by references to publications and projects in which they participated and in the acknowledgements.

What is now apparent as a 20-year project on scientific data practices began as a five-year (1999-2004) effort to study the design and use of digital libraries in physical geography, conducted in collaboration with geographers and computer scientists. That project, known as the Alexandria Digital Earth Prototype (ADEPT), was not designed to frame a longitudinal study that would span many scientific domains, field sites, and research questions. As our data and our findings accumulated, their collective value become apparent. While our findings on uses of ADEPT in physical geography are reported in numerous publications (Borgman et al., 2000, 2005, 2004; Mayer et al., 2002), the ethnographic notes, documentation, and interviews on which those papers are based languish in boxes of paper, printouts, and legacy formats such as cassette tapes. The materials enshrine the body of work, but cannot be readily repurposed or integrated with data collected in subsequent studies.

Mid-way into the ADEPT project, we became founding members of the Center for Embedded Networked Sensing (CENS), a National Science Foundation Science and Technology Center from 2002 to 2012 (Borgman et al., 2012, 2015; Borgman, Golshan, et al., 2016; Mayernik et al., 2013; Wallis et al., 2013a). Because our information-studies-based team was studying how CENS scientific teams collected and managed their data, we became more deliberate in managing our own data. We have digital records of our CENS data, along with codebooks for documenting them.

As we expanded from geography, environmental sciences, biology, and seismology into astronomy and astrophysics, undersea science, and the biomedical sciences in later years, with other grants, more collaborators, and more staff, our methods became more systematic – and more problematic (Borgman, 2019; Borgman, Darch, et al., 2016; Darch & Borgman, 2016; Pasquetto et al., 2017).

Collaboration and Continuity

Conducting each of these projects individually, and starting anew with data collection each time, would have been far simpler than combining them into a long-term research program that requires continuous data management. We experienced many of the data-handling problems encountered by our research participants, such as tradeoffs between resources spent on data management vs. new data collection, protecting data of dissertation-stage students vs. sharing data with our faculty collaborators, discontinuities in research questions between projects and sites, maintaining continuity in our research program while proposing innovative new directions to obtain new funding, recoding data to make comparisons, migrating data to new platforms, dealing with software and hardware upgrades, handoffs between personnel, and so on.

We discuss these challenges and tradeoffs, comparing our experiences with those of the scientists we study. As Paul N. Edwards (2010) learned in the climate science community, “making global data” is a prerequisite to “making data global.” Global data are those collected in consistent forms, usually based on agreements of methods and measures, so that they can be combined or compared. Investments in making global data span the entire research life cycle, from research design to data reuse. To make data global, which is the process of comparing, combining, and integrating data for scientific purposes, requires data science expertise. Whereas the importance of such data science expertise is now being recognized in scientific domains, the prerequisite skills in curation and stewardship necessary to make global data rarely are part of graduate training in the sciences or social sciences.

After nearly 20 years of investing in global data, we are achieving the rewards of making data global in our own research, while facing similar challenges in data integration and stewardship as those of our scientific research participants. This article reflects on the process of maintaining a long-term research program based on qualitative methods, the challenges and pitfalls, and the rewards and limitations encountered along the way.

Open Science, Data Reuse, and Knowledge Infrastructures

Open science policy, which includes open access to publications, data management plans, and data release with publications, is based on arguments for the value of replication, reproducibility, transparency, and reuse of research data for education and innovation (Borgman, 2015; Organisation for Economic Co-operation and Development, 2007). However, releasing scientific data often creates large burdens on researchers due to the labor and expertise involved in managing, curating, documenting, and providing access to those data (Mayernik, 2016a; Mayernik et al., 2013; Wallis et al., 2013b). Many scientists view data sharing and release as unfunded mandates. Thus, the larger questions that drive our research agenda are to identify where the value lies in data acquisition and reuse, how costs and benefits are distributed among the many stakeholders in those data resources, and the practices by which scientists steward their data.

Disciplines and Data

Some disciplines invest heavily in centralized maintenance of data resources, such as astronomy, genomics, seismology, and certain areas of the environmental sciences. Other disciplines are characterized by local data management, sometimes keeping samples and digital data indefinitely and sometimes discarding them after associated publications are released. Our most consistent finding about data practices across disciplines is heterogeneity. Individuals keep some kinds of data and discard others. Disciplinary repositories acquire some kinds of data and reject others. Scale is also a factor. Larger teams, especially those that generate larger volumes or varieties of data, are better able to invest in data management. Identifying these patterns, and theorizing relationships among them, is central to our agenda.

Another finding of our research is that data practices are embedded in complex social and technological contexts. The theoretical lens through which we view scientific data practices is knowledge infrastructures, a term first coined by Edwards (2010, p. 17) as “robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds.” Data practices can be studied at spatial, disciplinary, and temporal scales (Edwards et al., 2013; Star & Ruhleder, 1996). Infrastructure concerns framed the launch of our research program, starting with international and comparative questions about the roles of digital technologies in information practice (Borgman, 2000). At the end of a large, multi-university grant project, we held a community workshop to develop the concept of knowledge infrastructures (Edwards et al., 2013). At the end of a 5-year grant to UCLA, we held another workshop to examine developments in research on knowledge infrastructures in the intermediate 8-year period (Borgman et al., 2020). The concept is robust and is diversifying across disciplines.

Interdependencies of institutions also arise, as do relationships between software, code, data, and tools. Infrastructures that may appear durable often are fragile upon closer inspection

(Borgman, Darch, et al., 2016). The very idea of "data" is problematized throughout our research (Borgman, 2019; Leonelli, 2019). Whereas science policies tend to imply that data are simply "facts," or otherwise static and bounded objects, they are more commonly malleable, mobile, and mutable (Edwards et al., 2011; Latour, 1987). The ability to generate, use, and reuse data in these collaborative and interdisciplinary environments often requires "interactional expertise" in addition to domain knowledge and technical skills (Collins & Evans, 2007; Pasquetto et al., 2019b).

Research Agenda

Our research agenda lies in one part of Pasteur's Quadrant, that of "use-inspired basic research" (Stokes, 1997). As members of a professional school, we are acutely aware of the benefits of engaging with the communities we serve (*ISchools*, 2020). We partner with the research groups we study, reporting back periodically on our findings, and offering guidance on their data practices upon request. We also publish and give talks in these scientific communities. Our studies of scientific data practices began as a subcontract to ADEPT, a five-year (1999-2004) digital libraries research project on the use of a digital collection of physical geography content for teaching undergraduate courses. Collaborators on ADEPT, which was funded by the U.S. National Science Foundation (NSF), spanned geography, earth sciences, computer science, education, and psychology. Our role was to address questions of what data were useful to physical geographers in their teaching and research and the degree to which the digital library would address those needs. Generally, we found that these geography faculty much preferred to draw data for teaching examples from their own research, rather than seeking external data resources. They were more interested in the digital library to manage their own research data than for its intended instructional purposes (Borgman, 2006; Mayer et al., 2002).

Our research design from the ADEPT project laid the foundation for studying data practices in the Center for Embedded Networked Sensing (CENS), an NSF Science and Technology Center from 2002 to 2012. CENS, with five participating universities and 300 collaborators at its peak, spanned computer science, engineering, biology, environmental science, seismology, medicine and health, and other areas. Findings from ten years with CENS include four doctoral dissertations (Mayernik, 2011; Pepe, 2010; Shilton, 2011; Wallis, 2012), two masters theses, and approximately 100 publications. Overall, we identified a complex array of practices for data management, sharing, and reuse; mixes of incentives, disincentives, costs, and benefits of investing in data that varied by domain and team; vastly different concepts of "data" within and between collaborating research teams; and mapped social and authorial networks of CENS members and their external collaborators.

Starting in 2008, overlapping with CENS, we began to study data practices in astronomy as part of another large NSF center with collaborators from multiple physical and biological sciences, computer science, social sciences, and education. During this time, we started to use the term knowledge infrastructures, rather than associated concepts, such as information infrastructure (Borgman, 2000) or cyberinfrastructure (Atkins et al., 2003). One of our first comparative papers using the knowledge infrastructure framework emphasized the need to incorporate digital libraries into infrastructures to promote data reuse and fulfill the promise of data-intensive scholarship (Borgman et al., 2014). In one of our more recent papers, we examine how data reuse is a socio-technical process embedded within knowledge infrastructures and theorize the data creators' advantage, "that those who create data have intimate and tacit knowledge that can be used as barter to form collaborations for mutual advantage" (Pasquetto et al., 2019b).

We partnered with the Sloan Digital Sky Survey (SDSS) as they sought ways to curate, manage, and maintain access to a massive data resource as they neared the end of their funding for data collection. Astronomy differs greatly in scientific practices, infrastructure, scale, and other features from those of our partners in ADEPT and CENS. We leveraged our SDSS research to partner with other sites in astronomy and astrophysics. Overall, we find that astronomy has the most integrated knowledge infrastructure of any domain we have studied, spanning observational data, catalogs, bibliographic records, archives, thesauri, software, and other resources (Borgman, Darch, et al., 2016). Yet, they too struggle with many aspects of data collection, processing, management, and reuse of data (Boscoe, 2019; Darch et al., 2020b, 2020c; Sands, 2017).

Concurrent with our astronomy research, we studied two other large distributed collaborations at the invitation of their investigators. The first was in undersea science, where ocean drilling ships acquired rock samples (or, cores) for physical and biological research. This body of work builds upon our ecological studies of CENS, given the scientific commonalities, and on the astronomy research, given the large-scale infrastructure required (Darch, 2016, 2018; Darch et al., 2015; Darch & Borgman, 2016). The second collaboration is biomedicine, where a distributed and multidisciplinary array of labs, in a hub and spokes model, shares data about craniofacial abnormalities. While the biomedical collaboration is the farthest afield scientifically from our other sites, it has yielded striking comparisons in areas of data generation, reuse, and information policy (Pasquetto, 2018; Pasquetto et al., 2017, 2019a).

In sum, we are studying multiple knowledge infrastructures, each of which has many components, and relationships among those infrastructures. In domains such as astronomy, the community has funding and critical mass to maintain sophisticated infrastructures that span decades and countries. In domains such as undersea science, where data are sparse and disciplines are emergent, the community relies upon multiple infrastructures that are maintained by other stakeholders. Domains such as environmental sciences and biomedicine fall somewhere in between, each able to build some portions of their own infrastructures and to rely on multiple infrastructures that are controlled by other stakeholders

Investing in Data Assets

Acquiring and managing data in ways that they can be kept 'alive' for future reuse is a far different process than collecting data for a single grant project or a single dissertation in which data can be abandoned shortly after the publication of results. Commitments to data preservation pervade the process, from team building, research design, data collection, data management, and publication, to stewardship. By investing in our own data management, we gained opportunities to reflect on the data handling challenges of our research participants, and to construct more nuanced interpretations by comparing new and old findings continuously. The overhead is considerable, but necessary to study multiple knowledge infrastructures across many domains over long periods of time.

Following Edwards (2010), we distinguish between the process of making global data and making data global. In his framing example, making global data is the process of developing technical, social, governmental, and policy agreements by which weather services around the world could collect data in consistent forms that could be shared. Standards were lubricants in this century-long process, but friction remains a constant (Edwards, 2010; Edwards et al., 2011). Making data global is the process of integrating those data into computer models that could be used to model, predict, and theorize weather and meteorology. We have made similar investments in making global data, albeit on a significantly smaller scale. Subsequent generations of CKI researchers are now able to make these data global through comparisons over time and across projects.

Acquiring Global Data

In our case, the process of acquiring global data on scientific data practices can be grouped into several stages. To design research programs that produce reusable data, a first step is to take a team science approach and a second step is to build effective teams. Thereafter it becomes possible to pursue data reuse and integration across projects.

Team Science

The research groups we study reflect typical models of team science, with most teams consisting of two to ten individuals. Each of these teams may collaborate with teams of similar size elsewhere, domestic and international, resulting in coordinated efforts of dozens or hundreds of individuals. Team science, a much-studied research topic, offers benefits by assembling complementary expertise to address complex problems, balanced with the costs of communication and coordination (Bos et al., 2007; Gorman, 2010; Olson et al., 2008; Wagner, 2018). Among the features of team science that create challenges for research, as identified in a recent National Academies of Science (NAS) study, are high diversity of membership in terms of age, gender, culture, religion, or ethnicity; deep knowledge integration; and high task interdependence. Teams in all of our studies exhibited various combinations of these characteristics. The largest and most distributed teams we studied also exhibited the features the NAS study identified for larger teams, such as goal misalignment, permeable boundaries, and geographic dispersion (Committee on the Science of Team Science et al., 2015).

Our research participants faced similar challenges in acquiring data that would remain useful for their teams, and in distributing authorship credit (Borgman et al., 2012; Scroggins et al., 2020; Scroggins & Pasquetto, 2020; Wallis, 2012). Team science was the norm in the scientific groups we studied, which guided our collaborative approaches.

Team Building

The UCLA Center for Knowledge Infrastructures (CKI), as our team is now known, grew out of a CENS science team of three: a professor, a post-doctoral fellow, and a graduate student. Following the model of CENS teams, this group became known as ‘the Borgman lab,’ and later the ‘data practices team,’ in partnership with the CENS statistics group. Over the 20 years of research work discussed in this article, more than 20 people were part of the CKI team, such as PhD students, postdocs, professors, graduate student researchers, staff, and a few volunteers. We sometimes had joint grants with faculty at other universities, creating a much larger science team.

Thus, members of the CKI are a social science team that functions as a science team, with shared goals and infrastructure, collaborative writing practices, standing meetings every week, and joint responsibility for the data and other knowledge products we produce. Every grant proposal, paper, and talk is developed as a team and workshopped iteratively. Our practices reflect the benefits of integrating diverse knowledge and the overhead of coordination and building infrastructure.

Data Reuse Practices

Our team science approach is central to acquiring global data. We address these challenges by setting data reuse as a common goal for the team. As each new student or other collaborator joins the team, we establish expectations about data sharing within the team, while maintaining the confidentiality of our research participants. These are largely informal agreements, encoded in meeting notes but not subject to formal contracts. Our goal is to make our data reusable for asking new questions and combining old data with new data, not for reproducing the initial study. The distinction between reuse and reproducibility is a significant topic in our research agenda, and one we have addressed in depth elsewhere (Borgman, 2015; Boscoe, 2019; Pasquetto, 2018, 2019; Pasquetto et al., 2017, 2019a; Wofford et al., 2020).

When students reach the stage of developing dissertation proposals and conducting their data collection and writing, we give them a proprietary period for sole use of their data until the dissertation is filed. Thereafter, those data, which were acquired under grants to the university, become part of the CKI pooled resources for comparative research. In most cases, those who collected the data participate in writing joint publications which result, receiving authorship credit accordingly. Most of our publications are joint-authored, whether comparing data from multiple sites or addressing themes such as data reuse.

In one large collaboration involving faculty collaborators from multiple universities, each of whom were employing students and post-doctoral fellows on the project, we agreed that individuals who conducted interviews would always receive acknowledgements, but not necessarily receive co-authorship credit. The collaboration was too large and the number of papers too many to bring everyone ever involved into the writing process. That model has worked well. Our publications always acknowledge funding sources and include mention of anyone whose interviews were used, which might sometimes be a single quotation, but who did not participate in writing that paper.

Among the challenges of reusing qualitative data is that documentation tends to be highly personal, as each researcher develops relationships with participants over long periods of time. Individual ethnographers are often highly proprietary about their methods, notes, recordings, transcripts, and other field data collection. Sharing those data with others who have not participated in the research requires considerable explanation of context (Pasquetto et al., 2019a).

Documenting fieldnotes to make them reusable posed special challenges. We encouraged CKI members in the field to write extensive notes after each session of observation, following common ethnographic practice. Notes should include comments about what they had observed about scientific activities, interactions among research subjects, and personal interactions with research participants. CKI researchers often had privileged access to highly sensitive material such as hitherto unpublished research findings and participants' opinions of their collaborators. We are always careful to maintain confidentiality when using these materials and in sharing them between CKI members. In some cases, one member used data collected by other members, and in other cases two or more CKI members studied a single infrastructure such as CENS or SDSS (Darch et al., 2020a, 2020c; Scroggins & Pasquetto, 2020). Fieldnotes provide information about research practices that are not reported in interviews, and become contextual information useful to interpret interview transcripts later.

Sharing fieldnotes can be a fraught practice. Fieldnotes tend to be personalized, including frank impressions of research participants, reflecting relationships that the researcher has developed over long periods of time. The imperative to write fieldnotes as soon after observation as possible, perhaps after a long and tiring day in the laboratory, means that fieldnotes are not polished. To make global data, fieldnote authors must overcome a natural reluctance to share unpolished work with close professional colleagues.

When sharing fieldnotes, we sought to strike a balance between conserving the immediacy and detail of notes written during observation periods and allowing notes' authors to moderate frank impressions of research participants and to polish writing before sharing. Physical co-location in shared offices and trust built between CKI team members played a critical role in making fieldnote authors comfortable with sharing fieldnotes and with the exchange of contextual information necessary to interpret these notes. The contextual information necessary to interpret others' fieldnotes becomes less accessible over time, as students graduate and as memory fades (Pasquetto et al., 2019a).

Research Designs for Collecting Global Data

Collecting global data requires agreements on research designs that balance the need for continuity across grant projects and individual work such as dissertations. By anchoring our research designs in common protocols and human participants consent forms, we found that we could vary other aspects of investigations while creating a pool of data resources that could be reused by the CKI team in the future.

Common Protocols

Our protocols for ethnographic observation, interviews, and document analyses began organically, designed for our early grant projects. These were largely constructed, tested, and implemented by two students whose dissertations addressed data practices in CENS (Mayernik, 2011; Wallis, 2012). Core questions about data collection, analysis, sharing, reuse, and management provided continuity in other CENS studies. With that anchor, we could pursue new avenues and nuanced aspects of prior findings (Borgman et al., 2012, 2014, 2006, 2007; Mayernik et al., 2013; Wallis et al., 2013a, 2007).

Core protocols that were developed for CENS proved reasonably robust for application to astronomy, undersea science, and biomedicine. We developed complementary questions to explore the specifics of these domains. Each grant proposal and dissertation pursued new research questions, with our overarching questions remaining at the core of our inquiries. Similarly, our scientific research participants often pursued common goals throughout their careers, carving out pieces of the larger problem for individual grants and dissertations.

Complementary Approaches

We study infrastructures that are distributed both geographically, with work performed across multiple sites, and temporally, often extending years or decades into the past and projected long into the future. Whether using quantitative or qualitative methods, studies of infrastructures are snapshots in time, covering the period that researchers can be in the field.

Using teams to study infrastructures allows us to incorporate multiple methodological and epistemological approaches. In our early studies of ADEPT and CENS, two or more researchers worked together on ethnographies of a single infrastructure. In several cases these researchers overlapped in time periods, which allowed us to capture greater geographic and temporal scope of the infrastructure (Borgman et al., 2000, 2005, 2012; Bowker et al., 2009; Darch & Sands, 2017; Mayernik, 2011; Smith et al., 2003; Wallis, 2012). In other cases, a single CKI researcheer conducted all or most of the fieldwork (Boscoe, 2019; Darch & Borgman, 2016;

Pasquetto, 2018; Pasquetto et al., 2019a; Shilton, 2011). To address limitations of scale, our researchers use interviews to draw out the research participants' perspectives on their infrastructures' history and future plans and on challenges of coordinating and collaborating across multiple sites.

To study distributed infrastructures, our research team combines deep, long-term ethnographic observation of one or a few sites (e.g., a single laboratory or a suite of researchers' offices) with shorter periods of observations at other sites (e.g., collaborating laboratories, All-Hands' Meetings, conferences). This approach enables us to combine rich characterizations of research practices in individual sites with comparative analyses of how practices vary and how scientific information (e.g., data, software, physical samples, work plans) travels across collaborating sites (Beaulieu, 2010; Ribes, 2014; Traweek, 1988).

CKI members brought a range of epistemic, methodological, and philosophical commitments towards ethnography to the team. New members arrived with degrees or experience in art, anthropology, archaeology, astrophysics, biology, communication, computer science, ecology, education, engineering, history, mathematics, and philosophy of science, to name a few. Some drew more heavily on the interpretivist tradition, where the ethnographer approaches the field site as a naïve outsider and seeks to build an understanding of practices from the bottom-up (Hammersley & Atkinson, 2007). Other CKI members took a more top-down, positivist approach, entering the field site with a clear set of research questions guiding what to observe (Blomberg & Karasti, 2013). In some cases, our ethnography involved active participation in research subjects' work such as assisting in physical deployment of field sensors, while in other cases, ethnography was purely observational.

These multiple approaches proved complementary. Bottom-up approaches were suited to situations where a researcher was entering a single field site for an extended period of time, whereas top-down approaches using multiple short periods of observation allowed researchers to make the most of limited time in the field.

Researchers' training in ethnography also took a variety of forms. UCLA graduate students on the team took qualitative methods courses in multiple academic departments, drawing on psychology, sociology, communication, anthropology, and gender studies, in addition to courses offered in Information Studies. Mentorship and apprenticeship also played a significant role. Among our many collaborators over these 20 years were scholars well versed in qualitative methods, including Geoffrey Bowker, Paul Edwards, Thomas Finholt, Steven Jackson, Noel Enyedy, David Ribes, William Sandoval, Susan Leigh Star, and Sharon Traweek.

Informed Consent

To maintain access to our interviews, observations, documents, and other data from each project, we needed consent from the participants of our research. Working with our universities' Institutional Review Boards (IRB), we developed consent forms that asked participants for permission to reuse interview recordings, transcripts, and notes in subsequent research by the team. Research participants could opt out of allowing reuse, opt out of recording, and withdraw from the project at any time, but almost all of our participants granted permission for us to reuse their data in later studies. We promised confidentiality, following the usual IRB rules, and did not request permission to contribute the data to a public repository. In later studies, we asked both for consent to participate in the research and a "deed of gift" for the interview recording and transcript to ensure that these documents could be reused. The alternative, which we encountered in our later studies in biomedicine, is to "reconsent" participants to reuse their data in other projects. Locating, contacting, and getting permission from participants interviewed or observed years earlier is untenable. Rather, a broader initial consent process enhances the ability to acquire global data.

Neither audio recordings nor transcripts can be anonymized. We are studying well known scientific projects; others in the field could readily identify individuals if these materials were to be released. Thus, we struck a middle ground that maintained confidentiality while allowing us to reuse data for subsequent projects on related topics.

The human subjects research permissions granted by our IRBs must be renewed annually, with associated reports on data collection and analysis, both retrospective and planned. If we were to allow those permissions to lapse, we would not be able to analyze data from prior projects.

Making Data Global

Our efforts to maintain our data for reuse also began organically. Open science was in its ascendancy in the early days of our research program, and data management was a new topic in the field of Information Studies, known as 'iSchools' (*ISchools*, 2020). The CKI team has consisted largely of information studies students, a field that addresses the collection, selection, organization, curation, preservation, and accessibility of information. This educational background gives iSchools an advantage in acquiring global data and especially in making those data global. As individuals without a background in information studies later joined the team, we trained them in data management skills and in the ethics of data sharing. Even with expertise in data management and social studies of science, making our data global required dedicated effort.

Curation

Curation activities are necessary to make data global, by which we mean processes of standardizing data in ways that they can be integrated with other data into larger models (Edwards, 2010, Chapter 10). In scientific contexts, these are local processes to add metadata, map variant terms to common forms, organize files, store and migrate files, preserve and steward records, and other ways to add value for future use of the data. In the scientific teams we study, most of these curation activities are ad hoc, falling to individual graduate students or post-doctoral fellows who may have minimal (or no) training in data curation. The smaller the team, the more ad hoc the processes tend to be, as the few people involved are able to share their knowledge locally (Baker & Mayernik, 2020; Mayernik, 2016b, 2017; Mayernik et al., 2016).

Of the domains we have studied, astronomy has the most formalized processes of metadata creation, data reduction pipelines, and standard sets of analytical tools. In large astronomy endeavors, such as the Sloan Digital Sky Survey and the Large Synoptic Survey Telescope (recently renamed the Vera C. Rubin Observatory (Rubin Observatory, 2020), as much as half of the overall budget is devoted to data management. We also found that data curation in small astronomy teams tends to fall to graduate students and post-doctoral fellows (Borgman, Darch, et al., 2016; Boscoe, 2019).

Ad Hoc Data Curation

When the 'Borgman Lab,' precursor to the CKI, consisted of two to four individuals, we too took an informal approach to data curation. We stored files on multiple local computers for redundancy, stored paper records in file cabinets, and documented records as necessary for each publication. Each interviewer had full responsibility for transcribing and annotating interview records. After several years of observation, and our first large round of interviews, we had sufficient material that we needed codebooks and metadata to annotate our files. We chose NVivo for qualitative data analysis, importing Word files, and marking up our growing collection of interviews, notes, and other records as NVivo files (*NVivo 8*, 2008). At the time, this software had the best functionality available, despite its limitations in exporting files, as discussed further below.

Scaling Up to Professional Data Management

As grants and teams grew larger, we included data management responsibilities in the job duties of an individual. At first, we hired masters students in information studies at about 25% time to maintain our data resources. They had skills in metadata and records management, could process interview transcripts, keep track of records, and correspond with research participants to set up interviews and send corrected transcripts. Delegating data curation to part-time students sufficed for a few years, but lacked continuity as the team and the data corpus grew in size. When an opportunity arose in late 2013 to reorganize our staffing, we hired a full-time data manager with an MLIS degree and background in the sciences.

An essential, but often under-appreciated role a data manager can play in long-term research projects is to maintain bibliographies. Over the course of 20 years, we have accumulated more than 10,000 references in Zotero (*Zotero*, 2020) that represent the bibliographies of all of our publications, including dissertations and books; references to articles, documents, and books relevant to our research; and documents related to the teams and individuals we study. We mine this rich resource continuously as we write new papers, and as we assemble our annual reports to our funding agencies.

Data management is a growth area for individuals with graduate degrees in library, information, archives, and related areas of study. Data curators are “care givers” who preserve data for future work while maintaining policies, procedures, and promises associated with the context of the research (Baker & Karasti, 2018; Jørn Nielsen & Hjørland, 2014; Scroggins & Pasquetto, 2020).

Our investment in a professional data manager accelerated our ability to acquire global data and to make those data global. By delegating core data curation tasks, plus initial corrections to transcripts and coding, our intent is to give other team members more time for field research, data analysis, and writing. However, removing the responsibility for creating transcripts and routing access to the data corpus through a data manager can add distance between researchers and their participants.

We address this gap between qualitative researchers and their data in several ways. Individual researchers return to their data by listening to recordings, reading transcripts, adding more metadata, and writing interpretive memos to suit their inquiries. We involve the data curator in the research process by participating in selected interviews and observations and in writing. Our data curator, having listened to audio recordings of interviews, cleaned transcripts to correct scientific terminology and idioms, and conducted initial data coding passes, has intimate knowledge of our data resources that transcends that of any individual on the team.

Data Integration Practices

Integrating and reusing data globally requires long-term active management. Even with extensive experience in data management, our researchers find qualitative datasets difficult to integrate, frequently requiring additional cleaning and management prior to comparative analysis. Reflecting back on how we addressed these challenges, our approaches fall into two categories: crosswalks and software tools.

Metadata Crosswalks

Our efforts to make global data began organically, tested through iterative efforts to reuse data as our research on scientific data practices evolved. We began with a method widely used in library and archival practice, which is to build 'crosswalks' between the metadata in each of our protocols (Getty Research Institute, 2020). By comparing questions in different interview protocols, we could integrate our data descriptions into a common codebook, as shown in Figure 1. This continual reintegration reflected observations in our own research sites where teams actively manage historical data to work with newer data collected using modern methods (Boscoe, 2019).

Number	Topic	Code	Scope	Follow the Data Questions	Wind-down Questions
0:01:00	File Attributes	Researcher	Participant name		
0:02:00	File Attributes	Researcher role	This attribute should describe the employment status of the participant, using one of the following terms to describe them: Faculty, Student, Postdoc, or Staff.		
0:03:00	File Attributes	Project	This attribute should be used to record the project with which the participant identifies.	What is the main research project you are working on now?	What research project at CENS are you working on right now?
0:04:00	File Attributes	Domain	This attribute should be used to capture the domain with which the participant identifies	What type of research do you do?	
0:05:00	File Attributes	Interview date	Date of interview in YYYYMMDD		
0:06:00	File Attributes	Interviewers	Interviewers present for the interview		
1:01:00	Project	Project description	General description of the project, this should be a much thicker description of the project than File Attributes – Project.	What is the main research project you are working on now?	What research project at CENS are you working on right now?

Figure 1: Crosswalk of two early CENS protocols.

Following our initial exploration into data reintegration through crosswalks, we continued to modify and supplement our codebooks and protocols with each additional infrastructure studied. The gradual process, transpiring over almost twenty years, is mapped in Figure 2. Rather than abandoning previous research questions, codebooks, or protocols, we adjusted each iteration to incorporate new research topics thus simplifying comparative analysis. When feasible, we reanalyzed earlier data with new research questions in mind.

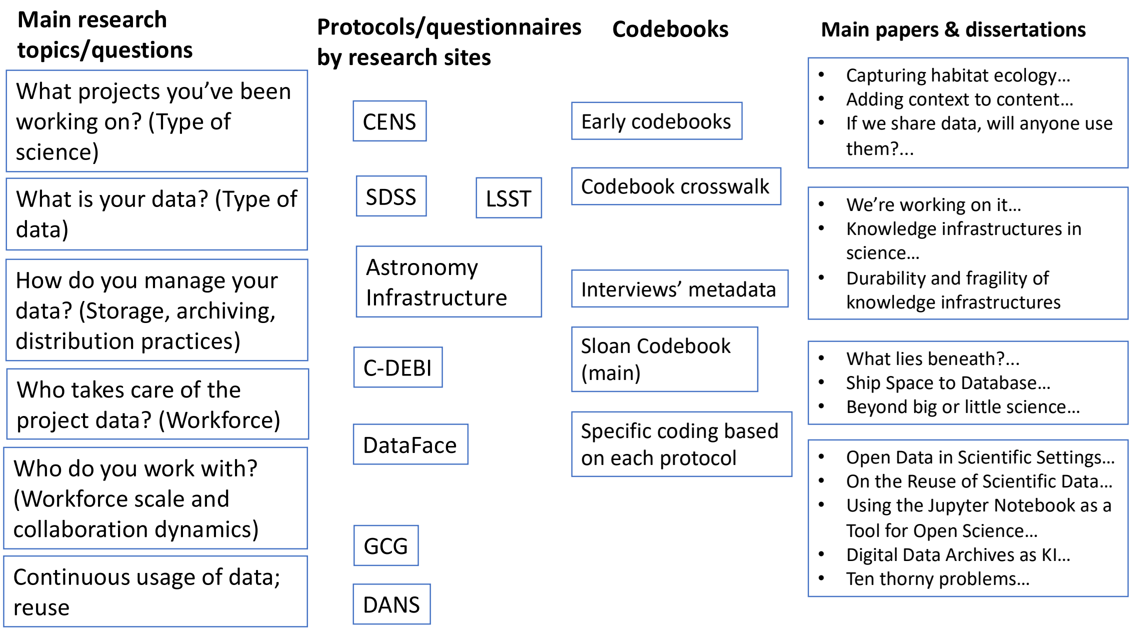
Main research topics/questions
What projects you've been working on? (Type of science)
What is your data? (Type of data)
How do you manage your data? (Storage, archiving, distribution practices)
Who takes care of the project data? (Workforce)
Who do you work with? (Workforce scale and collaboration dynamics)
Continuous usage of data; reuse
Protocols/questionnaires by research sites
CENS
SDSS
LSST
Astronomy Infrastructure
C-DEBI
DataFace
GCG
DANS
Codebooks
Early codebooks
Codebook crosswalk
Interviews' metadata
Sloan Codebook (main)
Specific coding based on each protocol
Main papers & dissertations
Capturing habitat ecology…
Adding context to content…
If we share data, will anyone use them?…
We're working on it…
Knowledge infrastructures in science…
Durability and fragility of knowledge infrastructures
What lies beneath?…
Ship Space to Database…
Beyond big or little science…
Open Data in Scientific Settings…
On the Reuse of Scientific Data…
Using the Jupyter Notebook as a Tool for Open Science…
Digital Data Archives as KI…
Ten thorny problems…

Figure 2: Mapping between research topics, protocols, codebooks, and papers.

Software Tools for Data Analysis

Building these crosswalks between protocols and mapping them to research questions, codebooks, and publications were essential steps in data integration, or making our data global. These steps were not sufficient, however, as our legacy files from NVivo did not integrate easily. We encountered difficulties combining large files that were created in NVivo versions spanning nearly 20 years, some on Apple and some on Windows computers (*NVivo 8*, 2008; *NVivo 9*, 2011; *NVivo 10*, 2013; *NVivo 11*, 2015). To be fair, NVivo and most other qualitative software tools are designed for the canonical situation of one researcher on one project. Our files are large and heterogeneous, and our analytical goals are complex.

One approach to the legacy problem was to start over with different analytical tools. As NVivo did not have the capability to export our files with full analytical markup, we ingested our text files from Word into ATLAS.ti, creating a shared analysis file (*ATLAS.Ti: The Qualitative Data Analysis and Research Software*, 2018). ATLAS.ti has features that facilitate making CKI data global. One feature is a standardized XML format that simplifies sharing a single analytic file among CKI researchers and allows files to be exported. Another feature of ATLAS.ti is a tool derived from discourse analysis that supports more granular markup. The semantic linking features of ATLAS.ti aided in exploring commonalities and differences between activities like "maintenance," thus leveraging the entirety of the CKI data corpus (Scroggins & Pasquetto, 2020).

Marking up interviews in ATLAS.ti proved useful for specific projects conducted by team members with expertise in the tool. However, the overhead of recoding data and learning a new tool proved burdensome for those team members whose materials were already indexed in NVivo. As a consequence, we are maintaining two analytical databases for future investigations.

Keeping Data Alive

By keeping data “alive,” or reusable over long periods of time, we are able to explore research questions at a much larger scale than would otherwise be possible. The ability to reuse data at scale is an inherent and underappreciated challenge of open science (Pasquetto et al., 2017). Here we summarize our lessons learned in collaborative qualitative research at scale by identifying the analyses made possible by our approach and the challenges raised.

Curation and Continuity

Individual professors, as principal investigators, often have long-term research agendas that they pursue with cohorts of students and post-doctoral fellows. Maintaining continuity in the research agenda is difficult due to the high turnover of staff and the short term of research grants, typically one to three years in duration. Grant funding can be a precarious existence. Unless funding periods for staff overlap, one cohort leaves before the next arrives, leaving substantial gaps in team knowledge (Jackson et al., 2011).

The CKI team’s investment in keeping our data alive has facilitated our continuity across cohorts and over long periods of time. In recent years, the data curator has trained each new graduate student and post-doctoral fellow in how to use our rich collection of data resources. As a consequence, new team members can build upon the work of prior staff. Alumni of the team also continue to collaborate, writing joint papers and helping to mine data they collected earlier.

Our challenges in continuity echo those of the teams we study. No matter how well documented, and how much knowledge is passed from one cohort to the next, the individual researchers who collected the data initially retain the deepest knowledge of context. As with our scientific teams, we contact our colleagues for further interpretation as needed, collaborating if substantial data integration will accomplish goals of mutual interest (Pasquetto et al., 2019b).

New Grants, New Research Questions

Each new grant proposal must promise something new and innovative. Rarely can incremental funding be acquired successfully. Here the challenge is to propose new work that builds on the prior, without losing the continuity of the larger research agenda. We have focused on questions of data practices; research teams’ abilities to share and reuse data; interactions between science policy and local practice; the concept of “data” as understood within and between scientific domains; and how knowledge infrastructures facilitate and constrain scientific work. This is a sufficiently broad agenda to allow us to ask new questions about how open science practices and policy play out in different domains, how old standards and new tools fit into knowledge infrastructures, and how infrastructures evolve and interact over time. We have sought funding from a wide range of sources, aligning our questions with funding agency interests in individual sciences, in education, in infrastructure, in policy, and in scholarly communication.

As our curated data resources accumulate, we pitch them as a competitive advantage in seeking new funding. These resources also provide competitive advantage in hiring new graduate students and post-doctoral fellows. By joining our team, they have access to these data for use in constructing their own research agendas. We put graduate students into the field in their first year of study, which gives them at least two years of research experience by the time they begin their dissertations.

Diversity of Data Analysis

Maintaining a core set of research questions about data, data practices, and infrastructure across our projects and grants provides the continuity necessary to study knowledge infrastructures at scale. At the same time, we are careful not to be overly prescriptive in the details of study design or data analysis. Research questions in an individual grant are general enough to allow students considerable flexibility in pursuing their interests and following their instincts in data analysis. Some of our graduate students have relied more heavily on ethnographic observation and writing memos, some on interviews, and some on document analyses. One dissertation was primarily quantitative, building upon the team's qualitative research findings (Pepe, 2010). The balance of methods varies intentionally. Some CKI members do extensive coding in NVivo or ATLAS.ti, which provides our best records for further analyses. Others do basic coding with these tools, then print out sections for hand-coding with colored markers. The latter approach provides flexibility, but does not scale well beyond the space of a table or office floor.

As qualitative researchers steeped in grounded theory (Clarke, 2005; Glaser & Strauss, 1967), we encourage hypothesis building and testing, and iterative analyses. These can be done in many ways, some of which provide better record trails than others. For continuity purposes, we much prefer granular documentation. One of the factors distinguishing analytic practices is the amount of training in the information fields. Researchers with library and information science degrees (MLIS or equivalent) have spent more time on developing documentation, tables, protocols, and codebooks than have team members who came from technical, historical, or social science backgrounds. Librarians on the team more often ensured that their data were organized, well-described, stored in the secured CKI archive, and available for sharing with other team members. Before we hired a professional data curator, MLIS-trained researchers developed the Zotero library and maintained detailed bibliographies of CKI publications. CKI researchers without a background in the information field have required more encouragement to manage their data as a collective resource. We make sure to ingest their coded data and bibliographies before they leave our employ, usually when completing their degrees.

Our approach to managing our data falls between two extremes in academic research. One, more common in the social sciences, is for individual students and post-doctoral fellows to maintain exclusive control over their data, not leaving copies behind for the supervisor or team. University practices vary widely in their expectations of students to have exclusive or non-exclusive rights over their knowledge products. The distinction may be a function of whether the research was conducted under grant funding or self-funded, and which open science policies may apply, whether by institution or government.

At the other extreme are open science policies that promote transparency throughout the entire research process, from registering hypotheses to depositing datasets and code (National Academies of Sciences, Engineering, and Medicine, 2018; Nosek et al., 2015). These approaches are intended to increase reproducibility, accountability, and the ability to reuse data. Complete transparency of research is particularly problematic with qualitative human subjects research. Audio recordings and transcripts cannot be fully anonymized, especially for interviews conducted with well known science teams. Fully transparent approaches to open science also are controversial because of the resources required for investigators to comply with regulations, competitive advantages of researchers subject to different rules, and economic benefits that may accrue to external parties (Mirowski, 2018).

Conclusions

Twenty years of collaborative qualitative research have enabled us to address big questions about how knowledge infrastructures develop, how they are used, when they are visible and when invisible, when they are robust, when they are fragile, and how they break down. We have learned that all infrastructures are fragile in the long run, no matter how robust they may appear in the present (Borgman, Darch, et al., 2016).

Similarly, “data” is the most complex concept in data science (Borgman, 2015, 2019). Throughout our research, we find that one person’s signal is another’s noise. Two researchers, working side by side, may not realize they have fundamentally different notions of essential variables such as “temperature,” as we found in CENS (Borgman et al., 2012). An astronomy team that removed gas clouds from their images as part of their data reduction pipeline later hired a specialist in gas clouds. Instead of treating these gases as noise, they began to treat them as signals, offering new insights into their research program. Examples abound of nuanced interpretations of data and datasets; policies about openness, sharing, and reuse; and responsibilities for stewarding scientific knowledge products.

Data sharing and reuse, in turn, depend on the availability of knowledge infrastructures, on characteristics of the research domain, and on how competing stakeholders implement (or not) such policies. Throughout our research, we have attempted to identify factors that distinguish data practices, whether by domain, discipline, institution, career stage, scale, temporality, or public policies. Often we are asked to compare practices by these or other factors, whether by funding agencies, policy makers, reviewers, or audiences at public talks. The larger the corpus of data practices and infrastructures we build, the more nuanced our conclusions become. Our efforts to distinguish data handling practices within a distributed interdisciplinary collaboration, for example, revealed that each individual researcher claimed multiple areas of disciplinary expertise, thus defying categorization. Of particular interest was how practices evolved as these people worked together, learning from each other. The disciplinary and methods training each person brought to the lab became part of an emerging set of practices. Any attempt to characterize how a discipline handles data may be a snapshot in time; generalizations are fraught. Focusing on the larger knowledge infrastructures in which these practices occur gives us a broader understanding of scientists’ experiences, technologies, and access to the resources necessary to manage their data.

Acknowledgements

Research reported here was supported in part by grants from the National Science Foundation (NSF) and Alfred P. Sloan Foundation (Sloan): (1) ADEPT was funded by National Science Foundation grant no. IIS-9817432, Terence R. Smith, University of California, Santa Barbara, PI); (2) The Center for Embedded Networked Sensing (CENS) was funded by NSF Cooperative Agreement #CCR-0120778; (3) CENSNet: An Architecture for Authentic Web-based Science Inquiry in Middle and High School, NSF #ESI-0352572; (4) Towards a Virtual Organization for Data Cyberinfrastructure, NSF #OCI-0750529; (5) Monitoring, Modeling & Memory: Dynamics of Data and Knowledge in Scientific Cyberinfrastructures, NSF #0827322; (6) If Data Sharing is the Answer, What is the Question?, Sloan #2015-14001. Microsoft Research, Technical Computing Initiative, led by Tony Hey, provided additional support for CENS research. Bernadette Boscoe and Michael Scroggins provided helpful commentary on earlier drafts. Joshua Greenberg made constructive suggestions on methodological issues to address. Most of all, we acknowledge the generosity of the scientific research teams we study for welcoming us into their communities as observers and for their thoughtful discussions of data practices.

References

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messina, P., Messerschmitt, D. G., Ostriker, J. P., & Wright, M. H. (2003). *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon panel on Cyberinfrastructure* (p. 84). National Science Foundation. http://www.nsf.gov/cise/sci/reports/atkins.pdf

ATLAS.ti: The qualitative data analysis and research software (Version 8). (2018). [Computer software]. ATLAS.ti Scientific Software Development GmbH. https://atlasti.com/product/

Baker, K. S., & Karasti, H. (2018). Data Care and Its Politics: Designing for Local Collective Data Management as a Neglected Thing. *Proceedings of the 15th Participatory Design Conference: Full Papers - Volume 1*, 10:1–10:12. https://doi.org/10.1145/3210586.3210587

Baker, K. S., & Mayernik, M. S. (2020). Disentangling knowledge production and data production. *Ecosphere*, *11*(7), e03191. https://doi.org/10.1002/ecs2.3191

Beaulieu, A. (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. https://doi.org/10.1177/0306312709359219

Blomberg, J., & Karasti, H. (2013). Reflections on 25 Years of Ethnography in CSCW. *Computer Supported Cooperative Work (CSCW)*, *22*(4–6), 373–423. https://doi.org/10.1007/s10606-012-9183-1

Borgman, C. L. (2000). *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. MIT Press.

Borgman, C. L. (2006). What can Studies of e-Learning Teach us about Collaboration in e-Research? Some Findings from Digital Library Studies. *Computer Supported Cooperative Work*, *15*(4), 359–383. https://doi.org/10.1007/s10606-006-9024-1

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT Press.

Borgman, C. L. (2019). The lives and after lives of data. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.9a36bdb6

Borgman, C. L., Darch, P. T., Pasquetto, I. V., & Wofford, M. F. (2020). *Our knowledge of knowledge infrastructures: Lessons learned and future directions* (Alfred P. Sloan Foundation, p. 27). University of California, Los Angeles. http://escholarship.org/uc/item/9rm6b7d4

Borgman, C. L., Darch, P. T., Sands, A. E., & Golshan, M. S. (2016). The durability and fragility of knowledge infrastructures: Lessons learned from astronomy. *Proceedings of the Association for Information Science and Technology*, *53*, 1–10. http://dx.doi.org/10.1002/pra2.2016.14505301057

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., & Traweek, S. (2015). Knowledge infrastructures in science: Data, diversity, and digital libraries. *International Journal on Digital Libraries*, *16*(3–4), 207–227. https://doi.org/10.1007/s00799-015-0157-z

Borgman, C. L., Darch, P. T., Sands, A. E., Wallis, J. C., & Traweek, S. (2014). The Ups and Downs of Knowledge Infrastructures in Science: Implications for Data Management. *Proceedings of the 2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, 257–266. https://doi.org/10.1109/JCDL.2014.6970177

Borgman, C. L., Gilliland-Swetland, A. J., Leazer, G. H., Mayer, R., Gwynn, D., Gazan, R., & Mautone, P. (2000). Evaluating Digital Libraries for Teaching and Learning in Undergraduate Education: A Case Study of the Alexandria Digital Earth ProtoType (ADEPT). *Library Trends*, *49*(2), 228–250.

Borgman, C. L., Golshan, M. S., Sands, A. E., Wallis, J. C., Cummings, R. L., Darch, P. T., & Randles, B. M. (2016). Data Management in the Long Tail: Science, Software, and Service. *International Journal of Digital Curation*, *11*(1), 128–149. https://doi.org/10.2218/ijdc.v11i1.428

Borgman, C. L., Leazer, G. H., Gilliland-Swetland, A. J., Millwood, K. A., Champeny, L., Finley, J. R., & Smart, L. J. (2004). How geography professors select materials for classroom lectures: Implications for the design of digital libraries. *JCDL '04: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries (Tucson, AZ, June 7-11, 2004)*, 179–185.

Borgman, C. L., Smart, L. J., Millwood, K. A., Finley, J. R., Champeny, L., Gilliland, A. J., & Leazer, G. H. (2005). Comparing faculty information seeking in teaching and research: Implications for the design of digital libraries. *Journal of the American Society for Information Science and Technology*, *56*(6), 636–657. https://doi.org/10.1002/asi.20154

Borgman, C. L., Wallis, J. C., & Enyedy, N. D. (2006). Building Digital Libraries for Scientific Data: An Exploratory Study of Data Practices in Habitat Ecology. In J. Gonzalo, C. Thanos, M. F. Verdejo, & R. C. Carrasco (Eds.), *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries: Vol. LINCS 4172* (pp. 170–183). Springer Berlin Heidelberg. https://doi.org/10.1007/11863878_15

Borgman, C. L., Wallis, J. C., & Mayernik, M. S. (2012). Who's got the data? Interdependencies in science and technology collaborations. *Computer Supported Cooperative Work*, *21*(6), 485–523. https://doi.org/10.1007/s10606-012-9169-z

Borgman, C. L., Wallis, J. C., Mayernik, M. S., & Pepe, A. (2007). Drowning in Data: Digital library architecture to support scientific use of embedded sensor networks. *Joint Conference on Digital Libraries*, 269–277. https://doi.org/10.1145/1255175.1255228

Bos, N. M., Zimmerman, A. S., Olson, J. S., Yew, J., Yerkie, J., Dahl, E., Cooney, D., & Olson, G. M. (2008). From Shared Databases to Communities of Practice: A Taxonomy of Collaboratories. In G. M. Olson, A. S. Zimmerman, & N. M. Bos (Eds.), *Scientific Collaboration on the Internet* (pp. 52–72). MIT Press. https://doi.org/10.7551/mitpress/9780262151207.003.0004

Bos, N. M., Zimmerman, A. S., Olson, J., Yew, J., Yerkie, J., Dahl, E., & Olson, G. M. (2007). From shared databases to communities of practice: A taxonomy of collaboratories. *Journal of Computer-Mediated Communication*, *12*, 563–582. https://doi.org/10.1111/j.1083-6101.2007.00343.x

Boscoe, B. M. (2019). *From Blurry Space to a Sharper Sky: Keeping Twenty-Three Years of Astronomical Data Alive* [Ph.D. Dissertation, University of California, Los Angeles]. https://escholarship.org/uc/item/2jv941sb

Bowker, G. C., Baker, K., Millerand, F., & Ribes, D. (2009). Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment. In J. Hunsinger, L. Klastrup, &

M. Allen (Eds.), *International Handbook of Internet Research* (pp. 97–117). Springer Netherlands. http://link.springer.com/10.1007/978-1-4020-9789-8_5

Clarke, A. E. (2005). *Situational analysis: Grounded theory after the postmodern turn*. SAGE Publications.

Collins, H. M., & Evans, R. (2007). *Rethinking Expertise*. University of Chicago Press.

Committee on the Science of Team Science, Board on Behavioral, Cognitive, and Sensory Sciences, Division of Behavioral and Social Sciences and Education, & National Research Council. (2015). *Enhancing the Effectiveness of Team Science* (N. J. Cooke & M. L. Hilton, Eds.). The National Academies Press. http://www.nap.edu/catalog/19007/enhancing-the-effectiveness-of-team-science

Cummings, J. N., Finholt, T. A., Foster, I., Kesselman, C., & Lawrence, K. A. (2008). Beyond Being There: A Blueprint for Advancing the Design. *Development, and Evaluation of Virtual Organizations: Report from an NSF Workshop on Developing VIrtual Organizations*. https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/PB2009100587.xhtml

Cummings, J. N., & Kiesler, S. (2004). *Collaborative research across disciplinary and institutional boundaries*. National Science Foundation. http://hciresearch.hcii.cs.cmu.edu/complexcollab/pubs/paperPDFs/cummings_collaborative.pdf

Darch, P. T. (2016, March 15). Many methods, many microbes: Methodological diversity and standardization in the deep subseafloor biosphere. *IConference 2016 Proceedings*. iConference 2016, Philadelphia, PA. https://doi.org/10.9776/16246

Darch, P. T. (2018). *Limits to the Pursuit of Reproducibility: Emergent Data-Scarce Domains of Science*. iConference 2018, Sheffield, UK. https://doi.org/10.1007/978-3-319-78105-1_21

Darch, P. T., & Borgman, C. L. (2016). Ship space to database: Emerging infrastructures for studies of the deep subseafloor biosphere. *PeerJ Computer Science*, *2*, e97. https://doi.org/10.7717/peerj-cs.97

Darch, P. T., Borgman, C. L., Traweek, S., Cummings, R. L., Wallis, J. C., & Sands, A. E. (2015). What lies beneath?: Knowledge infrastructures in the subseafloor biosphere and beyond. *International Journal on Digital Libraries*, *16*(1), 61–77. https://doi.org/10.1007/s00799-015-0137-3

Darch, P. T., & Sands, A. E. (2017). Uncertainty about the Long-Term: Digital Libraries, Astronomy Data, and Open Source Software. *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 1–4. https://doi.org/10.1109/JCDL.2017.7991584

Darch, P. T., Sands, A. E., Borgman, C. L., & Golshan, M. S. (2020a). Library Cultures of Data Curation: Adventures in Astronomy. *Journal of the Association for Information Science and Technology*.

Darch, P. T., Sands, A. E., Borgman, C. L., & Golshan, M. S. (2020b). Library Cultures of Data Curation: Adventures in Astronomy. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24345

Darch, P. T., Sands, A. E., Borgman, C. L., & Golshan, M. S. (2020c). Do the stars align?: Stakeholders and strategies in libraries' curation of an astronomy dataset. *Journal of the Association for Information Science and Technology*, 1–14. https://doi.org/10.1002/asi.24392

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. MIT Press.

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., Burton, M., & Calvert, S. (2013). *Knowledge infrastructures: Intellectual frameworks and research challenges* (p. 40). University of Michigan. http://hdl.handle.net/2027.42/97552

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science Friction: Data, Metadata, and Collaboration. *Social Studies of Science*, *41*(5), 667–690. https://doi.org/10.1177/0306312711413314

Getty Research Institute. (2020). *Introduction to Metadata: Crosswalk (Getty Research Institute)*. https://www.getty.edu/research/publications/electronic_publications/intrometadata/crosswalks.html

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine Publishing.

Gorman, M. E. (Ed.). (2010). *Trading Zones and Interactional Expertise: Creating New Kinds of Collaboration*. MIT Press.

Hammersley, M., & Atkinson, P. (2007). *Ethnography: Principles in Practice* (3rd ed.). Routledge.

ISchools. (2020). https://ischools.org/

Jackson, S. J., Ribes, D., Buyuktur, A., & Bowker, G. C. (2011). Collaborative Rhythm: Temporal Dissonance and Alignment in Collaborative Scientific Work. *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, 245–254. https://doi.org/10.1145/1958824.1958861

Jirotka, M., Lee, C. P., & Olson, G. M. (2013). Supporting Scientific Collaboration: Methods, Tools and Concepts. *Computer Supported Cooperative Work (CSCW)*, *22*(4–6), 667–715. https://doi.org/10.1007/s10606-012-9184-0

Jørn Nielsen, H., & Hjørland, B. (2014). Curating research data: The potential roles of libraries and information professionals. *Journal of Documentation*, *70*(2), 221–240. https://doi.org/10.1108/JD-03-2013-0034

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard University Press.

Leonelli, S. (2019). Data Governance is Key to Interpretation: Reconceptualizing Data in Data Science. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.17405bb6

Mayer, R. E., Smith, T. R., Borgman, C. L., & Smart, L. J. (2002). Digital libraries as instructional aids for knowledge construction. *Educational Technology*, *42*, 38–42.

Mayernik, M. S. (2011). *Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators* [PhD Dissertation, UCLA]. http://dx.doi.org/10.2139/ssrn.2042653

Mayernik, M. S. (2016a). Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, *67*(4), 973–993. https://doi.org/10.1002/asi.23425

Mayernik, M. S. (2016b). Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, *67*(4), 973–993. https://doi.org/10.1002/asi.23425

Mayernik, M. S. (2017). Open data: Accountability and transparency. *Big Data & Society*, *4*(2), 205395171771885. https://doi.org/10.1177/2053951717718853

Mayernik, M. S., Hart, D. L., Maull, K. E., & Weber, N. M. (2016). Assessing and tracing the outcomes and impact of research infrastructures. *Journal of the Association for Information Science and Technology*, n/a-n/a. https://doi.org/10.1002/asi.23721

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

Mirowski, P. (2018). The future(s) of open science. *Social Studies of Science*, *48*(2), 171–203. https://doi.org/10.1177/0306312718772086

National Academies of Sciences, Engineering, and Medicine. (2018). *Open Science by Design: Realizing a Vision for 21st Century Research*. The National Academies Press. https://doi.org/10.17226/25116

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

NVivo 8 research software for analysis and insight. (2008). QRS International.

NVivo 9 research software for analysis and insight. (2011). QSR International.

NVivo 10 research software for analysis and insight. (2013). QRS International. http://www.qsrinternational.com/products_nvivo.aspx

NVivo 11 research software for analysis and insight. (2015). QSR International. http://help-nv11.qsrinternational.com/desktop/welcome/welcome.htm

Olson, G. M., Zimmerman, A. S., & Bos, N. M. (Eds.). (2008). *Scientific Collaboration on the Internet*. MIT Press. http://ieeexplore.ieee.org/xpl/bkabstractplus.jsp?bkn=6267421

Organisation for Economic Co-operation and Development. (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding* (p. 24). Organisation for Economic Co-Operation and Development. http://www.oecd.org/dataoecd/9/61/38500813.pdf

Pasquetto, I. V. (2018). *From Open Data to Knowledge Production: Biomedical Data Sharing and Unpredictable Data Reuses* [Ph.D. Dissertation, UCLA]. https://escholarship.org/uc/item/1sx7v77r

Pasquetto, I. V. (2019, March 23). Do Scientists Reuse Open Data? [Blog Post]. *Sage Bionetworks Critical Assessment of Open Science (CAOS)*. http://sagebionetworks.org/in-the-news/do-scientists-reuse-open-data/

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019a). Uses and Reuses of Scientific Data: The Data Creators' Advantage. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.fc14bf2d

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019b). Uses and Reuses of Scientific Data: The Data Creators' Advantage. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.fc14bf2d

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the Reuse of Scientific Data. *Data Science Journal*, *16*. https://doi.org/10.5334/dsj-2017-008

Pepe, A. (2010). *Structure and Evolution of Scientific Collaboration Networks in a Modern Research Collaboratory* [Ph.D. Dissertation, University of California, Los Angeles]. http://dx.doi.org/10.2139/ssrn.1616935

Ribes, D. (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*, 158–170. http://dl.acm.org/citation.cfm?id=2531624

Ribes, D., & Finholt, T. (2009). The Long Now of Technology Infrastructure: Articulating Tensions in Development. *Journal of the Association for Information Systems*, *10*(5). https://doi.org/10.17705/1jais.00199

Rubin Observatory. (2020). *First national US observatory to be named after a woman!* https://www.lsst.org/news/vro-press-release

Sands, A. E. (2017). *Managing Astronomy Research Data: Data Practices in the Sloan Digital Sky Survey and Large Synoptic Survey Telescope Projects* [Ph.D. Dissertation, UCLA]. http://escholarship.org/uc/item/80p1w0pm

Scroggins, M. J., & Pasquetto, I. V. (2020). Labor Out of Place: On the Varieties and Valences of (In)visible Labor in Data-Intensive Science. *Engaging Science, Technology, and Society*, *6*(0), 111–132. https://doi.org/10.17351/ests2020.341

Scroggins, M. J., Pasquetto, I. V., Geiger, R. S., Boscoe, B. M., Darch, P. T., Cabasse-Mazel, C., Thompson, C., Golshan, M. S., & Borgman, C. L. (2020). Thorny problems in data (-intensive) science. *Communications of the ACM*, *63*(8), 30–32. https://doi.org/10.1145/3408047

Shilton, K. (2011). *Building Values into the Design of Pervasive Mobile Technologies* [Ph.D. Dissertation, University of California, Los Angeles]. http://ssrn.com/paper=1866783

Smith, T. R., Ancona, D., Buchel, O., Freeston, M., Heller, W., Nottrott, R., Tierney, T., & Ushakov, A. (2003). The ADEPT concept-based digital learning environment. In T. Koch & I. T. Solvberg (Eds.), *European Conference on Digital Libraries* (pp. 300–312). Springer.

Star, S. L., & Ruhleder, K. (1996). Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research*, *7*(1), 111–134. https://doi.org/10.1287/isre.7.1.111

Stokes, D. (1997). *Pasteur's Quadrant: Basic Science and Technological Innovation*. Brookings Institution Press.

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

Wagner, C. S. (2018). *The Collaborative Era in Science*. Springer.

Wallis, J. C. (2012). *The Distribution of Data Management Responsibility within Scientific Research Groups* [PhD dissertation, UCLA]. https://escholarship.org/uc/item/46d896fm

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

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013a). 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. https://doi.org/10.1371/journal.pone.0067332

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013b). 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. https://doi.org/10.1371/journal.pone.0067332

Wofford, M. F., Boscoe, B. M., Borgman, C. L., Pasquetto, I. V., & Golshan, M. S. (2020). Jupyter notebooks as discovery mechanisms for open science: Citation practices in the astronomy community. *Computing in Science Engineering*, *22*(1), 5–15. https://doi.org/10.1109/MCSE.2019.2932067

Zotero. (2020). https://www.zotero.org/