Harvard Data Science Review • 1.2

Uses and Reuses of Scientific Data: The Data Creators' Advantage

Irene V. Pasquetto, Christine L. Borgman, Morgan F. Wofford

Published on: Nov 15, 2019
Updated on: Nov 23, 2019
DOI: 10.1162/99608f92.fc14bf2d

ABSTRACT

Open access to data, as a core principle of open science, is predicated on assumptions that scientific data can be reused by other researchers. We test those assumptions by asking where scientists find reusable data, how they reuse those data, and how they interpret data they did not collect themselves. By conducting a qualitative meta-analysis of evidence on two long-term, distributed, interdisciplinary consortia, we found that scientists frequently sought data from public collections and from other researchers for comparative purposes such as "ground-truthing" and calibration. When they sought others' data for reanalysis or for combining with their own data, which was relatively rare, most preferred to collaborate with the data creators. We propose a typology of data reuses ranging from comparative to integrative. Comparative data reuse requires interactional expertise, which involves knowing enough about the data to assess their quality and value for a specific comparison such as calibrating an instrument in a lab experiment. Integrative reuse requires contributory expertise, which involves the ability to perform the action, such as reusing data in a new experiment. Data integration requires more specialized scientific knowledge and deeper levels of epistemic trust in the knowledge products. Metadata, ontologies, and other forms of curation benefit interpretation for any kind of data reuse. Based on these findings, we 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. Data reuse is a process that occurs within knowledge infrastructures that evolve over time, encompassing expertise, trust, communities, technologies, policies, resources, and institutions.

(*See Supplementary Materials for methodological and other details, including a full bibliography.)*

Keywords: data, science, reuse, biomedicine, environmental sciences, open science, data practices, science policy

1. Introduction and Problem Statement

Scientific practice and public policy continue to move toward open access to publications, data, software, code, and other research products. To provide open

access to research data, stakeholders must build digital archives, populate those archives, and maintain them. While all of these costly public investments are necessary for data reuse, they are not sufficient to ensure that those data are useful for further research, nor that those assets will be reused. Scientists and other scholars develop deep expertise in their research domain, methods, and tools, all of which become integral to the data they collect, analyze, interpret, report in publications, and may later deposit in digital archives. As a consequence of the expertise involved in their creation, data are difficult to extricate from the context in which they originated (Latour, 1987).

An important question for the sciences and for public policy is to ask what kinds of data reuse are made possible by access to public data archives and what kinds are not. When scientists seek data from sources beyond their own laboratories and current collaborations, under what conditions do public data suffice? When do scientists pursue interpersonal contact for further expertise about those data and their contexts of origin? How does data reuse vary by research domain, purposes for potential reuse, access to data creators, and time period? Answers to these questions can guide the design of digital archives, policies for data governance, and public policy for open access to data.

We have studied scientific data practices over a period of two decades. This article reports the findings of a qualitative meta-analysis of research on two large, long-term, distributed, and interdisciplinary scientific consortia. We found similar patterns of data reuse within and between consortia, despite considerable variation in research domains, access to data, research methods, and time periods. We combine ethnographic, interview, and documentary evidence to present a theoretical framework for a continuum of types of data reuses. At one extreme, comparative data reuse can be accomplished with access to publicly available data archives, which promotes greater equity in science. At the other extreme, integrative data reuse is most effective when accomplished in collaboration with the data creators, leading to a scientific advantage for these players. We explore the reasons and conditions for these different types of data reuses to theorize the data creators' advantage.

2. Data Practices in the Sciences

Data sharing policies presume that research data are useful to others and that others will reuse those data (European Commission High Level Expert Group on Scientific

Data, 2010; Hanson, Sugden, & Alberts, 2011; Wilkinson et al., 2016). A common argument for data sharing is to avoid duplication of research effort, thus accelerating the pace of science (Rung & Brazma, 2012).

A growing body of social science research reveals that sharing data is a complex sociotechnical process, making it hard to predict by whom, when, how, why, and whether scientific data will be reused (Borgman, 2015; Mosconi et al., 2019). These complexities are explored here to frame our empirical study.

2.1. Data Reuse and Reproducibility

Many stakeholders in the sciences and social sciences are concerned about a 'reproducibility crisis,' given the difficulties in reproducing research based on information in journal articles (Center for Open Science, 2015; McNutt, 2014). The reproducibility argument for data release is usually based on grounds of transparency, verifiability, or accountability. However, arguments for reproducibility often founder on disputes over what constitutes reproduction, such as reanalyzing published data, repeating the study, reprocessing the 'raw' data, or replicating the findings under different conditions. A recent National Academies of Sciences consensus study (2019) constrained these terms to distinguish between *computational reproducibility*, *replicability* as obtaining consistent results across multiple studies, and *generalizability* as the extent to which findings apply to other domains.

Data reuse is a broader concept that incorporates many different activities, such as returning to one's own data for later comparisons, acquiring datasets from public or private sources to compare to newly collected data, surveying available datasets as background research for a new project, or conducting reanalyses of one or more datasets to address new research questions. These activities vary widely in their implications for scientific practice, for the design of data archives, for public policy, and for data science. A starting point to conceptualize data reuse is to distinguish between 'uses' and 'reuses' of a dataset; distinctions are both contextual and temporal. In situations where data are collected by one individual (or one team) for a research project, the first use of a dataset typically occurs when an individual or team explores it for a specific question. If the same individual returns to that same dataset later, that is another use, as the dataset is still embedded in the creators' context. When that dataset is contributed to a repository, retrieved by someone else, and deployed for another project, the action becomes a reuse of the dataset. Reuse thus implies usage of a dataset by someone other than the originator (Pasquetto, Randles, & Borgman, 2017).

Data reuse is a process that may occur over long periods of time. It is not a single action or product that is readily countable. Thus, data reuse is best studied with qualitative methods that afford the opportunity to observe the circumstances in which researchers find themselves in need of others' data, how and when they pursue data, how they evaluate the usefulness of data for a given purpose, how they interpret others' data, and how they deploy those data in their own research. Quantitative methods such as surveys and content analyses can identify researchers' intentions to share or reuse data (Federer, 2019; Tenopir et al., 2011), rather than actual processes of reuse, which is our concern in this article.

2.2. Documenting Data for Reuse

Research data are not simply found in nature, but are crafted for specific research purposes. As a consequence, research data—like all data—are local and historically situated artifacts (Gitelman, 2013; Star & Griesemer, 1989). Studies grounded in the social sciences and philosophy continue to encounter tensions between the ability to repurpose research data in different contexts and the information loss that occurs when data are removed from their original contexts of production (Borgman, 2015; Leonelli, 2016; Loukissas, 2019).

To repurpose data collected by others, researchers need contextual information such as documentation about equipment, protocols, procedures for collecting and processing data, and experimental or laboratory conditions of data handling (Culina, Crowther, Ramakers, Gienapp, & Visser, 2018). Metadata schemas and ontologies are means to formalize and transfer such information (Mayernik, 2016; Mayernik & Acker, 2017). Leonelli (2010), a historian and philosopher of science, distinguishes between ontologies as "relevance labels" and metadata as "reliability labels" that situate many "small facts." Together, these mechanisms associate datasets with specific research objects (e.g., the biological entity under study) and provide information about the quality of the data, such as data format, organisms used in experiments, instruments and methods applied, and laboratory conditions under which data were obtained.

When successfully applied, metadata and ontologies help research data to perform as "mobile" objects (Latour, 1987), meaning objects that can move between different contexts of production while retaining sufficient evidentiary power. These labels also allow datasets to work as "boundary objects": objects that are both plastic enough to adapt to local conditions, yet robust enough to maintain a common identity across sites (Star & Griesemer, 1989). Investments in metadata and ontologies can make data more

amenable to integration and metanalyses, and may lead to higher rates of data reuse (Jones, Schildhauer, Reichman, & Bowers, 2006). Borrowing an economics term, Leonelli (2015) refers to transferable research data as "fungible objects" defined by "their portability and their prospective usefulness as evidence" (p. 810).

2.3. Trust in Data

Among the most essential but intangible aspects of data reuse is the ability to trust data collected by others. Scientific practice depends upon the ability to trust knowledge claims and products of others, a concept known as 'epistemic trust' (Darch, 2019; Porter, 1996; Shapin, 1994). Epistemic trust has several dimensions, and is relational, rather than something inherent in a dataset. One dimension is interpersonal trust, such as trust in the team that created a dataset (Prieto, 2009). For example, Jirotka et al.'s (2005) study of distributed readings of mammograms revealed strategies for assessing trustworthiness based on whether the data creator was known to produce reliable data. Similarly, ecologists assess data by disciplinary standards involved in their production and by reputation of the data creator (Yakel, Faniel, Kriesberg, & Yoon, 2013).

Other dimensions of trust include the ability to evaluate the quality of data, the reputations of the archives that host relevant datasets, and organizations responsible for the data curation process (Bietz & Lee, 2009; Borgman, Scharnhorst, & Golshan, 2019; Faniel & Jacobsen, 2010).

In an influential policy report, the U.S. National Science Board (2005) categorized data collections along a continuum from local to global uses. *Research data collections* are those that result from focused research projects; curation is limited. *Resource collections* serve a community, have more extensive curation, and establish community-level standards. *Reference collections* are broader in scope, serve large communities, and conform to robust and comprehensive standards. The latter are intended to promote epistemic trust by their communities.

Our hypothesis, posed in the 2015 grant proposal that supported this research, is that centralized data collection requires early agreements on data management, resulting in particular kinds of expertise and disciplinary configurations, whereas distributed data collection is more flexible and adaptive to local conditions, but the resulting datasets are more difficult to integrate later. Centralized data collection and curation, such as reference collections and sky surveys, results in standardized datasets that are

valuable for comparative reuses. Investigator-led projects are based in specific sets of research questions, models, theories, methods, and knowledge infrastructures. The resulting datasets, whether or not contributed to public collections, will be more idiosyncratic than those designed and curated for standardized comparisons (Borgman et al., 2015; Boscoe, 2019; Darch & Borgman, 2016; Sands, 2017).

2.4. Data and Infrastructure

The ability to create, use, document, and reuse scientific data depends on access to knowledge about those data and how they were created. Data creators usually have the most intimate knowledge about a given dataset, gained while designing, collecting, processing, analyzing and interpreting the data. Many individuals may participate in data creation, hence knowledge may be distributed among multiple parties over time. Later reusers of datasets seek whatever knowledge is needed through metadata, documentation, contact with data creators, or other means. The types of knowledge needed will vary by reusers' distance from the origin of the data, whether distance in time, domain, expertise, resources, or other factors; and by intended purposes of reuse (Borgman, 2015).

Data, knowledge, and expertise. Differences in knowledge and expertise, a core problem of epistemology, help to explain data reuse. The most general distinction is between "knowledge that" and "knowledge how" (Ryle, 1949). An individual might have descriptive knowledge, or knowledge that, but not be able to explain how something works. Procedural knowledge is a more advanced form of expertise, where the person has knowledge how to do something. As people become more skilled at a task, they gain tacit knowledge, a form of expertise that is difficult to articulate (Polanyi, 1966). Hilgartner and Brandt-Rauf (1994) discuss the transition of a scientific innovation from the "magic hands" of the expert who developed tacit knowledge in the innovation process, to "kits" as procedures become more standardized. The expert has a competitive advantage in the initial stages until the process becomes common and repeatable.

Collins and Evans (2007) distinguish "ubiquitous tacit knowledge" of the general population and "specialist tacit knowledge" of specialists such as scientists. The latter category is further divided between "interactional expertise" and "contributory expertise." Interactional expertise is the "ability to master the language of a specialist domain in the absence of practical expertise." Peer review is among their examples. "Contributory expertise" is "what you need to do an activity with competence" (p. 14).

Collins, Evans, and Gorman (2007) addressed relationships between interactional expertise and Galison's (1997) "trading zones" to examine how scientists collaborate across areas of specialist knowledge. While their model is complex and multifaceted, the general idea relevant here is that scientists trade knowledge about theory, models, tools, and other forms of expertise to achieve common ground. Trading conditions involve degrees of collaboration and coercion, homogeneity and heterogeneity, and other features.

These distinctions between the types of expertise required to understand, or to perform, certain scientific tasks help to explain what kinds of knowledge are needed to reuse data. Leonelli (2013), for example, identified scientific knowledge specific to plant biology that were necessary to integrate data.

Notions of descriptive, procedural, tacit, and other forms of knowledge and expertise have a complex history that spans centuries and disciplines. Schmidt (2012), in an extensive review of tacit knowledge in science, found more than 100,000 references, spanning "Philosophy of Science, Sociology of Science, Theology, Philosophy of Sociology, Knowledge Management, Organization Theory, CSCW (Computer Supported Cooperative Work), and so forth" (p. 164). Thus, tacit knowledge is a nuanced concept with respect to the reuse of data.

Prospective data reusers often fill gaps in their knowledge by requesting the help of the data producers to reuse the data, and may credit them as coauthors in return (Pasquetto, 2018; Wallis, Rolando, & Borgman, 2013). Questions about the influence of knowledge gaps on decision making appear in economics, psychology, statistics, information science, and many other fields (Kahnemann, Slovic, & Tversky, 1982; Newell & Simon, 1972; Paisley, 1980). Differential degrees of information affect the ability of each party to interpret a problem. In statistics, for example, the concept of "uncongeniality" arises, wherein "the analyst and the imputer have access to different amounts and sources of information, and have different assessments (e.g., explicit model, implicit judgements) about both responses and nonresponses" (Meng, 1994, p. 539).

Knowledge infrastructures. Neither expertise nor data exist in a vacuum. Data practices are best understood within the rubric of knowledge infrastructures, which are "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds" (Edwards, 2010, p. 17). They are living systems influenced by complex sociotechnical factors. While data are fundamental parts of the research process, they are difficult to extract as products

to be shared. Exchanging data between individuals and laboratories usually requires labor, expertise, and expense beyond the conduct of the research per se. The ability to create, process, and exchange datasets depends not only on scientific expertise, but on infrastructure to discover, retrieve, interpret, and use them (Borgman, 2015; Bowker, 2005; Edwards et al., 2013; Karasti & Blomberg, 2017).

Tensions over ownership and control of data also influence data sharing and reuse. Community norms vary by domain, method, and types of data. Kohler (1994), building upon Thompson (1971), argued that early 20th-century drosophila biologists operated in a "moral economy" of openness, in which scientists openly shared their data, fly stocks, and tools with other laboratories in the same domain (p. 12). For these scientists, openness was a response to practical needs when labs were producing more data than they could analyze. Later, Kelty (2012) argued that this community was, at the same time, both open and closed. It was not open "to just anyone"; rather, members had unrestricted access to others' data on the condition that they first share their own data. Today, new kinds of moral economies are emerging in science (Mirowski, 2018). Openness requires governance, whether for common grazing areas or data repositories, lest "free riders" undermine community norms (Hess & Ostrom, 2007). Those who reuse data without giving adequate credit to the original creators of data are viewed as "data parasites" in some circles (Longo & Drazen, 2016).

3. Research Design

Studies of data sharing and reuse practices over the last two decades have informed science policy and the design of knowledge infrastructures. We have explored how data reuse varies by factors such as scientific domain, scale, heterogeneity, temporality, research design, goals of reuse, and levels of data processing. These are qualitative studies that explore the processes by which scientists are able to reuse others' data and the obstacles they encounter along the way. Among the domains we have studied are sensor networks, environmental sciences, ecology, biology, seismology, astronomy, earth sciences, and biomedicine. Due to our own investments in data stewardship, we have deep troves of evidence available for meta-analyses. By comparing data reuse practices across contrasting arrays of disciplines, in different time periods, we generalize our findings and propose a new theoretical framework for data practices.

Our overarching research question asks what kinds of data reuse are made possible by access to public archives of scientific data and what kinds are not? Three sub-questions address scientific practice, policy, and data science concerns:

> RQ1: Where do scientists find reusable data?
>
> RQ2: How do scientists reuse others' data?
>
> RQ3: How do scientists interpret others' data?

3.1. Research Sites

We conducted a meta-analysis of data from multiple studies of two scientific consortia, both of which were large-scale, long-term, distributed, multi-sited, data-intensive, and interdisciplinary. Studies included in the meta-analysis were led by the same principal investigator, using the same general research design (Borgman, Wofford, Golshan, Darch, & Scroggins, 2019). Each individual study varied in research questions, sources of grant funding, and personnel, involving graduate students, postdoctoral fellows, and collaborators from other universities.

Center for Embedded Networked Sensing. During its decade (2002–2012) as a U.S. National Science Foundation Science and Technology Center, the Center for Embedded Networked Sensing (CENS) open science requirements were minimal and few data archives existed in the scientific or technical areas of their research. CENS facilitated multidisciplinary collaborations among faculty, staff, and students of five partner universities (UCLA, University of Southern California, Caltech, UC-Merced, and UC-Riverside) to conduct research on developing and implementing innovative wireless sensor systems. More than 300 individuals were associated with the center over the course of its operations, drawn from computer science, engineering, robotics, geophysics, seismology, environmental sciences, oceanography, ecology, biology, design and media arts, education, medicine and health sciences, information studies, and other areas.

Much of CENS research was exploratory, resulting in data that were diverse in character and small in volume. Teams went into the field with research questions about particular phenomena and returned to their laboratories to test or to generate hypotheses. Some researchers modeled systems and others used models of phenomena to design their data collection methods. The overarching goal of these collaborations

between technology researchers and scientists was to develop new instruments, new methods, and new measures.

CENS was a productive endeavor for studying data practices due to the array of interdisciplinary collaborations, geographic distribution of field sites, diversity of data production, and continuity over the course of a decade. As a founding co-investigator of CENS, Borgman and her team conducted data practices research throughout the life of the center. We studied knowledge production in interdisciplinary contexts, data sharing, stewardship, and access to information. The challenges of reusing research data created by others came to the fore (Borgman, Wallis, & Enyedy, 2007; Borgman, Wallis, & Mayernik, 2012; Borgman, Wallis, Mayernik, & Pepe, 2007; Mayernik, Wallis, & Borgman, 2013; Wallis, Borgman, Mayernik, & Pepe, 2008).

Sensors were unreliable, especially in the early stages of the center's research program, and trust in the data was a major concern (Hamilton et al., 2007; Wallis et al., 2007). CENS researchers produced fewer datasets than anticipated, and encountered myriad difficulties managing, sharing, and reanalyzing those data (Mayernik, Wallis, Borgman, & Pepe, 2007; Pepe, Borgman, Wallis, & Mayernik, 2007; Wallis, Mayernik, Borgman, & Pepe, 2010).

DataFace Consortium. The DataFace Consortium (a pseudonym for privacy reasons) was funded by the U.S. National Institutes of Health in 2009 to advance knowledge on development, diagnoses, and treatment of craniofacial syndromes in humans. The consortium concluded its second 5-year grant phase in 2019. Participants collect, process, and deposit diverse sets of craniofacial biomedical data in a public repository that was developed and is managed by the consortium. DataFace scientists collect highly heterogeneous datasets such as 3D facial images, anthropometrics, gene expression data, ChIP-seq, RNA-seq, and animal and human tissues. By the completion of our study of DataFace in 2018, about 100 participants had released a total of more than 700 research datasets.

DataFace investigators spanned molecular and developmental biology, computational biology, genomics, clinical genetics, medicine, bioinformatics, dentistry, plastic surgery, and computer science. Members were geographically distributed across nine academic laboratories, one national lab, and three international labs. Some labs were concerned with diagnosing and preventing rare craniofacial syndromes in humans, others studied evolutionary processes of facial variation in humans, chimpanzees, mice, and zebrafish.

3.2. Research Methods

The research methods for this article are two-fold: methods by which we conducted individual studies from 2002 to 2018, and methods by which we conducted the meta-analysis reported here. Throughout this body of research, our investigations focus on data reuse as a process, for which we applied multiple qualitative methods. Each study began with ethnographic observation to understand what researchers were doing with data. After one to two years of observation, we understood these processes well enough to design interview protocols. Our interviews, which were tailored to each study and sample, provided evidence on what researchers said they do with data. Lastly, we analyzed publications, reports, protocols, websites, and other documentation to identify what they actually reported doing during their research processes. All three methods were iterated throughout the studies, providing contrasts between what we observed scientists doing, what scientists told us they do, and how these scientists reported their work. Data were analyzed using grounded theory, where we tested hypotheses iteratively in our own datasets (Strauss & Corbin, 1998).

Our data resources for CENS consist of notes from 10 years (2002–2012) of frequent observation, several waves of interviews (stored as audio files and as textual transcriptions), and a deep trove of documents. Our data resources for DataFace similarly consist of observation records, interviews, and documentary analyses gathered from 2015 to 2018. These are human subjects studies certified by the Institutional Review Board of UCLA. Details of these research methods are presented in supplemental materials and in prior publications about CENS and DataFace, cited therein.

4. Findings

Our findings are organized by the three research questions, each subdivided by the two consortia. We report the outcomes of our meta-analysis, illustrating findings with quotations from interviews. This body of evidence lays the foundation for theoretical development in the discussion section.

4.1. RQ1: Where Do Scientists Find Reusable Data?

Scientists in both CENS and DataFace sought data from sources beyond their own laboratories and their current collaborations. Given the different time frames, fewer data public sources were available for CENS research than for DataFace. Data sources varied considerably by research domain and intended reuses.

Center for Embedded Networked Sensing. The majority of CENS researchers were in computer science and engineering; the remainder worked in scientific application areas of sensor research such as geophysics, biology, oceanography, ecology, and environmental sciences. Many researchers straddled science and technology, such as those in environmental engineering and seismology. Collecting one's own data was the norm across these disparate CENS communities. Few of their research domains, at least in that time period, required data sharing or deposit, thus few open datasets were available for reuse. Exceptions were seismology, which has a shared repository (IRIS, 2018), and some genomic data collected for biology and oceanography research that were contributed to genomic databases.

As these were early days of open science, data sharing was not yet a common practice in these communities. Researchers expressed a variety of conditions for sharing data, the most common of which was that they retain first rights to publish and that they receive attribution for the data. Reuse practices varied by discipline, methods, infrastructure, and a variety of other factors (Wallis et al., 2013).

Fewer than half of the researchers interviewed for the CENS studies mentioned a specific collection from which they obtained data for purposes of reuse (Wallis et al., 2013). Most of these collections were environmental observatories or regional repositories containing records on irrigation, ocean water conditions, tide charts, solar radiation, and other metrics; others contained photos, images, or bird sounds. They also deposited data in some of these collections, such as Incorporated Research Institutions for Seismology (IRIS) for seismology, SourceForge for software code, and the University of California James Reserve, which was a CENS partner. Collections existed only for a small portion of the many domains covered by CENS, hence, searching options for relevant data were few. These data sources spanned the continuum from research collections of individual projects, resource collections with more curation, and established reference collections, such as the U.S. Geological

Survey, National Oceanic and Atmospheres Administration, and NASA. The data most commonly mentioned were those of established reference collections.

Many researchers in the CENS community received requests for their data, thus, interpersonal exchange was an important vector. As our CENS studies addressed data release and sharing, we asked specific questions about when and how they responded to requests for reuse of their own data. Researchers mentioned a number of occasions when others had requested data for reuse. In some cases, requestors were known colleagues and in others they were unknown researchers who had identified datasets through publications, talks, or other means.

DataFace Consortium. DataFace scientists reported using open data from online repositories every day as an integral part of their research workflow. Contrary to the situation of CENS researchers, DataFace researchers had access to a wide array of online repositories and bioinformatics tools, which vary by specialization. Among those resources mentioned were repositories of human and animal sequence data from the National Center for Biotechnology Information, GenBank, Exome Aggregation Consortium (ExAC) browser, UCSC Genome Browser ('the genome browser'), Online Mendelian Inheritance in Man (OMIM) (a catalog of human genes and genetic disorders and traits), ClinVar (lists the relationships among human variations and phenotypes), 1000 Genomes (a catalogue of human variation between ethnic populations), and PubMed (a search engine of references and abstracts on life sciences and biomedical topics).

Like CENS, DataFace researchers reported receiving requests for their data, and asking others for their data at least a few times in their career.

4.2. RQ2: How Do Scientists Reuse Others' Data?

The second question addresses the purposes for which scientists use data acquired from others. We asked how data reuse varies by research domain, purposes for potential reuse, access to data creators, and time period. Our goal is to develop a typology of data reuses, the conditions under which these reuses occur, and relationships between data sources and reuses.

Center for Embedded Networked Sensing. CENS researchers sought external data to learn about a field site, such as current and historical weather conditions, to

determine trends of temperature, water, sunlight, and other factors relevant to planning new data collections. Often, they would combine literature and data reviews to gather information such as measurements of birds, microclimates, and soil conditions. CENS scientists referred to these processes as "ground truthing," a method to verify and compare real-world conditions for their empirical or experimental data collection (Borgman et al., 2012).

Our meta-analysis revealed myriad uses of external data for ground-truthing that crossed the boundaries of scientific domains. Biologists used recordings of bird sounds acquired from a data archive to calibrate instruments that detected the presence of those species. This team also used these recordings to attract birds of those species so they could be observed in situ. In another case, environmental engineering researchers used near-real-time data from the California Digital Exchange Center on river conditions to determine where to place sensors most safely and effectively. One CENS team employed a dataset on military truck movements to test localization algorithms for predicting the routes of marmots and woodpeckers. An electrical engineering group with experience in weapons guidance systems reused those algorithms to detect harmful algal blooms. In yet another example, a team acquired GIS data from weather, traffic, and fast-food restaurants to test sensor algorithms (Wallis et al., 2013).

Comparative uses of external data could also reveal errors or faulty detection systems, such as whether chemical concentrations were within an expected range. One researcher in 2006 explained how he compared measurements from his standard "trusted" method to those of a CENS nitrate sensor: "Historically and currently there's no readings getting a third of what it was by using this other trusted method…So that was kind of my first indication that 'Hey, this is probably not doing what it's supposed to be doing.'" We found ground-truthing to be an innovative, idiosyncratic, and scientifically effective practice of data reuse, but hard to generalize.

While most CENS participants reported using others' data for ground truthing, only a few mentioned reusing others' data for analysis, such as testing new hypotheses through statistical analysis. These latter cases usually involved meta-analyses. For example, researchers in the five Mediterranean climates combined raw datasets on specific species or variables to compare findings in California with those of other Mediterranean climates. As explained by this CENS scientist in 2006, "Data sets are getting shared in a big way, although at a high level, and at a very detailed level. ...

One complaint about ecology is that there is no grand unified theory. It's hard to get general principles out of it."

DataFace Consortium. Pasquetto (2018), in examining socio-technical, epistemic, and ethical challenges of making biomedical research data openly available and reusable, found that data reuse in DataFace remained a complex, delicate, and often time-consuming process. Patterns of data reuse among DataFace researchers were similar to those of CENS, despite differences in scientific domains, methods, and time periods. DataFace researchers often reused others' data for activities comparable to ground truthing, while relying almost entirely on their own data for analysis and hypothesis testing. Typically, DataFace scientists would reuse data acquired from online repositories for comparison and control. As explained by this DataFace bioinformatician in 2016, "When we map our data back on to the human genome or mouse genome, being able to visualize it on [the Genome Browser] is really valuable…PubMed just for literature, search…OMIM [when] trying to figure out gene function or genetic basis of various human diseases." In another lab, another biologist explained, still in 2016: "[For] a margin of interest, we found mutation in the gene and we want to look at the expression (to see) if there is any animal model for the gene… the database basically makes it easier for you to see if…anything…was done before, so you don't repeat."

DataFace researchers relied on data from open archives to compare, interpret, and summarize statistical findings about biological functions of certain genetic variations. When significant associations between genotypes and phenotypes were found, researchers compared their preliminary findings with other information known about the genetic markers and their biological functions. A postdoctoral researcher in genomics interviewed in 2017, for example, reviewed data and literature for 358 genes previously associated with the genetic markers identified by the team. The OMIM database was used to investigate whether the identified genes or single nucleotide polymorphisms (SNPs) variants are associated with syndromes that affect the formation of mouse or human faces. The VISTA Enhancer browser was used to check whether genes are located in 'enhancers.' The DECIPHER database, an online clinical database that contains human genome variants and phenotypes of thousands of patients worldwide, provided other clues.

Like CENS researchers, the DataFace community rarely reused data collected by others for analysis. Researchers mentioned new analyses conducted on others' data as anecdotes. One example of a secondary analysis conducted on existing data involved a

Genome Wide Association Study (GWAS) (SNPs, 3D facial images, etc.) craniofacial dataset that contained phenotypical and genotypical information on the variations of craniofacial features among a certain population. The dataset was originally collected by one of the DataFace labs to investigate how genetic predictors for craniofacial syndromes (e.g., orofacial clefting) might be different within and across populations. A second lab specialized in computer science and physical anthropology repurposed this dataset to test whether a machine-learning algorithm for spatial clustering would be able to predict facial features from genetic markers—a process called "DNA facial phenotyping." This new effort combined the GWAS dataset collected by the DataFace lab with four other similar data sources. The analysis conducted on the combined genomic and phenotypic datasets was completely new and it employed software pipelines not previously tested on these four datasets to extract new information.

4.3. RQ3: How Do Scientists Interpret Others' Data?

The third, and most complex, question examines the conditions under which available documentation is sufficient for scientists to reuse data, when scientists pursue interpersonal contact for further expertise about those data, the processes by which they interpret others' data, and the types of knowledge they employ for interpretation.

Our findings for the first two research questions demonstrate that researchers in both CENS and DataFace sought external data for use in their research and that they employed those data in myriad ways. The primary distinctions between comparative reuses and integrative reuses, such as conducting new analyses, are the greater difficulties in interpreting others' data and the knowledge required to do so.

Center for Embedded Networked Sensing. A typical research scenario in CENS was one in which scientists and technology researchers jointly developed and deployed a wireless sensor network in an environment that a scientific team wished to observe. Both the sensor network and the environment were studied—the sensor for its effectiveness and ability to collect accurate data, and the environment for trends and patterns that could be found in data collected by the sensors. Science and technology teams worked in the field together, spending a day, several days, or several weeks in ecological reserves, public lands, or private lands such as farmers' fields. Most of these sites were in California, but some were part of international collaborations, such as

large-scale seismic installations in Mexico and Peru, or ecological habitat reserves in Costa Rica.

Despite working closely together, and pursuing interdependent research questions, each of the CENS teams tended to collect, store, manage, and interpret their data independently. As shown in Figure 1, each team needed the measurements from the sensor network (center circle), but each team compared those measurements to others from their own equipment or methods. Biologists (hand-collected application data, bottom) dipped water samples from the lake, stream, or river; centrifuged the specimens; adjusted the pH; and took their own values of nitrates and other variables. Roboticists (sensor-collected proprioceptive data, top left) observed how their own devices could locate nitrate concentrations, algae, or other features based on the sensor-collected application data. Electrical engineers concurrently were testing the health of the sensor network (top right), using information-theoretic algorithms. Despite this interdependence, the teams made little attempt to interpret each others' datasets. Rather, they returned to their labs with their own datasets, and published their findings in journals of their own research domains (Borgman, Wallis, & Enyedy, 2007; Pepe, 2010).

Later research in CENS affirmed the complex relationships between data collected collaboratively by multiple teams and the uses those teams made of these data. CENS teams could use datasets from other teams for comparative or "background" uses, but generally preferred to collect their own data to conduct new inquiries, deemed "foreground" uses of data (Wallis et al., 2013).

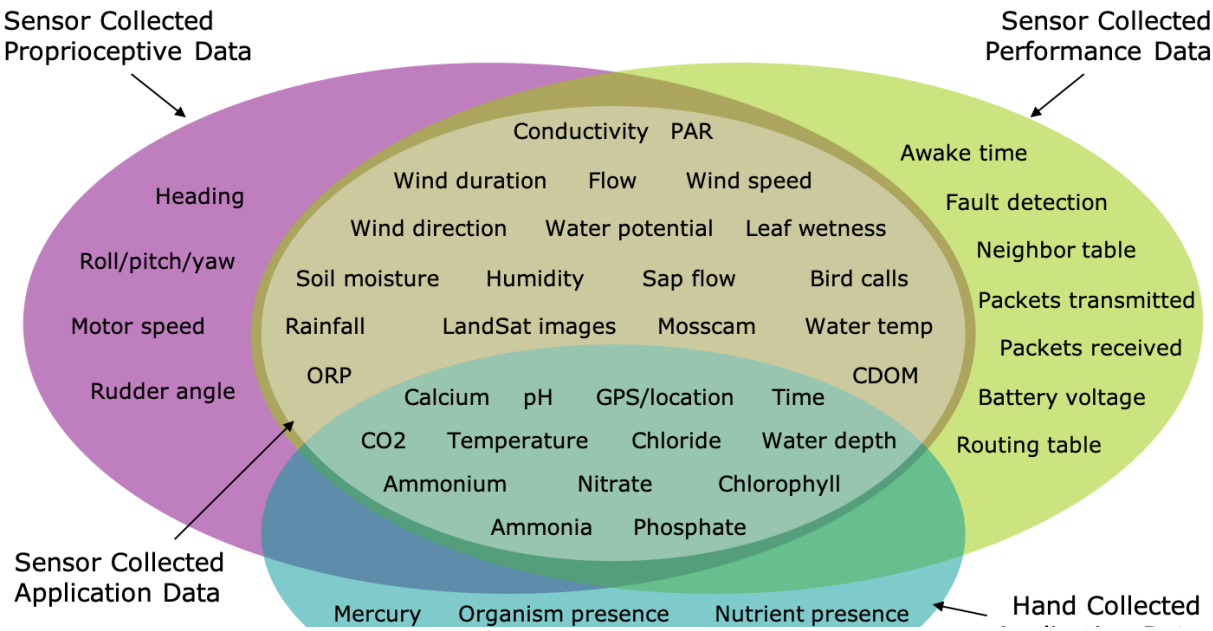

Figure 1. Data variation in CENS (Borgman, Wallis, & Enyedy, 2007).

Another striking example of the difficulty of interpreting others' data, even when researchers are collaborating on complementary problems, arose in the competing ways that CENS teams interpreted the concept of temperature. The measurement of temperature appeared throughout CENS research in biology, environment, ecology, health, and other areas because temperature gradients influence so many other variables. Through extensive observation of teams working together, followed by individual interviews in private settings, we came to realize that researchers conceptualized temperature in very different ways, and measured the variable accordingly. Teams often were unaware of these differences or the effects of measurement parameters on the work of other teams in CENS.

A simple example of the differences in interpretation is the succinct statement of an engineering researcher in 2006 that "temperature is temperature." His concern, from an engineering perspective, is whether measurements of temperature from the sensor network are consistent. A CENS biologist, however, told us in 2006 that*:*

> There are hundreds of ways to measure temperature. "The temperature is 98" is low-value compared to, "the temperature of the surface, measured by the infrared thermopile, model number XYZ, is 98." That means it is measuring a proxy for a temperature, rather than being in contact with a probe, and it is measuring from a distance. The accuracy is plus or minus .05 of a degree. I [also] want to know that it was taken outside versus inside a controlled environment, how long it had been in place, and the last time it was calibrated, which might tell me whether it has drifted.

The biologist's concern is whether the instrument meets the accuracy standards that are necessary to publish in his domain's journals. Despite the best intentions of his engineering partner, the sensor network measurements of temperature were useless to the biologist until (and unless) the instruments could be calibrated to biology standards. For biology research purposes, the new instrument had to be installed and operational next to his own instruments for a continuous 365 days to be considered trustworthy (Borgman et al., 2012).

CENS researchers also commented on the labor that would be required to assist others in interpreting their data sufficiently for reuse. An ecologist said in 2006, for example: "Oh gosh, it would be substantial [labor for us to work with the reusers]. I

think a lot of hand holding [would be needed] until people got used to it [the data]. Maybe third or fourth-time users would probably start to get a feel for it, but first-time users you're going to probably be answering two or three emails a day from this person." Another researcher, considering whether to reuse data collected by others, said "The data are out there for people to use, but you really probably want to talk to that person who collected it before you just use it. … for ecological data, I think that oftentimes there is a whole huge context—the way the data are collected" (2006).

Central to the interpretation challenges is the ability to know how, how much, and why data were processed. Only the data creators know all the minute details of data handling, some of which are difficult or impossible to document, as a CENS graduate student told us in 2012: "I know the site, I installed the sensors and I have looked at the data for a very long time, so if I want to do another paper I don't have any problem. … It's very hard to share those … because most of them will just be in a notebook and even if the metadata tells you the basic things of what it was installed and what was done, all of the little things [are] really, really, really hard to share. So usually people will not share the raw data, they will just share with you the processed data, which to get to that product you need to do all the processing and all the cleaning, and everything else that was involved with this information that is usually not shared."

Among the most revealing comments about data interpretation came from a CENS statistician in 2006:

> It's certainly part of how I've been trained that you can be a statistician and never look at data, but to me that's not very fun. So you've got to go talk to somebody. You've got to go collaborate... The whole point of me going on this deployment is I can now see and ask questions and do all that stuff that I end up doing after the fact.

DataFace Consortium. The goal of the DataFace consortium was to collect and release high-quality, structured data for reuse by the community. Despite their devotion to this goal, researchers preferred collaborative work for integrative reuse of data. Given this puzzling finding, we conducted a second round of interviews to investigate the circumstances in which they needed collaboration to run new analyses on others' data. This second set of interviews and observational evidence allowed us to gain a better understanding of what constitutes the little things that are hard to share, and conditions when one has to talk to somebody to reuse their data.

DataFace researchers found collaborative work particularly useful when they needed to acquire new and specialized skills, expertise, or knowledge to reuse data. When asked why they preferred collaborative reuse instead of independent reuse, a typical response was:

> Collaboration is better because it makes sure that everybody's on the same page and they know what's going on with the data. ... The groups have complementary skills ... for human data, we have our own GWAS and we participated in a different GWAS ... we could just download the data from dbGaP but our preferred method is just to collaborate with the group that did that GWAS and … then we share the results and we help each other on different analyses and we're always talking about who's doing which things and where are the priorities for different groups (2016).

Collaboration between the two labs was crucial to accomplishing the GWAS reanalysis. Members of the first lab cleaned the original dataset to make it reusable for the new research design and for searching the literature to identify additional variance. The second lab computed the analysis of the dataset and interpreted the statistical findings. The two labs coauthored the resulting articles, which were published in a prominent journal.

In these cross-disciplinary or cross-specialty situations, DataFace researchers preferred (although not necessarily required) collaboration even when data are structured, properly curated, interoperable, and of high quality. Typically, data reusers would ask data creators either to run entire new analyses for them (and then send them the refined results), or to provide them with a dataset that has been processed in a specific format for the intended analysis. Prospective data reusers reported that they often lacked time and resources necessary to become sufficiently familiar with the data to run new analyses themselves. Asking data creators for analytical assistance was more effective.

We found numerous examples in DataFace of how data interpretation required knowledge of the theoretic framework for the study, how experiments were conducted, how data were processed, and how signals were separated from noise. To collect and analyze their experimental data, these teams accessed specialized knowledge that they had accumulated over time. For a lab studying orofacial clefting in mouse embryos, for example, the teams' expertise can be clustered into three categories. First is scientific domain knowledge, which includes the biology of facial development, biology of the model organism (mouse), and data types (tissues, sequences, etc.), and the associated

theoretical framework. Second is research methods, including research design and methods for data collection (microarray). Third is the tools and technologies, including data processing and quality control workflow, databases for annotating data based on previous data, and computational packages and tools for statistical data analysis (Python, R, Jupyter notebooks). As one of the team members explained:

> If people want to go into our data and do the type of analysis that we want to do, well, they can do that too. I think they would be at a disadvantage though, because they do not know … all the meta-analysis that's associated with it. They weren't involved necessarily in the design and the execution. So I think it's more difficult for people to get in there and make sense of this.

5. Discussion

Researchers in both CENS and DataFace reused data created by other researchers, by government agencies, and by other trusted sources. The most common reuses of data were for comparative purposes such as ground truthing, calibration, and identifying baseline measurements of phenomena. Less commonly, they reused other's data for integrative purposes to ask new questions or conduct new analyses.[1] Here we develop the theoretical framework for types of data reuses and the data creators' advantage.

5.1. Comparative and Integrative Reuses of Research Data

In our initial studies of CENS, we identified a dichotomy of background and foreground reuses of data (Wallis et al., 2013). Background data are "those that are important to research activities but that are not necessarily reported in publications nor kept for future use or reuse." These include ground-truthing data from public sources and comparative uses of data from collaborating CENS teams, as illustrated in Figure 1. Foreground data, in contrast, were the focus of a CENS research project, whether a field deployment or laboratory study. Researchers referred to these forms of data as "core" or "primary" data. We tested our hypothesis of the background–foreground dichotomy of data reuse in the DataFace Consortium. Based on our meta-analysis of the CENS and DataFace studies, we propose a continuum of data reuses from comparative to integrative, which better describes the types of reuse and reflects the

variance in activities that fall between these end points. Table 1 summarizes our theoretical framework for the comparative–integrative continuum.

Table 1. Comparative–Integrative Data Reuse Continuum

	Comparative Data Reuse ⟷	**Integrative Data Reuse**
Goal	'Ground truthing': calibrate, compare, confirm	Analysis: identify patterns, correlations, causal relationships
Example	Instrument calibration, sequence annotation, review summary-level data	Meta-analyses, novel statistical analyses
Frequency	Frequent, routine practice	Rare, emergent practice
Interpretation	Interactional expertise, 'knowledge that'	Contributory expertise, 'knowledge how,' tacit knowledge

Comparative Reuses of Research Data. At the comparative end of the continuum, researchers reuse others' data to assess similarities and differences for purposes such as ground-truthing, calibration, and experimental controls. Observations of earth, oceans, and atmosphere are readily amenable to ground-truthing, as are many kinds of biomedical data. Reference collections, such as the databases of observations provided by USGS, NOAA, NASA, and similar resources provided by the National Institutes of Health, are curated to robust community standards (U.S. National Science Board, 2005). These are trusted sources for comparison and were essential resources for the communities studied. Epistemic trust in knowledge products (Darch, 2019; Porter, 1996; Shapin, 1994) may stem from the design and stewardship of collections, to the extent that they align with community practices.

Researchers in CENS and DataFace also sought data from sources that could be classified as research or resource collections, when those best served their needs. When reusing data for comparative purposes, researchers appear to rely on interactional expertise (Collins & Evans, 2007; Collins et al., 2007) to interpret data. That is, a scientist who is an expert in the same field can understand someone else's

data for comparative purposes, given sufficient metadata, ontologies, and other documentation. In these cases, researchers are able to obtain enough information about the data, whether through documentation, familiarity, personal contacts, literature, or other means, to assess the quality and value of data for a particular reuse. Comparative reuses of data are such commonplace occurrences that datasets, or the data collections from which they were acquired, may not be cited in publications (Borgman, 2015; Wallis et al., 2013). Expert readers of scientific publications may place varying degrees of epistemic trust in samples, field sites, methods, instrumentation, and other factors necessary to assess their adequacy for particular reuses.

Integrative Reuses of Research Data. At the integrative end of the continuum of data reuses, much more extensive knowledge about the data is required for interpretation. To reanalyze datasets created by others, or to combine data from external sources, CENS and DataFace researchers appeared to draw upon contributory expertise to interpret data (Collins & Evans, 2007). Contributory expertise is the ability to perform the action, such as reusing others' data in a new experiment. To perform a scientific action such as a laboratory technique requires training and experience. These experts develop tacit knowledge that cannot be fully documented in metadata, ontologies, and other "small facts" (Leonelli, 2010). Some of that tacit knowledge involves artisanal expertise, embodied knowledge (Ryle, 1949), or "magic hands" (Hilgartner & Brandt-Rauf, 1994).

Integrative reuses of data appear to require greater levels of expertise in the scientific specialty, including theory, models, methods, tools, and technologies associated with the dataset, and deeper levels of epistemic trust in the knowledge products and those who created them. Our findings that researchers in both CENS and DataFace preferred to collaborate with data creators for purposes of integrative data reuse support this theory. These researchers have learned through experience that interpreting data for reanalysis requires more knowledge of the context and purposes for which those data were created than is available through public documentation.

Data Reuse along the Comparative-Integrative Continuum. In between the extremes of comparative and integrative reuses of data are practices that require varying degrees of knowledge about the data and the infrastructure for interpretation. Types and degrees of expertise and trust required for data reuse may vary considerably by specialty. In the DataFace example of interpreting data about oroclefting in mouse embryos, knowledge about theory and method are distributed

among multiple individuals, each with different skill sets. In plant biology, Leonelli (2010) identified field-specific types of data integration that are hard to generalize, but lead to new knowledge: "inter-level integration and model organism research," "cross-species integration and biofuels research," and "translational integration and plant-pathogen interaction."

In some situations, data reuse may be accomplished by brief exchanges of information between creator and reuser. The reused datasets may be cited in publications and their creators acknowledged. In other situations, researchers may conduct reanalyses on their own by investing extensive labor to replicate as much context as they can (Wallis et al., 2013). Time estimates vary widely; some scientists in our studies have suggested that a year or more of effort might be required to reuse data without direct collaboration (Borgman, 2015).

5.2. Sources and Characteristics of Reusable Data

In principle, any scientific dataset might be reused for unknown purposes by unknown individuals in unknown domains at unknown times in the future. However, our findings suggest that some kinds of data are more readily reusable than others. Data created by individuals and teams for specific research projects are necessarily more difficult to interpret and reuse than those intentionally created for comparative purposes, such as the reference collections popular in CENS and DataFace.

While investigators can standardize the form of their datasets for deposit in an archive, the full array of knowledge about the dataset, its context, and relationships to other information is likely to remain local knowledge. Depending on the size of the creating team, that knowledge may be embodied in one person or distributed across many people. Another factor in reusability is the maturity of the technique. In the early stages of development, one person may have the "magic hands" necessary to perform the technique; if the method becomes common practice, it may be commodified into "kits" that are readily transferrable within the scientific specialty (Hilgartner & Brandt-Rauf, 1994).

The scale of the community is another consideration in the likelihood of reuse for any given dataset (Borgman, Darch, Sands, Wallis, & Traweek, 2014; Borgman et al., 2016). Smaller communities such as CENS, particularly those conducting exploratory research with new methods, have fewer public data sources on which to draw.

Similarly, the community that might wish to reuse those datasets is relatively small. Conversely, DataFace is part of the biomedical community, which functions on a much larger scale, and has access to far more reference collections. Even so, the size of specialty communities varies widely within the biomedical sciences.

Open access policies also influence the reuse of data. When CENS began in 2002, few of their research domains were subject to data-sharing requirements and relatively few data collections were available. Even so, CENS researchers did reuse data from available collections for comparative purposes, and did seek data from other researchers for integrative purposes (Wallis et al., 2013). DataFace began in 2009 as a consortium dedicated to building a common data collection, in biomedical fields with a long history of data-sharing requirements and with numerous public collections available.

5.3. The Data Creators' Advantage

Ontologies, metadata, documentation, and other forms of curation all aid in transferring knowledge between contexts. Curation is necessary, but rarely sufficient, for integrative data reuse. In the empirical sciences included in our meta-analysis, the individuals and teams who create datasets retain an inherent advantage in interpretation over any other prospective data reuser who may come along later. One reason is that data creators possess tacit knowledge about the context and purposes for which those data were collected. Removing data from their original context necessarily involves information loss. In the social studies of science, this problem is known as "making data mobile" across contexts (Bowker, 2005; Latour, 1987).

Researchers in CENS and DataFace described numerous examples of small details that were nearly impossible to convey to researchers who had not participated in the data creation. We also observed many situations where researchers made minor adjustments to sensors, devices, or software that were not recorded, any of which can change interpretation considerably. These details involved research design, methods, anomalies of field and laboratory sites, calibrations, fragile technologies, failing batteries, equipment sensitivities, and environmental and atmospheric conditions. Some anomalies can be known, such as recorded electrical failures, tinfoil hats that were folded around sensors to add shade or reflection, or calibrations that were tweaked between observing runs. Others are only suspected, such as a sensor that may have been kicked by a cow, a laptop signal disrupted by a cellphone, or a lab experiment that may have been influenced by odors on a visitor's clothes.

Data are processed, cleaned, and reduced according to the best practices of a domain. Data creators incorporate these many decisions into the interpretations they report in publications. However, detailed choices about removing outliers from observing runs or customizing sensors on the fly may be omitted from journal articles due to space limitations, or simply because they are standard practices within a specialty. Similarly, the "uncongeniality" problem in statistics is that data creators and data reusers have different interpretations due to different amounts and sources of information about a dataset (Meng, 1994, p. 539).

A second reason for the data creators' advantage is that they master the cumulative, multifaceted, and collaborative expertise needed to analyze a specific dataset for a specific purpose. Even when metadata, ontologies, and other documentation meet the best quality standards for a domain, that contextual information may not be sufficient for integrative data reuse. Data creators develop expertise through experience in designing experiments, participating in laboratory and field exercises, selecting data collection protocols, learning to use instruments, writing software, and through data analysis.

5.4. Knowledge Infrastructures in Practice

The data creators' advantage rests on knowledge and experience with a dataset, which in turn rests on access to resources, tools, communities, and institutional arrangements. Epistemic trust in knowledge products such as datasets also involves the ability to assess their value for a particular reuse. Members of a research specialty, who have similar expertise and similar access to the knowledge infrastructure of a community, are in the best position to make those value judgments.

Collaboration for data reuse is often a function of community relationships. For integrative purposes, CENS and DataFace researchers preferred to analyze others' data in the context of a collaboration. Working together saves time and improves accuracy. When mutual benefits accrue, data are more likely to be shared and reused.

As good citizens of their communities, CENS and DataFace researchers responded to most requests for information when asked. However, they lacked the resources to provide service for all of their prior datasets, for indefinite periods of time. Members of a research community who have common interactional expertise, and who are known to each other, appear most likely to share data with each other.

Knowledge infrastructures are dynamic ecosystems that have temporal characteristics. Data reusers can contact data creators while they are alive and have access to their data and resources. That time window can be short, as graduate students and postdoctoral fellows who collected data may depart within months or a few years. When migrated to new hardware, software, and servers, data may be lost, difficult to find, corrupted, or no longer readable by new tools (Borgman, 2015; Jackson, Ribes, Buyuktur, & Bowker, 2011).

Shared data collections are a key component of knowledge infrastructure. CENS never developed a data repository, due to lack of requirements, lack of interest, and the disparate types of data collected (Mandell, 2012; Wallis et al., 2010). Individual investigators in CENS were responsible for managing their own data, per the requirements of their own fields. CENS' documented legacy exists as a collection in the University of California eScholarship repository (2011), which currently contains 671 papers, posters, and other CENS products. More than half of these are posters, which proliferated at frequent research community events, and often contain detailed data, diagrams, photographs, and other images. DataFace, in contrast, has a shared repository that contains contributed datasets, and will remain available at least for the duration of project funding. Access to datasets created by CENS and DataFace will decline inexorably, as investigators cease to maintain local copies and physical specimens, as computers and scientific instruments are replaced, as students graduate, postdoctoral fellows and staff change jobs, and as others retire.

6. Conclusions

"The value of data lies in their use" is the premise claimed by the National Research Council (1997, p. 10) for open science. For data to be used, they must be usable and useful, which requires investments in knowledge infrastructures. Many kinds of data reuse can be accomplished with publicly available data, especially for purposes of comparison. To reuse data for integrative purposes, such as combining data from external sources to ask new questions, requires more knowledge of the context of the original data production and deeper expertise in the scientific domain. Therein lies the data creators' advantage.

6.1. Data Reuse Is a Process

Reuse is not a binary variable; thus, counting the frequency of dataset downloads from public archives is a misleading metric. Just as scholars read far more literature than they cite in their publications, these researchers survey more datasets than they reuse. When datasets are reused for comparison or ground-truthing, authors may not cite those datasets or the collections from which they were acquired. As a consequence, the knowledge infrastructures that support these research communities may be invisible and their value underappreciated.

6.2. Data Reuse Requires Expertise and Trust

Trust in the value of a dataset may depend upon the ability to assess the integrity of individuals, institutions, methods, and contexts of creation. Metadata, ontologies, and other forms of curation promote epistemic trust in knowledge products. To assess whether a dataset is reusable, researchers draw upon their tacit knowledge about the scientific domain. For comparative reuses, interactional expertise may suffice, as experts can glean enough knowledge about the dataset from available documentation. To assess whether a dataset may be reusable for integrative purposes, higher levels of epistemic trust and tacit knowledge are required. Contributory expertise entails deeper knowledge of context and greater skills in performing the tasks necessary to create the dataset. Interpretation depends upon a much deeper understanding of how those data were created.

6.3. The Data Creators' Advantage Promotes Scientific Collaboration

Data creators have intimate knowledge of their datasets that cannot be fully explained to others. They know those small things that are difficult to share. To interpret a dataset, one needs knowledge not only of what is in a dataset, but also knowledge of models, theories, hypotheses, instruments, hardware, software, techniques, and circumstances associated with its creation. Prospective data reusers seek out data creators when contributory expertise is required. Successful new collaborations result when the parties find mutual benefit. Data creators in our studies were generally helpful in providing information to prospective data reusers. However, they were necessarily selective in forming new partnerships to reanalyze their data.

Collaboration is time-consuming and resource-intensive, but worthwhile when both parties see mutual benefit for shared research agendas. For individual scientists and teams, datasets are valuable scientific assets that can be deposited, shared, and brokered to form new collaborations. Datasets also are scientific liabilities, to the extent that they need to be managed, stored, maintained, and serviced for reuse by others.

The 'reproducibility crisis' can be partially attributed to the data creators' advantage. No matter how similar the methods, laboratories, or field sites, identical circumstances and expertise cannot be achieved. Replicating a study using different methods is a more powerful outcome than reproduction, per se (McNutt, 2014; National Academies of Sciences, 2019).

In cases where datasets can be standardized sufficiently to circulate through communities, science benefits. Our findings suggest that anyone with sufficient interactional expertise can use those datasets for comparative purposes. While datasets may be fungible or substitutable objects for comparative purposes, rarely are they fungible for integrative reuse. In the CENS and DataFace communities, integrative data reuse was rare, occurring only a few times in a career.

6.4. Investments in Knowledge Infrastructures Promote Data Reuse

Scientific data reuse occurs in communities, as scientists develop "trading zones" within which they can exchange knowledge. Data curation facilitates the exchange of data within and between communities, thus expanding the range of possible reuses. However, data archives are expensive to build, requiring resources, governmental and institutional commitments, and policies that encourage or require scientists to contribute their datasets. Data stewardship is also an expensive process, involving digital collections, hardware, software, instrumentation, samples, and human expertise for curation and maintenance. These investments, which are core to open science, are necessary conditions for data to be available for reuse. Individual scientists must have means to discover, locate, retrieve, and interpret those datasets. They need knowledge and trust, but they also need network access, equipment, software, and other tools.

Knowledge infrastructures evolve over time; thus, temporal factors are a critical concern for data reuse. Datasets in well-curated public collections may be useful indefinitely for comparative purposes, but the time window for integrative reuses is

much shorter. To the extent that data reusers must rely on information available only from the data creators, those individuals must be reachable, available, and still have access to original records. Even if principal investigators are reachable, their knowledge of datasets may be limited; often, the graduate students or postdoctoral fellows who collected data in laboratories and field sites are the individuals with the intimate knowledge necessary for integrative reuses. Once separated from the university where the research was conducted, students and postdoctoral fellows usually are separated from those data and computing resources as well.

In sum, data reuse in the sciences is a complex process that depends upon trust, expertise, policy, and knowledge infrastructures. Public investments in scientific infrastructure, especially in data curation, are necessary to make data available and reusable. For comparative reuses of data, these investments may suffice to increase the circulation of data within and between scientific communities. For integrative reuses of data, however, collaboration between data creators and reusers usually is necessary. The data creators' advantage is a consequence of investments made by individuals and teams in conducting their research to high standards, with deep knowledge of theory, method, context, tools, and instrumentation. When data creators and reusers find mutual benefit in collaboration, science benefits. Our findings from a qualitative meta-analysis of nearly two decades of research on scientific data practices is extensive, but by no means comprehensive of all scientific fields. To advance open science and data science, we strongly encourage more qualitative and quantitative research into data practices in other domains.

Supplements

A full bibliography of sources for this article, including prior publications in which our initial findings were reported, is provided in the supplemental materials available at https://hdsr.mitpress.mit.edu/pub/tn4j86t1

Acknowledgments

Research reported here was supported in part by grants from the National Science Foundation (NSF) and the Alfred P. Sloan Foundation (Sloan): (1) The Center for Embedded Networked Sensing (CENS) was funded by NSF Cooperative Agreement #CCR-0120778; (2) CENSNet: An Architecture for Authentic Web-based Science

Inquiry in Middle and High School, NSF #ESI-0352572; (3) Towards a Virtual Organization for Data Cyberinfrastructure, NSF #OCI-0750529; (4) Monitoring, Modeling & Memory: Dynamics of Data and Knowledge in Scientific Cyberinfrastructures, NSF #0827322; (5) If Data Sharing is the Answer, What is the Question?, Sloan #2015-14001.

The authors thank Bernadette Boscoe, Peter Darch, Milena Golshan, Michael Scroggins, and Cheryl Thompson of the UCLA Center for Knowledge Infrastructures for their guidance in project design, data analysis, and comments on drafts. CENS interviews reported here were conducted by Jillian C. Wallis, Matthew S. Mayernik, and Christine L. Borgman, with additional data analysis by Elizabeth Rolando. Most of all, we acknowledge the generosity of CENS and DataFace personnel for welcoming us into their communities as observers and for their thoughtful discussions of data practices.

References

Bietz, M. J., & Lee, C. P. (2009). Collaboration in metagenomics: Sequence databases and the organization of scientific work. In I. Wagner, H. Tellioğlu, E. Balka, C. Simone, & L. Ciolfi (Eds.), *ECSCW 2009* (pp. 243–262). https://doi.org/10.1007/978-1-84882-854-4_15

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

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., 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., Scharnhorst, A., & Golshan, M. S. (2019). Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *Journal of the Association for Information Science and Technology*, *70*, 888–904. https://doi.org/10.1002/asi.24172

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, *7*(1–2), 17–30. https://doi.org/10.1007/s00799-007-0022-9

Borgman, C. L., Wallis, J. C., & Mayernik, M. S. (2012). Who's got the data? Interdependencies in science and technology collaborations. *Computer Supported Cooperative Work*, *21*, 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

Borgman, C. L., Wofford, M. F., Golshan, M. S., Darch, P. T., & Scroggins, M. J. (2019, In Review). Collaborative ethnography at scale: Reflections on 20 years of data integration. *Science and Technology Studies*. Retrieved from https://escholarship.org/uc/item/5bb8b1tn

Boscoe, B. M. (2019). *From blurry space to a sharper sky: Keeping twenty-three years of astronomical data alive* (PhD Dissertation, UCLA). Retrieved from https://escholarship.org/uc/item/2jv941sb

Bowker, G. C. (2005). *Memory practices in the sciences*. Cambridge, MA: MIT Press.

Center for Open Science. (2019). Center for Open Science: Openness, integrity, and reproducibility. Retrieved from http://centerforopenscience.org/

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

Collins, H. M., Evans, R., & Gorman, M. (2007). Trading zones and interactional expertise. *Studies in History and Philosophy of Science Part A*, *38*, 657–666. https://doi.org/10.1016/j.shpsa.2007.09.003

Culina, A., Crowther, T. W., Ramakers, J. J. C., Gienapp, P., & Visser, M. E. (2018). How to do meta-analysis of open datasets. *Nature Ecology & Evolution*, *2*, 1053–1056. https://doi.org/10.1038/s41559-018-0579-2

Darch, P. T. (2019). *The core of the matter: How do scientists judge the trustworthiness of physical samples?* Manuscript submitted for publication.

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

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

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., … Calvert, S. (2013). *Knowledge infrastructures: Intellectual frameworks and research challenges*. Retrieved from http://hdl.handle.net/2027.42/97552

European Commission High Level Expert Group on Scientific Data. (2010). *Riding the wave: How Europe can gain from the rising tide of scientific data* [Final report of the High Level Expert Group on Scientific Data. A submission to the European Commission]. Retrieved from https://www.fosteropenscience.eu/content/riding-wave-how-europe-can-gain-rising-tide-scientific-data

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Journal of Computer Supported Cooperative Work*, *19*(3–4), 355–375. https://doi.org/10.1007/s10606-010-9117-8

Federer, L. M. (2019). *Who, what, when, where, and why? Quantifying and understanding biomedical data reuse*. Retrieved from https://drum.lib.umd.edu/handle/1903/21991

Galison, P. (1997). *Image and logic: A material culture of microphysics*. Chicago, IL: University of Chicago Press.

Gitelman, L. (Ed.). (2013). *"Raw data" is an oxymoron*. Cambridge, MA: MIT Press.

Hamilton, M. P., Graham, E. A., Rundel, P. W., Allen, M. F., Kaiser, W., Hansen, M. H., & Estrin, D. L. (2007). New approaches in embedded networked sensing for terrestrial

ecological observatories. *Environmental Engineering Science*, *24*(2), 192–204. https://doi.org/10.1089/ees.2006.0045

Hanson, B., Sugden, A., & Alberts, B. (2011). Making data maximally available. *Science*, *331*, 649. https://doi.org/10.1126/science.1203354

Hess, C., & Ostrom, E. (2007). *Understanding knowledge as a commons: From theory to practice*. Cambridge, MA: MIT Press.

Hilgartner, S., & Brandt-Rauf, S. I. (1994). Data access, ownership, and control: Toward empirical studies of access practices. *Science Communication*, *15*, 355–372. https://doi.org/10.1177/107554709401500401

Incorporated Research Institutions for Seismology (IRIS). (2018). Retrieved August 10, 2018, from https://www.iris.edu/hq/

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., Procter, R., Hartswood, M., Slack, R., Simpson, A., Coopmans, C., … Voss, A. (2005). Collaboration and trust in healthcare innovation: The eDiaMoND case study. *Computer Supported Cooperative Work (CSCW)*, *14*, 369–398. https://doi.org/10.1007/s10606-005-9001-0

Jones, M. B., Schildhauer, M. P., Reichman, O. J., & Bowers, S. (2006). The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics*, *37*, 519–544. https://doi.org/10.1146/annurev.ecolsys.37.091305.110031

Kahnemann, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristic and biases*. Cambridge, UK: Cambridge University Press.

Karasti, H., & Blomberg, J. (2017). Studying infrastructuring ethnographically. *Computer Supported Cooperative Work (CSCW) 27,* 233–265. https://doi.org/10.1007/s10606-017-9296-7

Kelty, C. M. (2012). This is not an article: Model organism newsletters and the question of 'open science.' *BioSocieties*, *7*(2), 140–168. https://doi.org/10.1057/biosoc.2012.8

Kohler, R. E. (1994). *Lords of the fly: Drosophila genetics and the experimental life*. Chicago, IL: University of Chicago Press.

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

Leonelli, S. (2010). Packaging small facts for re-use: Databases in model organism biology. In M. Morgan & P. Howlett (Eds.), *How well do facts travel? The dissemination of reliable knowledge.* Cambridge, UK: Cambridge University Press (pp. 325–348).

Leonelli, S. (2013). Integrating data to acquire new knowledge: Three modes of integration in plant science. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *44*, 503–514. https://doi.org/10.1016/j.shpsc.2013.03.020

Leonelli, S. (2015). What counts as scientific data? A relational framework. *Philosophy of Science*, *82*, 810–821. https://doi.org/10.1086/684083

Leonelli, S. (2016). *Data-Centric Biology: A Philosophical Study*. Chicago, IL: University of Chicago Press.

Longo, D. L., & Drazen, J. M. (2016). Data sharing. *New England Journal of Medicine*, *374*, 276–277. https://doi.org/10.1056/NEJMe1516564

Loukissas, Y. A. (2019). *All data are local: Thinking critically in a data-driven society*. Cambridge, MA: The MIT Press.

Mandell, R. A. (2012). Researchers' attitudes towards data discovery: Implications for a UCLA data registry. *Libraries in the Digital Age (LIDA) Proceedings*, *12*. Retrieved from http://ozk.unizd.hr/proceedings/index.php/lida2012/article/view/59/43

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

Mayernik, M. S., & Acker, A. (2017). Tracing the traces: The critical role of metadata within networked communications [opinion paper]. *Journal of the Association for Information Science and Technology, 69,* 177–180. https://doi.org/10.1002/asi.23927

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

Mayernik, M. S., Wallis, J. C., Borgman, C. L., & Pepe, A. (2007). Adding context to content: The CENS deployment center. *Annual Meeting of the American Society for Information Science & Technology*, *44*, 1–7. https://doi.org/10.1002/meet.1450440388

McNutt, M. (2014). Reproducibility. *Science*, *343*, 229. https://doi.org/10.1126/science.1250475

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*, 538–558. https://doi.org/10.1214/ss/1177010269

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

Mosconi, G., Li, Q., Randall, D., Karasti, H., Tolmie, P., Barutzky, J., … Pipek, V. (2019). Three gaps in opening science. *Computer Supported Cooperative Work (CSCW)*, *28*, 749–789. https://doi.org/10.1007/s10606-019-09354-z

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. https://doi.org/10.17226/25303

National Research Council, Committee on Issues in the Transborder Flow of Scientific Data. (1997). *Bits of power: Issues in global access to scientific data*. Retrieved from http://www.nap.edu/openbook.php?record_id=5504

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Paisley, W. J. (1980). Information and work. In B. Dervin & M. J. Voigt (Eds.), *Progress in the communication sciences* (Vol. 2, pp. 114–165). Norwood, NJ: Ablex.

Pasquetto, I. V. (2018). *From open data to knowledge production: Biomedical data sharing and unpredictable data reuses* (PhD dissertation, UCLA). Retrieved from https://escholarship.org/uc/item/1sx7v77r

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* (PhD dissertation, UCLA). Retrieved from http://dx.doi.org/10.2139/ssrn.1616935

Pepe, A., Borgman, C. L., Wallis, J. C., & Mayernik, M. S. (2007). Knitting a fabric of sensor data resources. *Proceedings of the 2007 ACM IEEE International Conference on Information Processing in Sensor Networks*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.596.2608

Polanyi, M. (1966). *The tacit dimension*. Garden City, NY: Doubleday.

Porter, T. M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, NJ: Princeton University Press.

Prieto, A. G. (2009). From conceptual to perceptual reality: Trust in digital repositories. *Library* Review, 58(8), 593-606. https://doi.org/10.1108/00242530910987082

Rung, J., & Brazma, A. (2012). Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, *14*(2), 89–99. https://doi.org/10.1038/nrg3394

Ryle, G. (1949). *The concept of mind*. London, UK: Hutchinson.

Sands, A. E. (2017). *Managing astronomy research data: Data practices in the Sloan Digital Sky Survey and Large Synoptic Survey Telescope Projects* (PhD Dissertation, UCLA). Retrieved from http://escholarship.org/uc/item/80p1w0pm

Schmidt, K. (2012). The trouble with "tacit knowledge." *Computer Supported Cooperative Work (CSCW)*, *21*(2–3), 163–225. https://doi.org/10.1007/s10606-012-9160-8

Shapin, S. (1994). *A social history of truth: Civility and science in seventeenth-century England*. Chicago, IL: University of Chicago Press.

Star, S. L., & Griesemer, J. (1989). Institutional ecology, "translations," and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–1939. *Social Studies of Science*, *19*, 387–420.

Strauss, A., & Corbin, J. M. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: SAGE.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, *6*, e21101. https://doi.org/10.1371/journal.pone.0021101

Thompson, E. P. (1971). The moral economy of the English crowd in the eighteenth century. *Past & Present*, *50*, 76–136.

University of California eScholarship Repository. (2011). *Center for Embedded Network Sensing*. Retrieved from https://escholarship.org/uc/cens

U. S. National Science Board. (2005). *Long-lived digital data collections: Enabling research and education in the 21st century* (No. US NSF-NSB-05-40). Retrieved from https://www.nsf.gov/pubs/2005/nsb0540/

Wallis, J. C., Borgman, C. L., Mayernik, M. S., & Pepe, A. (2008). Moving archival practices upstream: An exploration of the life cycle of ecological sensing data in collaborative field research. *International Journal of Digital Curation*, *3*(1), 114–126. https://doi.org/10.2218/ijdc.v3i1.46

Wallis, J. C., Borgman, C. L., Mayernik, M. S., Pepe, A., Ramanathan, N., & Hansen, M. A. (2007). Know thy sensor: Trust, data quality, and data integrity in scientific digital libraries. *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., Mayernik, M. S., Borgman, C. L., & Pepe, A. (2010). Digital libraries for scientific data discovery and reuse: From vision to practical reality. *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, 333–340. https://doi.org/10.1145/1816123.1816173

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLOS ONE*, *8*, e67332. https://doi.org/10.1371/journal.pone.0067332

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018.

Yakel, E., Faniel, I. M., Kriesberg, A., & Yoon, A. (2013). Trust in digital repositories. *International Journal of Digital Curation*, *8*(1), 143–156. https://doi.org/10.2218/ijdc.v8i1.251

This article is © 2019 by Irene V. Pasquetto, Christine L. Borgman, and Morgan F. Wofford. The article is licensed under a Creative Commons Attribution (CC BY 4.0) International license (https://creativecommons.org/licenses/by/4.0/legalcode), except

where otherwise indicated with respect to particular material included in the article. The article should be attributed to the authors identified above.

Footnotes

1. A summary of our findings by the three research questions is included in the supplemental materials: https://hdsr.mitpress.mit.edu/pub/tn4j86t1/branch/2 ↩