Engaging Science, Technology, and Society

Engaging Science, Technology, and Society 7.1 (2021), 12-21. DOI:10.17351/ests2021.777

# Data Durability:
# Towards Conceptualizations of Scientific Long-Term Data Storage

ESTRID SØRENSEN
RUHR-UNIVERSITY BOCHUM
GERMANY

LAURA KOCKSCH
RUHR-UNIVERSITY BOCHUM
GERMANY

## Abstract

With the increased requirement for open data and data reuse in the sciences the call for long-term data storage becomes stronger. However, long-term data storage is insufficiently theorized and often considered as simply short-term data that are stored longer. Interviews with scientists at a German university show that data are not in themselves durable; they are made durable. While Science & Technology Studies data research has emphasized the relational character of data, always embedded in local contexts and infrastructures, we propose to add the temporal dimension of data durability to this understanding. We replace notions of long-term and short-term stored data with notions of *publication data* and *project data*, because the latter terms point to the practices through which data durability is made in a variety of ways, contingent on the kind of research phases in which the data are embedded, and on their infrastructures and practices. With the notion of *data durability devices* we inquire into technologies and tools, techniques and skills as well as organizational arrangements, cultural norms and relations that contribute to making data durable. We define scientific data as durable as long as they can operate in a socio-technical apparatus and uphold their capacity to make claims about the world. The scientists' data practices revealed what we term *media data durability devices* and *scientific data durability devices*. The former were media materiality, the care for this materiality, and the compatibility between data and the data apparatus, which all contributed to shaping data durability. *Scientific data durability devices*, on the other hand included concealment and competition, through which data durability was prolonged, but also distributed unevenly among researchers. With these proposed concepts we hope to initiate discussions on the making of long-term data storage, just as we believe the concepts to be helpful for making realistic and relevant decisions about what data to store and for how long.

## Keywords

data infrastructure; data management; Germany; materiality; temporality



**To cite this article:** Sørensen, Estrid, and Laura Kocksch. 2021. "Data Durability: Towards Conceptualisations of Scientific Long-term Data Storage." *Engaging Science, Technology, & Society* 7.1: 12–21. https://doi.org/10.17351/ests2021.777.

To email contact Estrid Sørensen: estrid.sorensen@rub.de.

## Introduction

> "We store them forever and hope that we'll never need them."

This quote from an interview with a chemist interlocutor about his research data echoes the pervasive tension in the sciences between requirements for endless data storage and the practical complexity of making data accessible. As we will discuss in this paper, interviews show data durability varies considerably across scientists' data practices. We coin *data durability—the period in which scientific data can operate in a socio-technical apparatus and uphold their capacity make claims about the world.* In conjunction with a *relational* and *local* approach to data, which recognizes that data are always embedded in local contexts and extended infrastructures, and thus are interdependent on situated and changing relations (Loukissas 2019; Vertesi and Ribes 2019), we add a *temporal* dimension to the study of data practices. Moreover, we hold that data durability presupposes assessments about what *kinds* of data will be made durable for future reuse. It is important to study *data durability* along with the *devices* for making data durable to gain a better understanding of the temporal aspects of data. Additional to providing a nuanced empirically founded picture of how data durability is achieved in practice, the terms we offer contribute timely resources to STS for thinking about and discussing temporal dimensions of data. The notion of *data durability* draws attention to durability as an *achieved quality* of data, rather than being simply about storing data longer. It urges STS scholars to attend to data durability as an important area of research. With the notion of *data durability devices* and with our characterization of different data durability devices, we contribute important theoretical tools to STS for how to study the efforts, infrastructures and resources that are invested into making data durable.

As funder and government requirements for open data and data reuse grow, the need for long-term data storage becomes greater. In 2016, G20 leaders issued a paper advocating for the application of FAIR principles in research to make data findable, accessible, interoperable, and reusable (Wilkinson et al. 2016). Research grant applicants are required to submit a description of their data management plan for making data publicly available long-term. As is the case in many other countries, a ten-year standard for storage of research data has been established in Germany (Deutsche Forschungsgemeinschaft 2013; 2019). However, Tempini points out that *"data can be productively reused only if they [are] related with other data, and this relation is stabilized in a new relational dataset object"* (2020, 258, italics in original, insertion by authors). This insight suggests that long-term storage is not enough to ensure data reuse. For this, data also need adequate preparation and "packaging" (cf. Leonelli 2016) to make them interoperable both with other data and with data processing procedures, techniques and infrastructures. Even though the dominant discourse emphasizes the need to store *all* scientific data (Deutsche Forschungsgemeinschaft 2015; Tenopir et al. 2020), interviews show that scientists' data practices differentiate between which data are saved for longer or shorter periods of time, or not at all. We discuss this differentiation along the lines of a distinction between two *kinds of data*: publication data and project data. Secondly, we inquire into the data practices that make data last for longer or shorter and point to two different types of *data durability devices*: media and scientific.

Devices for making data durable included hard drives, repositories and data care, but the durability of data to be able to support scientific claims also appeared to depend on the development of the scientific discourse and on the sharing of data. One scientist stressed that if he shared his data, other scientists may

make groundbreaking discoveries with the data before his own team, thus decreasing data durability for them. These and other practices involved in achieving data durability will be discussed as *data durability devices* below. Beyond the empirical insights, the aim of this article is to work out concepts for thinking about data durability. We do so through analysis of interview data. Inquiring into temporal aspects of data raises more fundamental questions about what long-term and short-term storage mean, terms whose meaning and value in the literature and in funding organizations are generally regarded as straightforward. Our research suggests a more complex picture of what "long-term" and "short-term" mean in scientific data storage practices and how these temporal qualities are achieved.

Prior to the analysis of our empirical material, we briefly introduce the current STS literature relating to data and temporality. We then outline the genealogy of the discourse of long-term scientific data storage in Germany. The analyses of data durability are divided into a section about kinds of data and their particular durability and a section on durability devices for making data last. We conclude that attention to the complexity of data durability and to how it is made through data practices is crucial for developing a vocabulary that enables a discussion of which data are relevant to store for how long and to offer better support for relevant long-term data storage.

## Data and Durability: The Literature

A central objective of STS data research is to problematize the popular imaginary of data that treats data as "objective 'givens' which are 'out there' like natural resources such as water or oil" (Pinel et al. 2020, p. 178), and that portrays data as holding universal epistemic facts (Loukissas 2019). In contrast, STS scholars have shown how data production and data (re)use require packaging efforts (Leonelli 2016; Leonelli and Tempini 2020; Plantin 2019), organizational adjustments and work (Birnholtz and Bietz 2003; Gitelman 2013; Pine et al. 2016; Ribes and Finholt 2009), epistemic competences and negotiations (Hine 2008; Marres 2017; Passi and Jackson 2017), technical infrastructures and their socio-technical negotiations and maintenance (Bowker 2008; Cohn 2019; Dourish 2017; Edwards 2003), as well as local sensitivities (Loukissas 2019; Pinel et al. 2020; Zhou et al. 2010; Zimmerman 2008). We join these critical approaches to data. They reject the view of data as given objects and universal facts. Instead, they acknowledge data as emergent and relational compositions that are shaped through socio-material practices and infrastructures situated in historical and local contexts.

Data's locality as well as their situated and spatial nature are central tropes in STS literature. Loukissas (2019) emphasizes that "all data are local," while Leonelli (2016) has coined the notion of "data journeys," also drawing on a spatial metaphor. Studies of data most often attend to the production and use of data, including the local and trans-local practices and infrastructures these practices involve. Fewer authors have attended to the temporal dimensions of data. Among these, Ribes and Finholt (2007; 2009) have researched the long-term durability of data infrastructures, and more recently Cohn (2019) studied data temporality in terms of their decay. Both accentuate the extensive work involved in making data infrastructures endure. Based on similar insights, Faniel and Jacobsen (2010) critique the naïve assumption implied in the greater emphasis on scientific data sharing, that "making data more widely available will ensure reuse. It will not." (p. 356). The authors underline those practices, routines and infrastructures for reuse need first to be established. Bowker (2008) has studied the complexities sustaining data management and their content over time, and in his research of climate science, Edwards (2003) asserts that even if

infrastructures are made durable (cf. Bowker and Star 1999; Star and Ruhleder 1996), data's meaning and operationality change over time because both epistemic, technical, and social infrastructures become modified. As these latter references show, data durability has already been addressed in STS literature on data practices, even if this issue is rarely of focal interest. Due to the increased requirement of long-term data storage and the growing emphasis and interest in long-term storage of scientific data, it is crucial to develop a better understanding of data durability in relation to actual scientific data practices and infrastructure. The current trend towards long-term data storage has pushed scientists to engage more with long-term storage as a practical task for which, however, few are trained. Furthermore, when considered a purely practical matter, and most frequently viewed as invisible work (Star and Strauss 1999), data durability remains under the threshold of scholarly debate necessary for making both realistic and relevant decisions about data storage. To remedy the lack of reflective attention given to long-term data storage in the sciences, we work out a number of terms in this article to help develop a vocabulary that will support academic discussion of data durability.

## Genealogy of the Ten-Year Data Storage Standard

Before attending to our interview material, we will introduce the German standard for long-term data storage, since it makes up a core institutional context for the data practices we discuss. Since 1998, the German Research Foundation (DFG) has required all research data be stored for ten years. This standard has its roots in research ethics. As a reaction to cases of scientific data falsification in German cancer research discovered in 1997, the DFG assembled a think tank to deal with fraud and ethics in science and published the document "Safeguarding Good Scientific Practice: a memorandum" in 1998, in both German and English. The memorandum contains sixteen recommendations, of which number seven is titled "Data Handling." The pretext to this recommendation addresses the handling of "primary data" and states that: "[p]rimary data as the basis for publications shall be securely stored for ten years in a durable form" (Deutsche Forschungsgemeinschaft 1998, 55). As an explanation for the specified time period, it is indicated that:

> In the USA it is customary that such policies require the storage of primary data (with the possibility of access by third parties entitled to it):
> - in the laboratory of origin
> - for eight to ten years after their generation (ibid. 56).

Shortly after the turn of the century, scientific institutions extended their interest in long-term data storage from being solely related to research ethics to also being an asset for creating more value out of data reuse. A 2008 publication by the European Heads of Research Councils (EUROHORCs) and European Science Foundation (ESF), later referred to as a foundation for DFG (2015) policies, and states with respect to "research data" that:

> The collection of research data is a huge investment. Permanent access to such data, if quality controlled and in interoperable formats, will allow better use to be made of this investment because it allows other researchers to (re)use them. Furthermore, it allows re-analysis and could play a role in ensuring research integrity (EUROHORCs and ESF 2008, 17).

The primary focus of this publication is the prospect of data storage as a means for better data exploitation; research ethics and verification are secondary. This twin rationale of both ethics and data exploitation for long-term data storage appears again in the 2015 "DFG Guidelines on the Handling of Research Data" (Deutsche Forschungsgemeinschaft 2015). Over the course of two decades the rationale for the ten-year standard for long-term data storage thus moved from being about preventing scientific fraud to also—or even primarily—being a device for enabling secondary and tertiary data reuse.

While in DFG's 2019 "Guidelines for Safeguarding Good Research Practice" no justification for storage is given, the duration of storage is now more specifically addressed in guideline seventeen, titled "Archiving":

> Researchers back up research data and results made publicly available… and retain them for an appropriate period of time… the research data (generally raw data) on which they [the results] are based are generally archived in an accessible and identifiable manner for a period of ten years… In justified cases, shorter archiving periods may be appropriate; the reasons for this are described clearly and comprehensibly (Deutsche Forschungsgemeinschaft 2019, 20, brackets added).

This document is of particular significance. From 1st August 2019 all research institutions are required to implement the guidelines in a legally binding manner to be eligible to receive DFG funding. The formulation of this current guideline points to an increased attention to the duration of storage itself, and to it not always being feasible to store "research" and "raw" data for ten years.

### The Ten-Year Standard in Data Management Practice

Experts from three German data storage and data management institutions whose core assignment is to make scientific data available for reuse highlighted in conversations with us the importance of long-term storage, but to a lesser significance, of the exact length of the duration. The GESIS Institute in Cologne stores and provides access to data from many social science studies on the population in Germany. The head of its secure data center explained that only recently, in a process of professionalizing their services, they started operating with the ten-year standard. Earlier, they negotiated an ad hoc storage time for each data set with their scientific clients. As they decided to develop "service packages" for scientific data storage they had to define the service conditions more precisely, including the storage period. They decided on ten years, less to meet the DFG standard and more, because ten years was assessed to be a period that the institute could expect to have funding to support their infrastructure and thus guarantee the commissioned service. "But in reality, we don't have a horizon for storage duration. Our understanding is that we store data forever," our interlocutor from GESIS explained. An expert at the Institute for Education Information (DIPF) in Frankfurt, which offers similar services as GESIS for educational data, confirmed this practice. At the Technology, Methods, and Infrastructure for Networked Medical Research (TMF) in Berlin, an advisory institute for medical data infrastructures, an expert underlined the right of patients to have their data deleted, which makes the idea of storing data forever less applicable to medical data. It is noteworthy that the specification of data durability emerged in relation to an economic rationale, and data ethics also worked as a device for limiting data durability. In addition, the experts' assessments point to the different relevance of long-term storage in different disciplines and emphasize the pragmatic interpretation and handling of

the ten-year standard (cf. Lampland and Star 2009). Ten is not treated as an exact number of years; it works rather as a heuristic for "long time" or even "forever."

This outline of the genealogy of the DFG standard for long-term storage of scientific data in Germany points to the increased requirement for specification of long-term storage. Data management experts, however, treat durability specifications rather pragmatically. For our discussion below it is worth noticing that the specifications applied for the notions of data—"research data," "raw data" and "primary data"—remain vague. As we show in our empirical analyses, the relevance and feasibility of long-term storage varies with the kind of research data. A better understanding of data durability in relation to actual scientific data practices and infrastructures is needed to make both realistic and relevant decisions about data storage.

## Methods and Materials

Methodologically we approach temporality as a matter of "doing time." STS scholars in the wider area of research into time and temporality in relation to technology have recently developed this research style (e.g. Mol 2006; Sager and Zuiderent-Jerak 2021). It has been critiqued that most literature on time and technology treat time as linear, while in practice various temporalities are enacted that take different shapes (e.g. Sørensen 2007; Cohn 2019; Maguire 2020). These authors point to the importance of applying a practice approach—or praxeography (e.g., Sørensen & Schank 2017; Mol 2003)—to the study of the temporality of technology as emerging out of socio-technical practices. A practice approach is particularly relevant when the ontologies of the object in question have only sparsely been subjected to empirical research and thus lack detailed description of their variations and diversity. In our interviews we did encounter nonlinear temporalities, such as backup routines and delete-days that followed repetitive rather than progressive rhythms. However, the focus of this article is long-term storage and data durability that as phenomena imply a linear temporal imaginary. In their discussion of the tension between long-term data infrastructure building in the sciences and short-term funding schemes, career development etc., Ribes and Finholt (2009) maintain that the establishing and maintaining of linear time is dependent upon practices involved in sustaining the operation and meaning of this temporality. Similarly, we inquire into the practical achievement in science of linear data durability although it is also interlocked with nonlinear temporal practices.

The analyses presented on the following pages emerged from a study of data practices around the establishment of a new datacenter at a German university. We conducted 13 interviews at the university over a period of one year about data storage practices in the sciences. The interlocutors were asked about their own data storage and that of their teams and fields; about what data they stored; how they decided upon data storage; which social and material practices and infrastructures supported their data storage and data durability; as well as the challenges and controversies they experienced in relation to data storage. The semi-structured interviews lasted between 60 and 120 minutes, were all conducted in German and in person in the interviewees' offices, and all but two were recorded and subsequently transcribed. Interview citations in this article are all translated by the authors. Out of this body of material, four interviews with established scientists in their fields were selected as the core material for the study of long-term data storage and data durability. These interviews provided more detailed insights into data storage practices than the interviews with less-established scientists and non-scientific staff. The four interviews were coded according to the

principles of grounded theory (Strauss & Corbin 1998). This procedure implies attention to what was of concern in processes of scientific practice rather than providing a generalized view of data from the perspective of scientific results or from how data would normatively be expected to be handled (cf. Sørensen 2009). A first coding focused on utterances about temporality and endurance of storage and data. This coding was further refined to code the kind of data scientists talked about when discussing data temporality and what devices they mentioned that supported or undermined storage for longer or shorter periods of time.

Since the number of interviews was small and covered only four disciplines, we were careful in the interviews to ask interlocutors to what extent they viewed their practices as idiosyncratic or typical for the field, and in our interpretation to draw primarily on utterances that held relevance beyond the scientists and fields in question. Nonetheless, we would surely have come to draw a different image of data durability had we instead interviewed psychologists, computer scientists, anthropologists or scientists from other fields. Yet, the empirical material is not used to present typical data practices in the sciences but to suggest concepts we believe may help scholars to think about, discuss and develop data durability measures.

## The Interlocutors

The analyses of data durability to follow are based on interviews with four scientists whose names and locations are pseudonymized:

*Ralph Anderson* is a senior lecturer of geomatics, a discipline concerned with the collection, distribution, storage, analysis, processing and presentation of geographic data. He is also the head of the geography department's geo-information systems. Anderson notes in the interview that 80% of all data have territorial references and that, accordingly, a great variety of data are relevant for his research, which at its core is about identifying spatial patterns in data. The data he attends to include satellite photos, maps, digital spatial models, and simulations, as well as all kinds of geolocation data typically drawn from open data repositories such as, for instance, real estate cadaster. Most of the data he uses are available in repositories, but he also occasionally uses data collected by himself or at the department, such as weather data (temperature, air pressure, wind speed, wind direction, humidity), aerial photos and even survey data.

*Eric Monroe* is an assistant professor at the medical department in the interdisciplinary and data-centric[1] field of proteomics, the large-scale study of proteins. As a bioinformatics scholar, he typically receives data for analysis from his medical colleagues or works with mass spectrometry data available in data repositories. The data may be standard medical measurements such as blood values (lactate, leucocytes, calcium etc.) and vital parameters (heart rate, blood pressure etc.) along with proteomics data from blood plasma samples, based on which algorithms are trained to identify patient survival chances. Other data are mass spectrometry data based on which algorithms are developed to detect unknown protein variants. For

[1] Although data are important to all empirical sciences, Leonelli (2016) coins the notion "data-centric science" to point to a specific recognition of data in some branches of science that have intensified over the past decades, where data figure not only as key components of scientific inquiry, but whose production and circulation constitute top priorities for the scientific community.

these procedures the data quality is crucial for Monroe's team's ability to train and develop useful algorithms.

*Gregory Rosenberg* is a professor of theoretical physics and specializes in the study of intermittency and singular structures in turbulent plasma, which is research into the chaotic dynamics of turbulence. Rosenberg's research consists of running numerical simulations to "develop new physical laws," as he put it. Here, data do not emerge from measurements or observations. Instead, a theory in question—for instance on turbulence—is simulated numerically. Because these simulations are run in the six physical dimensions relevant to plasma physics, everything is calculated to the power of six. The numerical outcome of a simulation works as data that enable physicists to explore the theory further and compare it to experimental data. Data reduction is a core challenge, as a single CPU-hour[2] simulation may produce a data volume of tens of terabytes, and simulations may involve up to millions of CPU-hours.

*Simon Roth* is a professor of theoretical inorganic chemistry. He specializes in the development of computer simulation methods for the prediction of the behavior of porous solids, such as how much CO2 crystalline can absorb. The computational methods his team develops along with the documentation of the developmental process do not require much storage space. However, in order to test the methods they develop, his team runs simulations that generate binary data, text data or ASCII data that indicate the positions of the atoms in space of the solid under investigation, including their energy and the velocity of the particles. This results in "a huge list of numbers" of up to several terabytes, Roth noted, since both the solid molecule and the gas may contain tens of thousands of atoms, and the position of each atom is additionally indicated by three spatial dimensions and computed in sometimes millions of configurations.

## Kinds of Data

The policy documents discussed above recommending and requiring long-term scientific data storage used notions of "research data," "raw data," and "primary data." The coding of our interviews showed that the way in which scientists talked about data storage and data durability depended very much upon what kind of data they were talking about. A more nuanced discussion of long-term data storage thus requires a vocabulary that differentiates between data with respect to their durability. Following the scientists' own differentiation, we divide the data related to durability and mentioned by our interlocutors into two different kinds: *publication data* and *project data*.[3]

[2] A CPU-hour refers to the number of processor units used to run a simulation multiplied by the duration of the job in hours.

[3] The notion of "kind" can be applied with reference to many different parameters. Morgan (2020) also talks about kinds of data. She differentiates between data with reference to what measuring instruments, principles and strategies that are applied to "recognise, collect, code, assemble, and organise the information from raw observations into numbers" (ibid., 108). Our focus is not on these aspects and "kind" refers here solely to different types of data durability.

### Publication Data

The term "publication data" was our interlocutors' most frequently used and shared term for stored data. They used it to refer to the specific data on which claims made in a publication are based. An example of this is a recent publication by the proteomics scholar Eric Monroe and co-authors in which they claim that their functional metagenomic approach was proved useful to assign in vivo functions to representatives of thousands of proteins. Attached to the online publication in a section on "supporting information" readers can download three Excel tables, two pdf files and five zip folders containing different types of data about the proteins involved in the study. In our terminology these were the publication data that allowed the authors to make the said claim. It was a matter of course for all the scientists we talked to that publication data are stored and made available for verification and in some cases also for reuse. While our interlocutors shared this general understanding, their infrastructures for publication data and the practices for handling publication data differed considerably, as we discuss below.

### Variations of Publication Data

Of our four interlocutors, Eric Monroe from proteomics had the most well-established and standardized infrastructure for storage of publication data. He told us how they (like most other bioinformatics researchers) store their publication data in the data center of the European Bioinformatics Institute in Hinxton, Cambridgeshire in the UK. The value granted to bioinformatics research data and to their free accessibility and circulation is evidenced by the Institute's legal status: it is exempt from UK jurisdiction including immigration control, which means that both people and data can circulate without customs or law enforcement control inside as well as on their way to and from the data center (Treaty No. 29 1995). Authors who publish in a selection of central protein, genomics and bioinformatics journals associated with the European Bioinformatics Institute are required to store their publication data in one of the Institute's data repositories. Accordingly, the duration of publication data is not a practical concern for Monroe and his colleagues since long-term storage is delegated to the journals' publication infrastructures. Also in geographer Ralph Anderson's case, long-term data storage is mostly delegated to repositories. Anderson and his colleagues rarely have to package and prepare data for storage in repositories, because most of their research data originate from repositories. Authors simply refer to the repository in question in their publications and no data must be prepared separately or stored with the publications.

According to our inorganic chemist interlocutor, similar infrastructures are available in his discipline, even if not as standardized as in bioinformatics or geography. An additional common practice, Simon Roth explained, is to store publication data locally on hard drives to be available on request, a practice also observed by Faniel and Jacobsen (2010) among earthquake engineers. Roth had never experienced colleagues requesting publication data from his local storages, and he did not know if anyone had ever used the data available online. Neither had he ever requested such data from colleagues. Nonetheless, Roth in no way questioned the moral obligation for long-term storage of publication data, and their data were indeed always stored: "we store them forever and hope that we'll never need them," Roth stated. The lack of any actual use of his publication data along with the temporal imprecision of "forever" and "never" points to the vagueness of the discourse around long-term storage of scientific data.

Greg Rosenberg stressed that what is usually called publication data are, in theoretical physics, not data in any narrow sense of the term. Similar to Roth, Rosenberg's analyses do not draw on data from

measurements or observations. Instead, data are produced through digital simulations of the theory in question, which result in a data volume that often by far exceeds petabytes. Due to the amount of data produced in his team's calculations, storage of data for reference in publications is mostly too costly and thus not feasible. In most cases it is considered easier and cheaper for other scientists to produce the data again than it would be to carry the cost of storing such a large amount of data. A crucial component in this calculation is the expectation that in the future, CPU power will be both stronger and cheaper. Instead of storing publication data for verification and reuse purposes, theoretical physicists store the codes per point in time and the parameters applied to produce the data, i.e. the "data recipe," as Rosenberg called it. These codes and parameters are presented in publications.

### The Made Character of Publication Data

All four scientists made it clear that publication data are only a tiny proportion of the data involved in their research processes. Monroe took us downstairs to the department's 1.2 petabyte data center in the basement of their newly constructed building. The publication data stored, Monroe explained, is only a minor fragment of the amount of protein spectrometry data his team produces in experiments and analyses and which they store in this cool and noisy data center. Our interlocutor from inorganic chemistry described publication data as the result of the "evaporation" [*eingedampft*] of the total amount of data produced. In this original chemical imagery Roth made us understand publication data in analogy to little grains of calcium on the bottom of a kettle left for hours on the stove. This imagery surely points to the quantitative difference between the vast original amount of data and the small amount of publication data. However, it also points to publication data as a result of a process of transformation. Many STS scholars emphasize the *made* character of data (cf. Boellstorff 2013; Bowker 2008; Gitelman 2013; Knox 2020). Similarly, Roth's evaporation metaphor indicated that their durability is made. Publication data were not just a subset of the total amount of data produced in the research process—as the notions of long-term and short-term storage tend to suggest. Listening carefully to Roth's metaphor we noticed that he did not parallel publication data with a jar of water compared to the vast volume of the kettle. Publication data were not just a quantitative subset of the total amount of data, but indeed a transformed matter. Like grains of calcium on the bottom of the kettle have become hard and durable only thanks to the process of evaporation, publication data have become durable through processes of cleaning, sorting and analysis that allow them to specifically support the claims made in a publication. They do not resemble the whole of the data more than the grains of calcium resemble the water in the kettle.

What our interlocutors understood as publication data were routine practice to them. These data differed, across disciplines, each drawing on different data practices and infrastructures and delegating the packaging and storing in different ways: The geographer delegated both the packaging and the storage of publication data to a data repository. The bioinformatics scholar packaged publication data according to repository standards and delegated the storage to them. The chemist built websites for simulation data and also stored publication data locally. The physicist neither stored nor provided publication data, but published the "recipe" of how to produce the data. These differences substantiate that publication data are not simply data stored longer, as the notion of long-term data suggests, but that they are made durable.

### Project Data

"Publication data" was a term used commonly by our interlocutors, yet what we term "project data" were also referred to as "rubbish data" and *Nebendaten*, as we discuss below. The non-standard terms for project data mirror their non-standard character. They change in the course of the research process, considered relevant at one point, yet later often excluded, forgotten, deleted, or transformed into publication data. In this section we present two aspects that characterize our interlocutors' project data: their dependency on ongoing research and their auxillary character.

### Project Data as Entangled in Ongoing Processes

Our geography interlocutor talked about project data that result from calculations, combinations, trial runs, and are often also traces from machines' calibrations or from experimenting with data combinations. "Do you store project data?" we wanted to know. "Well, you know," Anderson explained, "if you are enmeshed in the project, and you make your processing, and you save all the intermediate operations, then you know what is going on. If you look at it a year later and you look for a specific operation, then you go nuts." The intelligibility of the data was dependent on Anderson's temporal proximity to the data analysis. Bowker points to the difficulties of dealing with old data and to "ensure that one's data doesn't rot away" (2008, 121). Accordingly, we can understand the sensation of "going nuts" when looking at one-year-old data as arising from the lack of synchronicity between theory and data. When "enmeshed in the project," the theory for sorting and aggregating data are in sync with the data. Yet, contrary to publication data, whose state is fixated through their highly processed character of infrastructures and strong bonds to specific scientific claims, project data are relevant for analysis exactly because they are underdetermined and open to various interlinkages with theory. At each step in the data analysis (sorting, cleaning, rearranging, interpreting, deciding, assessing etc. (cf. Dumit and Nafus 2018)) project data and theory are brought together; they are synchronized. If synchronization is paused without fixation and the scientist returns a year later, there is a good chance that she will not "know what is going on." Anderson's utterance pointed to the lack of durability of project data not as a quality of the data itself. One-year-old data would make him "go nuts," less because of the data "themselves," but because he lost the ability to engage with the data. Project data are characterized by their intelligibility being deeply contingent upon particular temporary and situated research processes.

### Project Data as Rubbish Data or *Nebendaten*

"Sometimes you calculate an incredible amount, and then you realize that it is all just rubbish," the chemist, Roth, explained. The tinkering with and testing of data analysis does not always lead to publishable results and thus to publication data. Geomatician Anderson used the word *Nebendaten* for that which Roth termed "rubbish data." When working for contractors, Anderson forwards data to the contracting company or institution, but there will always be additional *Nebendaten* that you cannot communicate to the contractor. The literal translation of *Nebendaten* would be "side-data." He explained: "You compute in one direction and then in another direction in order to find out what the ideal processing step is for your data." Contrary to metadata, which are second order data indexing or cataloguing of first-order data, *Nebendaten* are first-order data, no different from the data resulting from the "ideal" processing. Only their destiny is different as they are left behind and not packaged or documented as are data from the "ideal" processing. Computing

"in other directions" is essential to eventually reach the "ideal" processing, yet the moment the "ideal" step has been identified, the *Nebendaten* lose their value and can be put aside and indeed forgotten.

The DFG defines that "[r]esearch data might include measurement data, laboratory values, audiovisual information, texts, survey data, objects from collections, or samples that were created, developed or evaluated during scientific work" (Deutsche Forschungsgemeinschaft 2015, 1), which resonates with "project data," "rubbish data" and *Nebendaten*. In the 2019 "Guidelines for Safeguarding good Scientific Practice," DFG acknowledges that some data may not be feasible for long-term storage, and that for these data, a comprehensive justification for nonstorage should instead be delivered. "And this is when I very much wonder, how is all that really going to be indexed so that you can rationally exploit those entire heaps of data in the future that are lying there?" Anderson commented. Our interlocutors found comprehensive justification and documentation of "rubbish data" or *Nebendaten* to be neither realistic nor relevant.

### From Long-Term and Short-Term Data to Publication and Project Data

According to our interlocutors project data is the core material of scientific knowledge production, yet nonetheless only of value in the process of knowledge production, not when having reached a result. The lack of durability is not a deficit of these data, but a necessary component of their underdetermined character. It grants them their value for the research process. This fundamental difference in kind between project data and publication data underlines that publication data are not simply project data stored longer, since data lose their value if they are only stored. While project data are certainly stored short term and publication data are stored long term, the categorization of these data as short-term and long-term data obscure the many other differences between publication data and project data that are much more important for the research process. By contrast, the notions of *publication data* and *project data* point to these data as being related to entirely different phases of scientific practice and invite discussions of the practical specificities of these data and the relevance of their storage along with discussions of how the short-term and long-term characteristics of data are indeed made. In the following section we turn to this question.

## Data Durability Devices

Even if the focus of the previous section was not on how data durability is achieved, we hinted at publication infrastructures, smaller data volumes, financial resources and moral obligations as devices for making data durable. In this section we focus specifically on what we call *data durability devices*.

Leonelli (2016, 3) quotes Rebecca Lawrence, Managing Director of the open access publishing platform F1000: "If you have useful data quietly declining in the bottom of a drawer somewhere, I urge you to do the right thing and send it in for publication—who knows what interesting discoveries you might find yourself sharing credit for" (Lawrence 2013). This quote nicely expresses a core vision of the open science movement of rejuvenating old data that are otherwise left unexploited. Moreover, it suggests that data storage for reuse is only a matter of opening a drawer and digging out the valuable data. As stated, STS authors generally contest this imagery and have shown that data storage and data reuse are deeply dependent on the extensive work of packaging and infrastructuring data for circulation. In this section we discuss how data storage is contingent upon *durability devices* as well. We present the data durability devices our interlocutors talked about and also explore the relevance of *data durability devices* as a concept.

We divide the data durability devices we found in our interviews into two categories: media and scientific. This division is inspired by Tempini's (2020) differentiation between scientific and computer data. Since our interlocutors did not only refer to data stored on digital or computer media but also on other media materialities, we prefer the notion of "media" over "computer." With this variation, yet still strongly inspired by Tempini, we define the two types of data durability devices, which are here only indicated conceptually while over the following pages the phenomena giving rise to these definitions will be presented:

- *media data durability devices* are characterized by the capacity to support the endurance of data's ability to operate in a socio-technical apparatus through which they can be used to access or generate information (cf. op. cit., 243); and
- *scientific data durability devices* are characterized by their capacity to support the endurance of data's ability to make claims about the world that a social actor considers to be scientifically usable (cf. op. cit., 242).

In the following we present a non-exhaustive number of data durability devices we learned about in the interviews. An overview of these is given in Table 1.

| **Media durability devices** | |
|---|---|
| Materiality | Hard drives and servers<br>Backup routines<br>Underground |
| Care for materiality | A temperature-controlled and humidity-regulated room<br>Affective relationship to servers |
| Keeping data and apparatus compatible | Backward compatible processing techniques regarding older media materialities and data formats |
| **Scientific durability devices** | |
| Epistemic devices | Symbols with little historical contingency<br>Backward compatible theories |
| Competition | Data concealment<br>Transparency about data digitalization and publication plans |

Table 1. Durability devices mentioned in the interviews.

### Media Data Durability Devices

Our interview transcripts are full of references to how making and keeping data durable depend on the materiality of data media, on continuous care for this materiality, and on maintaining mutual compatibility of data media and data apparatus.

### Media materiality and the care for materiality

STS data studies scholars have recently asserted the materiality of the digital (e.g. Dourish 2017; Hogan 2015; Maguire and Winthereik 2021; Vertesi and Ribes 2019; Vonderau 2019), otherwise characterized by words like "cloud" or "the virtual" that suggest data's immaterial nature. Our interlocutors also pointed to the crucial role materiality plays in producing data durability. As an answer to how he would secure long-term

storage, our inorganic chemist interlocutor Simon Roth got up and started searching the cabinets of his office. He drew out a few hard drives that appeared to be quite aged. While he insisted that the drives and thus the data were only at risk in case of fire, he also mentioned the material fragility of the hard drives:

> When you take it out and plug it in and let electricity run through it, then the motor will have to start up again and everything will need to move. Most likely this goes well a thousand times, and then time one thousand and one it suddenly goes "schrrphthh" and the hard drive is down the drain [*im Eimer*].

As media data durability devices, the hard drives involve a mixture of storage and hope: "we always save the data on two disks and hope that only one of them breaks," Roth continued indicating the interdependence between data durability and the durability of the media materiality. A core media data durability device in Roth's team was the routine of Monday backups where all team members would back up data on hard drives or on a small server Roth had installed in the corner of his office.

A quite different media data durability device pointed to the scope and variety of this phenomenon reaching far beyond socio-technical infrastructures: due to the mining history of the region, the ground was in constant risk of subsidence. It was therefore a challenge to find a building site on campus for the new datacenter that would keep data from collapsing into the ground. Here, underground itself served as a media data durability device.

Anderson pointed to media data durability as not only dependent on materiality but also on the care for materiality. The geography department hold a large archive of aerial photos from the 1950s. Photographic paper is climate sensitive and may stretch, shrink, and bend a great deal, thus bending the dimensions of the photographic data and skewing researchers' exact reading of the images. To avoid this the department had installed a temperature- and humidity-controlled room that offered the care needed for these data to endure. Pinel et al. (2020) maintain that care for data also consists in continuous affective relationships with data. Both Roth and Anderson illustrated to us how a caring relationship to servers can be a crucial durability device. When showing us the department's server room Anderson announced "My assistant passes by regularly to listen to the servers. You need to learn to hear if something is wrong. Right now, everything sounds fine." Roth, on the other hand, worried about requirements to move servers to a new datacenter: "Why should I move my nine-year-old machine? It may not survive it," he lamented.

### Keeping Data Media and Data Apparatus Compatible

Media data durability is not only dependent on media materiality and the care for it. Likewise, the socio-technical apparatus (Barad 2007; Østerlund et al. 2020) of metaphors, tools, classifications, scientific models, formats, etc. needs to stay interoperable with the data media. For about a century the geography department's climate station has been writing weather data on paper rolls. Anderson said that this data still holds scientific value, but since the analytical methods in geography today can only process digital data, the paper data have become obsolete. A similar issue is well known in relation to digital data formats: both Monroe and Anderson highlighted the problem that updates of proprietary software are sometimes not backwards compatible with earlier data formats. A crucial device for media data durability for data on both paper and older digital data formats are backward-compatible processing techniques. Edwards (2003) points to a similar difficulty of making data endure. Even if 160 years of climate data are available, data cannot easily be brought together since "practically everything about the weather observing systems has

changed" (ibid., 6): weather stations change their location, and their environment changes through new buildings and new streets. Manufacturers go out of business and with them the instruments that are in use. In addition, the algorithms and models change for how weather is calculated. Cohn (2019) notes that data durability can also be achieved by sticking to software that maintain data relations and point backwards rather than racing for innovative and cutting-edge software.

### Scientific Data Durability Devices

Not only material entities but also scientific discourses, models and practices work as devices for rendering data more or less durable. We call these *scientific data durability devices* and we discuss a few of them below.

### Epistemic Devices: Symbols, Models and Analytical Conventions

In a longer line of reasoning, our interlocutor from inorganic chemistry pointed to the complexity and interdependency of media and scientific data durability devices. He first indicated that different kinds of symbols more or less efficiently carry data through time: those symbols in need of the least-specific historical context to interpret, will be best transported over time. Based on a widespread trust in numbers (cf. Porter 1995), Roth stated that these symbols work better as data durability devices than do the more historically contingent images and words, and that, together with optical character recognition (OCR), the comma separated variable data format (CSV) and Microsoft's Excel software (together making up a media data durability device), it is even possible to rejuvenate "ancient" numerical data:

> "in principle you can convert number forms [*Zahlenformen*] into all sorts of formats. After all, if you think about it, you have a somewhat ancient data series on a sheet of paper, then I scan it, run an OCR on it and get a CSV data set that I can import into Excel, and then I have arrived in the present time. And that is no problem, I'd say."

It may be debated whether it is particularly the case for numbers (cf. Edwards 2003), but it seems plausible that a better interpretability over time of the symbols a specific area of science uses as data will surely support the durability of this data's capacity to make scientific claims; or in other words, work as a scientific data durability device.

After discussing the durability of symbols, Roth added that this seemingly effortless transportation of data across time rarely happens in practice. Data, which are considered epistemically relevant when new formats are introduced, will be converted into new formats as soon as such are introduced. Data, which on the other hand happen to be ignored at that time, will also not be converted later. According to Roth, the reason for this is that what he called "ancient data" depend on "ancient models" that from the perspective of current inorganic chemistry seem almost an imposition: "the raw data they simulated back then: my GOD! That are quasi models... we would say, 'hey, come on: THAT?' No way that we'll commit ourselves to THAT anymore." Even if one considers numerical data universal, their dependency on theoretical models and how they undermine data's ability to make scientific claims when models become outdated must be acknowledged. New theoretical models that are compatible with former models may, on the other hand, work as scientific durability devices supporting the transfer of data from one model to the next and thus the endurance of data's ability to make scientific claims. What Cohn (2019) accentuates about software also

applies to theoretical modes: scientific data durability can be achieved by sticking to theories that maintain data relations and point backwards rather than by racing for innovative and cutting-edge theories.

### Competition as Scientific Durability Devices

As mentioned in the discussion of publication data above, our theoretical physics interlocutor Gregory Rosenberg would usually not store his research data due to large data volumes. However, he specified one study in which he was granted funding to run a 60 million CPU-hours calculation with a research center specialized in supercomputing. Due to the high costs of this calculation, a repetition for future studies was out of the question, and thus the project data were in this case stored for a while on local servers at the university for reuse. While Rosenberg would normally share data and calculations with his closest collaborators in Venice, Montpellier and Lille, the data of this extremely valuable calculation were at first kept for his own research team to exploit. This would prevent the collaborators from "inventing new physical laws" with the data before his own team did. Only after his team had run the relevant analyses were the data shared with the Italian and French partners. The strategy of concealing data from collaborators can be considered a scientific data durability device: it prolongs the period in which data enable the team to make scientifically usable claims. Notably, this data durability is limited to the team holding the data, since only they can exploit them. This indicates that when discussing the durability of data, it may be necessary to ask *for whom* specific data are durable. Concealing data from others may make them durable for oneself and sharing data may make data expire for the one doing the sharing (cf. Tempini and Leonelli 2018).

Anderson pointed to a different way in which competition in science intervenes with data durability. In our discussion above of the geography department's climate room for aerial photos as a media durability device we did not mention that the department also considered digitalizing these photographical data digitalization came with the promise of less care work and increased media data durability. However, the department was aware that similar aerial photos existed elsewhere, and they were uncertain whether other institutions were planning to digitalize them. Digitalizing geographic photos is a comprehensive task. In the midst of the geography department digitalizing their photographic data, if another institution were to make their photos available online, the department's data would become obsolete, and the work would be a waste. In a world of open data repositories there is no epistemic need for more than one copy of a data set. This one set will endure, while others will expire immediately.

## Conclusion

By attending to scientists' data practices we have shown that data durability turns out to be far more than "data stored longer." It is valued and treated differently for data adhering to different phases in the research process, which we have accordingly termed *publication data* and *project data*. These terms were induced from our interview study on scientists' data practices along with the notion of *data durability devices*. The latter term draws attention to the social, material, and discursive components of how data durability is achieved.

Our empirical study is limited in several ways, including only attending to four disciplines: inorganic chemistry, theoretical physics, geography, and proteomics. Further studies are needed to determine how scientists in other areas differentiate between kinds of data with reference to their durability and which data durability devices they apply. Even though the proposed concepts need specification and

adjustment in relation to other and broader empirical cases, in their current development they offer a sensitivity to the achievement of data durability. As we have seen, scientists already engage in complex and differentiated practices for selecting which data to store and how to achieve durability, but these are mostly idiosyncratic procedures, and not subject to systematic scholarly debate.

Due to their abstract temporality, notions of long-term and short-term storage are of little help for promoting such a debate. Pointing to phases in research processes, the differentiation between publication data and project data draws attention to the specificities of these practices and how data durability is achieved here. More specifically it suggests the need for clarification of several issues through systematic inquiry. Among these are questions about how to decide which project data to store and which to delete: which are the relevant criteria for deciding which project data are worth preparing for storage and which count as rubbish data and can be deleted? How do scientists decide on when it is relevant to keep synchronizing data and theory to prolong the usability of project data? Concerning media durability devices, it would be helpful for scientists to reach a consensus on what level of guarantee they should offer that their media materiality will endure for ten years: how many backups and how much investment into the care of media data durability is considered adequate? In relation to scientific data durability devices, could science profit from scientists being more careful to make new theoretical models compatible with older ones? Would scientists profit from agreements on how to deal with the concealment of data? Would a temporary moratorium restricting other researchers from exploring specific research questions with a named data set make an earlier sharing of the data with colleagues more likely? Is there a need for transparency about ongoing procedures for digitalization of data to avoid digitalization of the same data happening in several research groups at the same time, and to avoid a scenario where researchers refrain from digitalizing their data due to the risk of others doing so as well? The notions suggested in this article draw attention to these and other questions, which help stimulate a debate on the relevance and feasibility of making which data durable (cf. Imeri 2018; Meier zu Verl and Meyer 2018), for how long, and how.

## Acknowledgements

We would like to thank participants of the Data-Times Conference at the IT-University Copenhagen in 2020 as well as the RUSTlab at the Ruhr-University Bochum and in particular Miriam Bachmann and Jan Schmutzler for their extremely helpful comments on an earlier draft of this paper. Two anonymous reviewers from *Engaging Science, Technology, and Society* provided very considerate and careful reading of our manuscript and along with wonderful support by the *ESTS* editors they helped immensely to improve the paper. Our thanks also go to our exceptionally cooperative interlocutors.

## Author Biographies

Estrid Sørensen is a professor of cultural psychology and anthropology of knowledge at the Ruhr-University Bochum (Germany) with a core interest in data practices, experimentalist data studies and the integration of ethnography and digital methods.

Laura Kocksch is a PhD scholar at the Ruhr-University Bochum (Germany) researching into the fragility of corporate IT-security practices from a care-perspective, and she is also developing participatory and interdisciplinary methods for engaging with environmental data.

## References


Barad, Karen. 2007. *Meeting the University Halfway: Quantum Physics and the Entanglement of matter and meaning.* Durham: Duke University Press.

Birnholtz, Jeremy P., and Matthew J. Bietz. 2003. "Data at Work: Supporting Sharing in Science and Engineering." In *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work*, Sanibel Island, Florida, November 9–12 2003.

Boellstorff, Tom. 2013. "Making Big Data, in Theory." *First Monday* 18(10): 1–17.

Bowker, Geoffrey C. 2008. *Memory Practices in the Sciences.* Cambridge (MA): MIT Press.

———, and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences.* Cambridge (MA): MIT Press.

Cohn, Marisa Leavitt. 2019. "Keeping Software Present: Software as a Timely Object for Digital STS." In *Digital STS: A Field Guide for Science & Technology Studies*, edited by Janet Vertesi and Daniel Ribes, 423–446. Princeton: Princeton University Press.

Deutsche Forschungsgemeinschaft. 1998. *Proposals for Safeguarding Good Scientific Practice: Memorandum.* First edition. Weinheim: Wiley-VCH Verlag GmbH & Co. Accessed July 22, 2021.

———. 2013. *Proposals for Safeguarding Good Scientific Practice: Memorandum.* Enlarged edition. Weinheim: Wiley-VCH Verlag GmbH & Co. Accessed August 13, 2021. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9783527679188.oth1.

———. 2015. *DFG Guidelines on the Handling of Research Data.* Accessed October 27, 2020. https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/guidelines_research_data.pdf.

———. 2019. *Guidelines for Safeguarding Good Research Practice: Code of Conduct*. Accessed July 5, 2021. https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp_en.pdf.

Dourish, Paul. 2017. *The Stuff of Bits: An Essay on the Materialities of Information.* Cambridge (MA): MIT Press.

Dumit, Joseph, and Dawn Nafus. 2018. "The Other Ninety Per Cent: Thinking with Data Science, Creating Data Studies—An Interview with Joseph Dumit." In *Ethnography for a Data-Saturated World*, edited by Hannah Knox, and Dawn Nafus, 252–274. Manchester: Manchester University Press.

Edwards, Paul N. 2003. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming.* Cambridge (MA): MIT Press.

EUROHORCs and ESF. 2008. "The EUROHORCs and ESF Vision on a Globally Competitive ERA and their Road Map for Actions." *Science Policy Briefing* 33. Accessed July 5, 2021. https://www.esf.org/fileadmin/user_upload/esf/EUROHORCs-ESF-Road-Map-Report_2009.pdf.

Faniel, Ixchel M., and Trond E. Jacobsen 2010. "Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data." *Computer Supported Cooperative Work (CSCW)* 19(3-4): 355–375.

Gitelman, Lisa, ed. 2013. *"Raw Data" is an Oxymoron.* Cambridge (MA): MIT Press.

Hine, Christine. 2008. *Systematics as Cyberscience: Computers, Change, and Continuity in Science.* Cambridge (MA): MIT Press.

Hogan, Mel. 2015. "Data flows and Water Woes: The Utah Data Center." *Big Data & Society* 2(2): 1–12.

Imeri, Sabine. 2018. "Archivierung und Verantwortung. Zum Stand der Debatte über den Umgang mit Forschungsdaten in den ethnologischen Fächern." *RatSWD Working Paper Series* 267, 69–79. Accessed July 22 2021. https://doi.org/10.17620/02671.35.

Knox, Hannah. 2020. *Thinking like a Climate: Governing a City in Times of Environmental Change.* Durham: Duke University Press.

Lampland, Martha, and Susan Leigh Star, editors. 2009. *Standards and their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life.* New York: Cornell University Press.

Lawrence, Rebecca. 2013. "Data: Why Openness and Sharing are Important." *F1000 Research Blog*. Accessed July 5, 2021. https://blog.f1000.com/2013/03/14/data-why-openness-and-sharing-are-important/

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study.* Chicago: University of Chicago Press.

Loukissas, Yannis Alexander. 2019. *All Data are Local: Thinking Critically in a Data-Driven Society.* Cambridge (MA): MIT Press.

Maguire, James. 2020. "The Temporal Politics of Anthropogenic Earthquakes: Acceleration, Anticipation, and Energy Extraction in Iceland." *Time & Society* 29(3): 704–726.

Maguire, James, and Brit Ross Winthereik. 2021. "Digitalizing the State: Data Centres and the Power of Exchange." *Ethnos* 86(3): 530–551.

Marres, Noortje. 2017. *Digital Sociology: The Reinvention of Social Research.* Malden: Polity Press.

Meier zu Verl, Christian, and Christian Meyer. 2018. "Probleme der Archivierung und Sekundären Nutzung Ethnografischer Daten." *RatSWD Working Paper Series* 267: 80–90.

Mol, Annemarie. 2003. *The Body Multiple: Ontology in Medical Practice*. Durham: Duke University Press.

———. 2006. "Proving or Improving: on Health Care Research as a Form of Self-Reflection." *Qualitative Health Research* 16(3): 405–414.

Morgan, Mary S. 2020. "The Datum in Context: Measuring Frameworks, Data Series and the Journeys of Individual Datums." *Data Journeys in the Sciences*, edited by Sabina Leonelli and Niccolò Tempini, 103–120. Basel: Springer International Publishing.

Østerlund, Carsten, Kevin Crowston, Corey Jackson, and Mahboobeh Harandi. 2020. "Building an apparatus: Refractive, reflective and diffractive readings of trace data." *Journal of the Association for Information Systems (JAIS)* 21(1), Art. 10.

Passi, Samir, and Steven J. Jackson, 2017. "Data Vision: Learning to See Through Algorithmic Abstraction." In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.* New York: ACM.

Pine, Kathleen H., Christine Wolf, and Melissa Mazmanian. 2016. "The Work of Reuse: Birth Certificate Data and Health Care Accountability Measurements." In *iConference 2016 Proceedings.*

Pinel, Clémence, Cécile Pauline, Barbara Prainsack, and Christopher McKevitt. 2020. "Caring for Data: Value Creation in a Data-Intensive Research Laboratory." *Social Studies of Science* 50(2): 175–197.

Plantin, Jean-Christophe. 2019. "Data Cleaners for Pristine Datasets: Visibility and Invisibility of Data Processors in Social Science." *Science, Technology, & Human Values (ST&HV)* 44(1): 52-7.

Porter, Theodor M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life.* Princeton: Princeton University Press.

Ribes, David, and Thomas A. Finholt. 2007. "Tensions Across the Scales: Planning Infrastructure for the Long-Term." In *Proceedings of the 2007 International ACM Conference on Supporting Group Work.* New York: ACM.

———, ———. 2009. "The Long Now of Technology Infrastructure: Articulating Tensions in Development." *Journal of the Association for Information Systems* 10(5): 375–398.

Sager, Morten, and Teun Zuiderent-Jerak. 2021. "Knowing Times: Temporalities of Evidence for Implantable Cardioverter Defibrillators." *Science, Technology, & Human Values (ST&HV)* 46(3): 628–654.

Sørensen, Estrid. 2007. "The Time of Materiality." *Forum Qualitative Social Research* 8(1), Art. 2. Accessed July 5 2021. https://doi.org/10.17169/fqs-8.1.207.

———. 2009. The *Materiality of Learning: Technology and Knowledge in Educational Practice.* New York: Cambridge University Press.

———, and Jan Schank. 2017. Praxeographie [Praxeography]. In *Science & Technology Studies: Klassische Positionen und Aktuelle Perspektiven*, edited by Susanne Bauer, Torsten Heinemann and Thomas Lemke, 407–428. Frankfurt am Main: Suhrkamp.

Star, Susan Leigh, and Karen Ruhleder. 1996. "Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces." *Information Systems Research* 7(1): 111–134.

———, and Anselm Strauss. 1999. "Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work." *Computer Supported Cooperative Work (CSCW)* 8: 9–30.

Strauss, Anselm, and Juliet Corbin. 1998. *Basics of Qualitative Research: Grounded Theory Procedures and Technique.* Newbury Park, London: Sage.

Tempini, Niccolò. 2020. "The Reuse of Digital Computer Data: Transformation, Recombination and Generation of Data Mixes in Big Data Science." In *Data Journeys in the Sciences*, edited by Sabina Leonelli and Niccolò Tempini, 239–263. Basel: Springer International Publishing.

———, and Sabina Leonelli. 2018. "Concealment and Discovery: The Role of Information Security in Biomedical Data Re-use." *Social Studies of Science* 48(5): 663–690.

Tenopir, Carol, Natalie M. Rice, Suzie Allard, Lynn Baird, et al. 2020. "Data Sharing, Management, Use, and Reuse: Practices and Perceptions of Scientists Worldwide." *PLoS One* 15(3): 1–26.

Treaty No. 29. 1995. Agreement between the Government of the United Kingdom of Great Britain and Northern Ireland and the European Molecular Biology Laboratory, July 24, 1994, *Miscellaneous* 33(1994) Cm 2595. Accessed on 22 July 2021. https://treaties.fco.gov.uk/awweb/pdfopener?md=1&did=69030.

Vertesi, Janet, and David Ribes, eds. 2019. *DigitalSTS: A Field Guide for Science & Technology Studies.* Princeton: Princeton University Press.

Vonderau, Asta. 2019. "Scaling the Cloud: Making State and Infrastructure in Sweden." *Ethnos* 84(4): 698–718.

Wilkinson, Mark D., et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3(160018). Accessed July 5, 2021. https://doi.org/10.1038/sdata.2016.18.

Zhou, Xiaomu, Mark S. Ackerman, and Kai Zheng. 2010. "Doctors and Psychosocial Information: Records and Reuse in Inpatient Care." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* New York: ACM.

Zimmerman, Ann S. 2008. "New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data." *Science, Technology, & Human Values (ST&HV)* 33(5): 631–652.