Scandinavian Journal of Information Systems

Volume 32 | Issue 1 Article 1

6-30-2020

Data Curation as Governance Practice

Elena Parmiggiani
Norwegian University of Science and Technology, parmiggi@ntnu.no

Miria Grisot
University of Oslo, miriag@ifi.uio.no

Data Curation as Governance Practice

Cover Page Footnote

The research presented in this paper was part of the project “InfraData” (2017-2018), funded by the IKTPLUSS program of the Norwegian Research Council (project nr: 270912). We gratefully aknowledge the support received by the unnamed organizations and interviwees who provided their insights. We are greatful for the valuable feedback provided on earlier draft of the paper by Eric Monteiro and by the Forskerfabrikken discussion group at the Department of Computer Science, NTNU, and by the Digital Innovation group at the Department of Informatics, University of Oslo. We also thank the Editor-in-Chief and the anonymous reviewers for helping improve the manuscript.

© Scandinavian Journal of Information Systems, 2020 32(1), 3-38

Data Curation as Governance Practice

Elena Parmiggiani
Department of Computer Science, Norwegian University of Science and Technology
parmiggi@ntnu.no

Miria Grisot
Department of Informatics, University of Oslo
miriag@ifi.uio.no

Abstract. Data governance is concerned with leveraging the potential value of data in data infrastructures. In IS research, data governance has developed as a management perspective, implying a narrow view of who makes decisions about the data in infrastructures. In contrast, we propose a data governance in practice view and focus on the day-to-day decisions of users working with the data. Drawing on an interpretive case study of three data infrastructures in the Norwegian public sector, we ask: How can we characterize data governance in practice? We find that the work of data curation is a fundamental element of data governance practice. Data emerge dynamically as assets, enfolding the involved users' interests and contexts. We contribute to the IS literature in two ways. First, we characterize three main practices of data curation: achieving data quality, filtering the relevant data, and ensuring data protection. In so doing we foreground the role of the users as contributing to shaping data infrastructures. Second, we develop an analytical framework which specifies the unfolding of user involvement in data infrastructures-in-use and conceptualizes this work as emergent. Our contributions have implications for developing training support for users as data curators, and for the ethics of data management.

Key words: data governance, work practice, infrastructure, data curation.

1 Introduction

Data governance, sometimes referred to as information governance, is a concept that describes an organization's capability to ensure data accessibility, consistency, and usability throughout their lifecycle (Otto 2011; Tallon 2013). Attention to data governance is

Accepting editor: Arto Ojala

growing among academics, practitioners, and mass media outlets because organizations are currently dealing with an increasing availability of data in this age of digitalization (Ekbia et al. 2015). (Big) data are now generated and circulated to an unprecedented extent so that they have become key drivers of the digital economy (Alaimo et al. 2020) and that organizations regard data as central resources for their businesses (Abbasi et al. 2016). As a result, many organizations in both public and private sectors are implementing data infrastructures[1] to collect, organize, and analyze (big) data and use the datasets as a basis for new digital services (Fitzgerald 2016; Kitchin 2017; Ylinen and Pekkola 2018).

However, as the current unprecedented scale of the data produced by data infrastructures is changing the involved users' roles and accountability (Boos et al. 2013), research needs a closer examination of who is involved in making decisions about the data in infrastructures over time (cf. Iivari et al. 2010). Despite enthusiastic calls for studies on the potential benefits of digitalization, there is still limited understanding of how organizations can take into account actual work practices and users' interests in their data governance structures (Günther et al. 2017).

Extant research in Information Systems (IS) does not provide a common definition of data governance but typically employs a management-oriented perspective. This view focuses on the business value of data as company assets (Benfeldt et al. 2019; Otto 2011). Thus, data governance is concerned with how managers can leverage the potential value of data, for instance, by developing tools to assess the quality of the data throughout their lifecycle (Otto 2011), and by designing data infrastructures that support an organization's core data processes. However, IS scholars have only to a limited extent focused on how data governance unfolds in practice (Alhassan et al. 2016), what the users' role is, and how data infrastructures are actually used. As Mikalef and colleagues (2020) observe, the IS data governance literature assumes a direct relationship between data and data governance as an organizational capability. In doing so, scholars tend to overlook the day-to-day work of users engaged in data governance practices (i.e., working with data, interpreting outcomes, and making decisions). This is unfortunate, because foregrounding such data work practices, either internally or externally, has strategic value for organizations (Plantin 2019). A recurrent example is cited by a New York Times report stating that cleaning and preparing the data for further use account for 50-80% of the workload of data scientists (Lohr 2014). Understanding these practices is crucial in ensuring that the actual usage patterns and user roles and interests are captured and tracked by management-oriented governance frameworks.

A shift from data governance as a matter of asset management to data governance as a matter of work practice is thus essential. This is important because the intangible nature of digital data challenges conceptualizations of data qua traditional resources, as the current discourse on digitalization in IS highlights (Henfridsson et al. 2018). Data are digital artifacts that are editable, programmable, synthetic (Kallinikos et al. 2013; Monteiro and Parmiggiani 2019) and constantly evolving with the shifting ecosystems where they belong. Digital artifacts tend to acquire value and are shaped as part of situated practices of use through which users engage with technology during their day-to-day activities (ibid). This perspective is echoed by Abbasi and colleagues (2016), who invite IS researchers to be explicit about the nature of the data and their underlying values and assumptions.

To argue for an understanding of data governance in practice, we build on research taking a technology-in-use perspective. In this view, as data infrastructures are used, they dynamically evolve, shaped by the daily practices through which users engage and re-engage with technologies and recursively enact structures of technology use (Orlikowski 2000). We also draw on a perspective on infrastructures that emphasizes their sociotechnical and evolutionary nature (Aanestad et al. 2017; Ribes and Finholt 2009). Based on these perspectives, we aim to develop a theoretical understanding of data governance in practice and address the following research question: *How can we characterize data governance in practice?*

To answer our research question, we conducted an exploratory case study of the work performed by the users of data infrastructures in the Norwegian public sector in three empirical domains: remote healthcare, environmental monitoring, and city governance.

Our study contributes to the IS literature in two ways. First, at the empirical level, we identify three main *data curation practices* (Karasti et al. 2006) through which users engage with the data: achieving data quality, filtering the relevant data, and ensuring data protection. These practices unearth the dependencies that emerge across user groups and systems as the users engage with the data. By exposing these practices, we show that users become actively involved in generating and shaping the data by making mundane daily decisions about the infrastructure.

Second, at the theoretical level, we use these practices as bases for developing an analytical framework that advances our knowledge about data governance in practice and suggests directions to take actual work practices and stakeholders' interests into account in data governance. Specifically, we argue that the work of data curation is a crucial element of data governance (Leonelli 2019). In doing so, we foreground user

involvement (Iivari et al. 2010), an issue that is often overlooked by data governance frameworks. Such a focus on user involvement in data governance aligns with the Scandinavian tradition in IS (Bjerknes and Bratteteig 1995; Bratteteig and Wagner 2016), aimed at sensitizing researchers and practitioners to focus on the actual work and decision-making practices to be included in technology design. This sensitivity is important because the users increasingly emerge as data curators, as opposed to data consumers. As such, they should be supported and trained to enable them to handle the data and understand the ethical dimensions of their data curation work.

The remainder of this paper is organized as follows. We begin by reviewing the main conceptualizations of data governance in IS and the related fields. Then, we present our theoretical framework, pointing to the work of data curation as an important form of data governance in practice. Subsequently, we illustrate our study of the ongoing implementation of three data infrastructures in Norway. After that we describe our research methods where we develop our analytical framework identifying three constructs of data curation. We then present our findings in light of our framework. Finally, we discuss the significance of our analytical framework for discourses of data governance in infrastructures in IS.

2 Data governance in Information Systems

Attention to governance is rooted in organizations' need to manage and mobilize their resource portfolios by defining standard structures, processes, and decision-making roles in order to deliver value (Sirmon et al. 2007). The advent of big data has led to a growing interest in information and data as sources of business value for organizations. Unsurprisingly, there has been an increasing focus on data governance as both a core concern of organizations and a legitimate research theme in IS (Alhassan et al. 2016; Benfeldt et al. 2019; Otto 2011).

Several scholars have conceived of data governance from a perspective that we label 'data as assets' (i.e., adapting traditional definitions associated with resource management and value generation). For example, Benfeldt et al. (2019, p. 1) write: "Data governance refers to the organization and implementation of rules and responsibilities, which enforce decision making and accountabilities regarding an organization's data assets [...]. Embedded is that data governance contributes to organizational goals by encouraging desirable behavior in the treatment of data as a resource." Similarly, Otto (2011, p. 47) defines data governance as a "companywide framework for assigning

decision-related rights and duties in order to be able to adequately handle data as a company asset".

Implicit in such definitions is a top-down approach to data governance aimed at providing an organization's top management with tools and rules for controlling the data flow, usually in terms of ensuring data quality (Ofner et al. 2012), access to data and metadata, and security assessment frameworks (Khatri and Brown 2010; Tallon 2013). According to this perspective, "designing data governance requires stepping back from day-to-day decision making and focusing on identifying the fundamental decisions that need to be made and who should be making them" (Khatri and Brown 2010, p. 148).

These approaches have been criticized. For instance, Mikalef and colleagues (2020, p. 9) observe that "the human component [...] is seldom taken into account, even though it is up to humans to interpret outcomes and make decisions". Over time, several actors become involved with the data and in informing decisions about them throughout their lifecycle (Iannacci 2010). Following the data might then reveal "the implications of local working practices and knowledge for the sharing and reuse of data collected across different sites" because "local variations in practices cannot be entirely eliminated by standardization of data lifecycle protocols" (Ure et al. 2009, p. 417).

Overall, the limitations of the conceptualizations of data governance as asset management are well summarized by Alhassan and colleagues (2016), who identify three main activities related to data governance—define, implement, and monitor. The authors observe that existing research has so far focused mostly on the define phase and that there remains an immature understanding of the day-to-day decision making in data governance in practice beyond this phase. Expanding on this insight, in the next section, we propose a theoretical focus to shed light on data governance in practice.

3 Toward data governance in practice

To shift the attention in governance from asset management to a practice-based understanding, we draw on a technology-in-use perspective that emphasizes the "recurrent, materially bounded and situated action engaged in by members of a community" (Orlikowski 2002, p. 256, see also Orlikowski 2000). Such a focus implies that work practice is understood as the locus of organization, and organizational phenomena are the effects of interconnected material, discursive, and social practices (Nicolini et al. 2003). Expanding on this view, scholars of information infrastructures have engaged

with technology-in-practice at scale (e.g., in the context of large-scale infrastructures) and examined how infrastructures are constantly shaped by the entangled daily relations between humans and non-humans, such as standards, data models, and data management practices (Aanestad et al. 2017; Grisot et al. 2014; Hanseth et al. 1996).

One important finding from this literature is about the "extended design" perspective "to capture how workplace technologies can be shaped across multiple contexts and over extended periods of time" (Monteiro et al. 2013, p. 576). Similarly, Ribes and Finholt (2009) describe these processes as the "Long Now" of infrastructures, namely "the varied compendium of work done today with an eye toward generating a sustainable future" (p. 377). Underlying this perspective is the effort to foreground and be specific about the role of data users, who, through their day-to-day decisions about the data, acquire an active role and participate in shaping infrastructure evolution to handle the dependencies that emerge over time across user groups and systems (Pipek and Wulf 2009).

In this paper, we examine these practices as legitimate practices of extended design. We adopt the lens of *data curation*, which involves a broad spectrum of activities related to cleaning, assembling, setting up, and stewarding the data to make them fit with the existing templates (Leonelli 2016). Infrastructure scholars remind us that data curation practices constitute a significant portion of data governance (Karasti et al. 2006; Ribes and Polk 2014). The importance of data curation for infrastructures has emerged over the last 15 years, primarily from studies of eScience, in connection with the awareness that the exponential increase in the availability of primary scientific data requires going beyond overly technocentric accounts of data governance and embracing a more nuanced understanding of the actual work of collecting and preserving the data (Karasti et al. 2006).

Ribes and Polk's (2014, see also 2015) account of the long-term infrastructure that studies HIV/AIDS significantly illustrates data curation in IS. They describe the longitudinal endeavor to collect biological data from voluntary donors to identify the agent causing AIDS. As HIV was discovered a few years after the existing data collection started, the infrastructure demonstrated remarkable flexibility in its data governance practices that enabled using the same data archives and collecting and analyzing new data to further characterize HIV/AIDS. One of their informants concisely described this attitude toward data governance as follows: "'we were ready to handle just about any cause, as long as it wasn't aliens" (Ribes and Polk 2015, p. 224). This flexibility is important to address because data governance must always meet concerns related to supporting present decisions while being flexible regarding long-term evolution and

future use (Venters et al. 2014). The consequence of this observation also confirms that data are not predefined assets in governance but tend to emerge dynamically as assets that are part of work practices, enfolding the involved stakeholders' interests and contexts (Monteiro and Parmiggiani 2019; Vassilakopoulou et al. 2017). Leonelli compellingly summarizes this perspective:

> Technological development, particularly digitization, has revolutionized the production, methods, dissemination, aims, players and role of science. Just as important, however, are the broad shifts in the processes, rules and institutions that have determined who does what, under which conditions and why. Governance, in a word. Data emerge from this reading of history as relational objects, the very identity of which as sources of evidence—let alone their significance and interpretation—depends on the interests, goals and motives of the people involved, and their institutional and financial context. Extracting knowledge from data is not a neutral act. (Leonelli 2019, p. 320)

Scholars of Computer-Supported Cooperative Work have included data curation as a legitimate part of governance in infrastructure-in-use, notably in studies on scientific work (Borgman et al. 2012), healthcare (Bossen et al. 2019), data science (Passi and Jackson 2018), and in the energy industry (Mikalsen and Monteiro 2018). These studies bring specific user roles to the fore beyond the data model development phase that should be considered in governance accounts (cf. Millerand and Baker 2010). In the healthcare field, data work related to ensuring sufficient data quality requires more time to handle the emerging dependencies between workflows and the new digital systems. This process translates into the development of new competencies and skills, as well as the creation of new functions and roles for professionals and patients alike. A 'medical scribe' has thus become a new occupation in response to increased demands for documentation and digitalization in healthcare (Bossen et al. 2019).

These studies are important because they also illustrate that the users curating the data end up making decisions about the data in infrastructures through their daily work. As we shall discuss, this requires providing the users with data management training to improve their practices of auditing, peer review, and quality control, among others. Building on the above perspectives on data as emergent through practice, we characterize data governance in practice by shedding light on and conceptualizing data curation practices.

4 Case description

We draw on a two-year (2017-2018) interpretive case study (Walsham 2006). Our unit of analysis consists of the practices of handling the data in the context of the emergence and the adoption of data infrastructures in the Norwegian public sphere. We adopted an exploratory, single case study-based research design strategy (Baxter and Jack 2008) because we were theoretically interested in developing an in-depth understanding and characterization of this novel process (Pettigrew 1990) in the context of the ongoing digitalization projects in the public realm in Norway.

The authors have a history of research activity investigating data infrastructures in different domains in Norway and Scandinavia. The selection of the data infrastructures presented in this paper was driven by pragmatic concerns of access in connection with a research project funded by the Norwegian Research Council that the authors were involved in at the beginning of the study (2017). Against this backdrop, we obtained access to initiatives in three different domains: environmental monitoring, remote healthcare, and city governance.

In environmental monitoring, we studied the Norwegian node of the European long-term ecological research network (eLTER)[2], which is currently adopting transnational and cross-disciplinary standardized data-sharing infrastructures. The node consists of several stations distributed across the country. Environmental monitoring stations are highly heterogeneous, characterized by different objects of interest (e.g., terrestrial, freshwater/saltwater, or air species), funding structures, and data management traditions. In this paper, we primarily draw on illustrations obtained from a field visit to one research station in southwest Norway, which conducts long-term monitoring of a fluvial ecosystem and focusing on assessing the health of local fish species, such as trout, eel, and salmon. The data are typically collected by environmental researchers, aided by technicians via combined manual and sensor-based approaches. The larger facility at the station consists of tens of indoor and outdoor water tanks containing different fish species of varying ages. For instance, water parameters, such as temperature and turbidity, are recorded via digital sensors and sent automatically to a shared database. The data about the fish are mostly generated either by manually catching each fish and measuring its length and other health parameters or by observing its behavior from a glass observatory positioned above the tanks. Paper-based records are kept for all the manually gathered data, which are later uploaded on different databases.

In remote healthcare, we studied a data infrastructure for patient-generated health data on primary care in Norway. In this paper, we draw on the fieldwork conducted in the primary care centers in two municipalities. We focused on how nurses and patients

interacted through the data and how the data enabled new forms of care and nurse-patient interaction. The data are generated at home with the use of personal digital devices by patients affected by chronic conditions, such as diabetes, heart disease, chronic obstructive pulmonary disease (COPD), and multi-morbidities. The digital devices include scales, thermometers, and spirometers, connected to a software system for remote care that provides access to both patients and nurses. Patients access the data via an iPad app. The nurses in primary care centers work via a web interface, where they access the data, evaluate the patients' conditions, and follow up on the patients mainly via text messages. For instance, if the data report an increasing body temperature of a COPD patient, the nurses would send a message, advising the patient to take antibiotics before a full infection develops.

In city governance, we studied the piloting of a data infrastructure for mapping and modeling green areas and natural ecosystems within a municipality. In this paper, we draw on fieldwork conducted to investigate the ongoing datafication of city governance in a large Norwegian city. In this context we focused specifically on the work on real-time remote sensing of tree crowns in a data infrastructure managed by a group of environmental scientists and computer engineers at a large Norwegian institution for ecosystem research. At the technical level, the data infrastructure consists of satellite- and radar-based measurements, algorithms to compute the tree crowns' locations and distributions, and the tree distribution models outputted by the algorithms. Other algorithms are then deployed for calculating and regulating ecosystem services, such as forecasting air quality, energy effects, and volatile organic compounds. However, the data generation methods still pose a challenge; satellite and radar data tend to be unreliable, and the institution responsible for this data infrastructure is currently exploring complementary solutions, such as citizen-generated GPS data from smartphones and wearable devices. For example, additional data are gathered from sensors placed on bicycles to track and analyze people's movements in the city areas.

5 Research methods

Based on an ethnographically inspired research strategy (Myers 1999), our primary data sources were qualitative. They included data from interviews, observations of work activities, and analysis of documents (project documents, press releases, official strategy documents, and spreadsheets). A detailed overview of our data sources is provided in Table 1.

We conducted a total of 16 semi-structured interviews. Given our interest in practices to handle the data, our main informants were users involved in collecting, cleaning, generating, and interpreting the data in the three data infrastructures: researchers at environmental research stations, nurses at remote care centers, and environmental scientists. We asked them to describe how they worked with the data, e.g., how the data were generated, stored, shared, and checked, how decisions about the data were made, and their concerns. To obtain a more detailed understanding of the context of these data practices, we also interviewed project managers, research station engineers, software developers, and doctors. We asked them how the data infrastructures were developed and set up, which instrumentation was used and why, and if and how the data were intended for secondary use and further analysis in other settings, and which practices of cleaning, preparing, and modeling this entailed.

Interviews overlapped in time with participant observations. We conducted 32 hours of participant observations of work practices. These were crucial for three reasons: first, to observe how the data were handled in practice, and thus meet our practice-oriented focus; second, to get a deeper understanding of the contextual conditions under which the data infrastructures we studied were developed and used; finally, the observations informed the interviews by guiding us to ask more specific and relevant questions, and facilitated the interviews by allowing us to meet other participants in the projects we followed, and thus allowing for a snowballing strategy for identifying new informants to interview. We would typically start by following the responsible persons at an observation site (e.g., an environmental station manager, a nurse operating the remote care system at the remote care center, an environmental scientist in charge of the city governance infrastructure). As we gained confidence with the site, we would also interact with the other involved personnel, including researchers and site engineers for the environmental monitoring and city governance infrastructures, and supervising doctors and other nurses in remote care. We observed how the informants worked during their work day. We looked specifically at the practices through which they handled or produced the data (e.g., which data tools were used, how the physical space was organized). In doing so, we also constantly interacted with them (e.g., asking them to explain aloud what they were doing and why). We also organized a full-day workshop with participants from the three different domains to identify and discuss cross-cutting concerns, as well as attended specialized practitioners' conferences.

Finally, the documents that we analyzed included the national digitalization strategies, project documents, and documentation issued by the Norwegian Research Council and the European Union (EU) that regulated data infrastructures in the 2008-2019 period. The data from these documents enriched our understanding of the policy con-

Data source	*Amount and role/site for each domain*
Semi-structured interviews (16; approx. 1h each)	Environmental monitoring: • 1 environmental station manager • 2 environmental station engineers • 3 project managers (environmental researchers) Remote care: • 4 nurses • 1 project manager (nurse) • 1 doctor • 1 software developer City governance: • 1 research manager (environmental researchers) • 1 data modeler • 1 project manager (data analytics service company)
Participant observations (32h)	Environmental monitoring: • 6h at 1 environmental research station • 4h at 1 conference for environmental researchers Remote care: • 12h at 2 remote care centers (4h + 8h) • 2h at workshop with patients City governance: • 4h x 1 seminars (smart city designers) • 4h visit at data analytics service company
Document study	Internal documents (2017-2018): • Project documentation for each domain • External documents (2008-2019): • Repositories of national regulations and policy papers • Strategy documents by the European Union (e.g., guidelines for establishing/funding data infrastructures) • Strategy documents by the Norwegian Research Council (e.g., digitalization road maps)

Table 1. Detailed overview of the interviews and observations done in each domain and the documentation retrieved.

text, the public discourse on digitalization, and the rationale and vision behind each data infrastructure that we studied.

Our data analysis followed an interpretive paradigm (Klein and Myers 1999) to make sense of data governance as a complex whole by iteratively going through our and the users' situated perspectives. We analyzed our material through a deductive-inductive strategy in three phases by iterating between theory and data (Eisenhardt 1989).

In the first phase, we scoped our analytical focus, driven by our interest and ongoing engagement in the research problems encountered in data governance in infrastructures (Parmiggiani and Grisot 2019).

In the second phase, we manually open-coded our material, following Emerson and colleagues' (2011) guidelines for coding ethnographic data. We used color-based codes, highlighters, and sticky notes. In line with interpretivism's actor-centric perspective, we sought to trace what was identified as a concern for whom and why. We particularly looked for concerns related to managing the data: which issues did actors face when ensuring the persistence of the data flow, when handling the data, and when ensuring that the data were meaningful and relevant? We also sought to identify what approaches people devised to deal with such concerns in practice. We developed codes describing our informants' perspectives. For example, one such descriptive code, "Strategies to perform visual monitoring", was used to label the following excerpt from fieldnotes taken during observations at a research station in the environmental project:

> My attention is caught by a red, long elevated 'palafitte', overlooking the large fish tubs. [The station manager] tells me it is used by researchers to monitor the fish visually and record the necessary data.

It illustrates the work done to ensure that enough data are collected to describe the health of the fish in the absence of better technology to do it. We gradually refined and clustered our codes into conceptual categories corresponding to sets of overarching concerns, first each author individually and then jointly in half-day and full-day data analysis sessions aided by a whiteboard. For example, a nurse in the remote care infrastructure and an environmental scientist in the city governance infrastructure similarly commented on the information that needs to be fed into respectively an EMR or tree rendering algorithms. Despite the very different domains, both these remarks pointed to concerns with how the choices made on what to include or not in the data impact subsequent results and interpretations. Such concerns were thus grouped under the concept "Learning how data choices affect resulting data values". This process was

iterative, as we continuously revised our categories. In the end, we had nine conceptual categories. See Table 2 for empirical illustrations referring to each concept.

In the third and last phase, to evaluate the novelty of the emergent findings, we compared and contrasted our clusters against the extant literature on data governance, thus including a more deductive phase (Eisenhardt 1989). In the previous phase, we had become aware of the practical issues and concerns that shaped the data in the infrastructures. In this third phase, informed by our conceptual categories, we focused on them as instances of technology-in-practice shaping the infrastructures. This triggered us to shed more light on and conceptualize how data governance unfolded in practice (Alhassan et al. 2016). We thus clustered our conceptual categories into three constructs, representing main practices and heuristics[3] that further specify data governance in practice, namely achieving data quality, filtering the relevant data, and ensuring data protection (Table 2). These three practices resonated with the examples of data curation discussed in the literature (Karasti et al. 2006; Ribes and Polk 2014).

Our theoretical specification of data governance in practice is presented in the interpretive template in Table 2. It constitutes an analytical framework whose goal is to provide an analysis and a description of data governance in practice as our phenomenon of interest, including the relations between the constructs and the concepts and the corresponding observations (Gregor 2006).

Constructs and definitions	*Conceptual categories*	*Examples and excerpts*
Achieving data quality Practices and heuristics for producing trustworthy data of sufficient quality for aggregated analysis	Learning to produce good enough data with digital devices	"(Patients) do not just get the equipment and sit at peace with it. If we see that they do not use it properly, that they do not measure, or that they do not master it, [...] we can take it back [...]" (nurse). "So, we spend the first fourteen days just charting, comparing measurements. The patients measure as much as they want. We ask them to measure frequently, not only to learn the use of the equipment, but also to provide us with a basis for setting the threshold value because we see that in those fourteen days, we set a very wide [range of] threshold values. And then we go in afterwards, and we adjust it. Then we see that okay, here is the average" (nurse).
	Assessing quality in the long term	"The system does not allow us to record additional information. So we keep paper-based archives, including handwritten notes on anything that happened in one day that might help us interpret the data values afterwards" (environmental research station manager).
	Assessing quality of data production process	"He wants to use the pulse oximeter several times a day. He uses it when he goes up the stairs and when he goes to the store. I understand the pulse oximeter, or oxygen saturation in the blood drops to eighty. It may not be so strange; isn't it true?" (nurse).

Table 2. Our analytical framework reporting the identified constructs, the corresponding concerns, and illustrations from the empirical material.

Constructs and definitions	*Conceptual categories*	*Examples and excerpts*
Filtering the relevant data Practices and heuristics for identifying well-formed data that can be useful, both in and out of a given context	Ensuring that the data are useful, both locally and globally	"We are provided with general metadata models to record information about forests. But not all forests in Europe are equal. We need to augment the general metadata models with site-specific modifications" (head of environmental research department, practitioner conference).
	Ensuring device calibration and granularity	"In my opinion, data filtering involves both getting the right data and getting the data right" (environmental scientist). "This [image] is by city districts, and it kind of shows the changes in the small trees and the changes in the big trees, and another aspect here is, you are maybe wondering where in Oslo do we have 50 meter-tall trees. Well, that's a noise in the data here. Some of the areas have a lot of building activities where you have cranes moving about. They reflect the laser back up, and they look like trees, so you have to discount part of the data" (environmental scientist).
	Learning how data choices affect resulting data values	"Then it is a bit like this: when should you, in a way, note in the [Electronic Medical Record (EMR)]? I don't write very much in [the EMR] about these things, that is, I write in the [EMR] when I've talked to them [the patients]. After all, it is reporting [...]; they have to come and tell us how to really work […]. These are things [that we] should agree on" (nurse). "You can render trees in different ways. Here, the trees are rendered as [a] continuous canopy, and you get a slightly higher score, but if you use the data about each individual tree, then they kind of show up like this, and the score goes down a little bit. So the way you interpret [a] tree canopy has a path later on, further down the information production chain, which changes the blue-green factor that you have to use to achieve a norm. So some assumptions you make up the information production chain have [a] physical impact on the ground" (environmental scientist).

Table 2. Our analytical framework reporting the identified constructs, the corresponding concerns, and illustrations from the empirical material (cont'd)

Constructs and definitions	*Conceptual categories*	*Examples and excerpts*
Ensuring data protection Practices and heuristics for identifying and flagging potential threats related to intellectual, technical, and privacy aspects	Flagging intellectual property rights	"I am reluctant to share the data resulting from my environmental analyses in the open databases. They are important for my career advancement, and I am afraid that I will not be properly cited by those who will reuse my results" (environmental researcher).
	Preventing sensitive data sharing	"There is an exchange of data between health professionals. It is completely closed, you know; the data is connected to the individual user in [remote care system name], and it will lie with the user [and] with us [the municipality]. There is no one but those who have access to see that user" (informant from municipal care services).
	Ensuring user privacy	"This is based on the consent of the user [patient]. The user has to say for himself that 'I want you to look at it [the data] and follow it,' so [the patients] have to give their consent. We have to make sure that all patients in the project have given consent" (informant from municipal health services). "With [a] sensor [...], it is clear that it can quickly become a form of surveillance because you constantly see that 'oh, now he is out; now he comes home.' Obviously, if you have a door sensor, for example, you will always know, almost, where the user is" (nurse).

Table 2. Our analytical framework reporting the identified constructs, the corresponding concerns, and illustrations from the empirical material (cont'd)

6 Findings: data curation in infrastructures

In this section, we present our findings on data curation as governance practice by characterizing it into three main constructs: *achieving data quality, filtering the relevant data, and ensuring data protection.*

6.1 Data curation as achieving data quality

One main concern of the users working with data is quality. In principle, data quality is critical for trusting the data and ensuring their further use. Thus, data quality encompasses the practices of assessing the data production process and disembedding the data from the context of their production and related challenges. The concern for data quality is addressed differently in the three data infrastructures; however, two common issues can be identified.

The first issue is that data quality assessment depends on the skills of the data generators who may be data experts, as well as people with no previous formal training, who must learn to produce 'good quality' data. While in the environmental monitoring data infrastructure, the data generators are typically trained research scientists and technicians, in both remote healthcare and city governance, data generation is delegated to lay people, comprising elderly patients and citizens. For instance, in remote healthcare, nurses work with patient-generated data. This means that the patients themselves—who are elderly and with chronic conditions—use the devices at home, take the measurements, and produce the data needed. They perform practices that are traditionally carried out by health personnel; usually, when they visit the hospital or the general practitioner's (GP's) office, a nurse would use the devices (e.g., a thermometer, a blood pressure meter) and take the measurements. Thus, in remote care, patients need to learn how to handle the devices, position them correctly, ensure that the batteries are sufficiently charged, and check whether the data are correctly sent to the system. Nurses need to check if their patients manage the devices correctly and if they become skilled enough to generate the data. In an interview, a nurse explains:

> [Patients] do not just get the equipment and sit at peace with it. If we see that they do not use it properly, that they do not measure, or that they do not master it, [...] we can take it back [...], but […] many of these [patients] […] have tablets themselves, for they are [getting] younger and younger. [They] are down in [their] sixties, and they, like us, are very used to using the technology, so I think that will be a smaller and smaller problem in the future.

The nurses are aware that mastering the devices and the tablet—which means mastering the data production process—is not for all patients. When patients are enrolled in the service, nurses spend the first two weeks teaching patients how to use the devices. A nurse would visit a patient at home, demonstrate the use of each device, and explain some easy tricks for troubleshooting. They would also advise the patient to take many measurements during the day, just to get used to handling the devices. Often, other

family members are also involved, especially in the case of an elderly patient with cognitive issues.

Similarly, in city governance, if citizens have to share GPS data, they need to be trained in generating the right data. However, how to do so is still an open question. Additionally, quality data are often equated with having them in large amounts, but in reality, this equation does not hold. Preparing the data so that they can be re-used according to the existing templates in systems and routines is also part of the required data curation practices. Moreover, as an interviewed environmental scientist working in the city governance infrastructure says:

> There is still little understanding of what it means to produce and work with good quality data. It is somehow an immature understanding to consider good data the data that are technically adequate for analysis because they often contain personal and sensitive information about individuals, which should not be used.

For instance, while our informants agree that city governance based on data infrastructures should rely on citizens' data-sharing practices, it is not a given that all citizens are willing to share data, as well as what the regulations currently allow. Our informants are aware that participative processes must be implemented to engage the public in exploring the balance between public service and private interests.

The second issue is that data quality depends on enriching the data with additional data, which are needed to assess the quality of the dataset. For instance, in environmental monitoring, the scientists must share the datasets about the fish they monitor, both locally at the environmental station and in centralized databases. In this latter infrastructure, the data must be in a sanitized format, stripped of their contextual information. Often, scientists need to review and question the quality of a dataset. To do so, they rely on additional mundane information about the conditions on the day when the measurements were taken. The available systems typically allow them to record information about the water temperature, the number of fish, and the amount of oxygen. However, it is often difficult to assess the quality of a measurement taken a few months or years ago. Why did the water temperature vary so much in one day? Why had the number of fish decreased so much during one week? Was it due to a sensor failure or a disease of the fish?

> I follow the head scientist of the environmental research station into the office. There, they keep a PC that they use to enter data, such as temperature and oxygen level, about the fish tubs contained in the nearby building. I ask him what

> sort of database it is, and he tells me it is something that was not developed for them, but general purpose. They use it to enter information about the fish [...] but they are frustrated because the user interface of the database does not allow for adding free-text comments regarding the context in which the measurements were taken, for instance the particular conditions in a given day. I spot a little office agenda next to the PC, so I ask what it is used for. He says that it is the agenda where they note by hand anything of interest that has occurred during a day. I take a closer look. It contains telegraphic notes in the local dialect, such as "Too much food in tub 1415. The blue fan was slowly warming up." (October 30); "Small water leak. Visit by the vet. 3 fish dead in tub 1345." (October 31) "Outside temperature 10 deg." (November 2) (Excerpt from fieldnotes, June 2018).

Local conditions are tracked at the station by noting down the conditions or events on the day when the measurements are taken on a simple paper agenda. The scientists then compare the notes with the available data to assess the data quality. In other words, for the scientists, the practices of assessing the data quality depend on having sufficient information about the history of the measurement process, not only about the datasets per se. The only index that links the datasets with the contextual notes is the date. There is no official rule on what should be noted on a specific day, but the environmental scientists' experience and training enable them to assess what factors might affect the data quality assessment on a specific day. Of course, such an approach has several drawbacks. In an interview, an environmental research station manager says, "[We] have a lot of paper records about all measurements taken since the station was started in 1975, but in case of a fire, everything would be lost."

Similarly, in remote care, nurses need contextual information about patients to interpret the quality of the datasets that they receive. They need to get to know the patients, for instance, by understanding their habits and daily routines, in order to read the data appropriately. An interviewed nurse describes a patient in the following way:

> He wants to use the pulse oximeter several times a day. He uses it when he goes up the stairs and when he goes to the store. I understand the pulse oximeter or oxygen saturation in the blood drops to eighty. It may not be so strange; isn't it true? We should have it in a hundred; we others, sit down and relax a little, and then it will come back, and then he can move on.

In this case, the nurse needs to know how the patient behaves and when he takes the measurements. With the additional contextual information, she can then understand why certain values are high and how these data should be interpreted. A similar practice is needed in the case of missing data. For instance, if a patient is staying in an area with no Internet connection (e.g., a cabin in the mountains), the nurses would not receive the patient's data for a few days, and then, they would receive all the data upon re-connection. The nurses need contextual information to interpret the absence of data and the new data and trust that these are correct, such as not mistaking them for erroneous data from a malfunctioning device.

In sum, the practices for assessing data quality are emergent and contextual, as well as involve stakeholders with different interests and training in data management and who actively participate in producing the data and ensuring their quality by means of several informal heuristics. Generating good quality data is thus open to interpretation and context dependent. The data curation practices for achieving quality shape the way that data are generated and have far-reaching consequences for how they are acted on, thus shaping the data infrastructure in use.

6.2 Data curation as filtering the relevant data

Data infrastructures typically deal with large amounts of data. One concern is then about how to filter the relevant data for a given context or issue. Data filtering encompasses heuristics where the involved users must sort out well-formed data from noise. At the same time, data filtering entails matching aims and datasets to know which data are relevant for which decision and at what level, as well as learning how to single out those data. Both issues emerged from our fieldwork.

Filtering data is a form of data curation that plays out between local and global data needs. For instance, in environmental monitoring, the work of filtering the globally relevant parameters from the local datasets is fundamental for research stations to receive attention and funding from national and international initiatives. As a result, environmental monitoring practices, at least in Europe, tend to play out across two complementary analytical levels. First, the data must be tailored and make sense with reference to the local research station. Second, the data must comply with the EU's policy for data collection and initiatives to standardize and share the environmental data centrally through pan-European research infrastructures (Parmiggiani et al. 2018). For example, eLTER and similar initiatives in Europe provide environmental scientists with standard metadata models (e.g., ontologies) that they can use to organize and record information regarding natural environments, such as rivers or forests. However, there are always lo-

cal variations on each site that require different or additional parameters. For instance, parameters that are relevant to a sub-Arctic forest are not relevant to a forest in Portugal. "Not all forests in Europe are equal," an environmental researcher complained during a practitioner conference. As a result, researchers must decide which additional parameters should be recorded for local use only, concurrently with a sanitized version of the data being uploaded to a centralized database based on standard metadata models. This is a crucial step when the data are formed; the researchers make situated decisions about what and how much to record locally, as well as globally.

Similarly, in remote care, nurses are given the task to filter patient-generated data for 'global' use. However, filtering decisions has consequences that are not yet well understood. For example, nurses in municipal health services have the duty to document all interactions with patients, such as visits and phone calls, in the Electronic Medical Record (EMR). With the use of digital devices, nurses are continuously receiving data in the remote care system. Should all notifications of received data be documented? The nurses discuss whether the data should be documented in the EMR system every time a new data instance is reported by patients, or on a daily basis, or less often. An interviewed nurse explains:

> Then it is a bit like this: when should you, in a way, note in the [EMR]? I don't write very much in [the EMR] about these things, that is, I write in the [EMR] when I've talked to them. After all, it is reporting [...]; they have to come and tell us how to really work […]. These are things [that we] should agree on.

As GPs have access to the EMR system, which data are reported has consequences beyond the patient-nurse interaction. For instance, if the data from the devices are reported in the EMR system, should GPs make decisions and take actions based on these data? Would they actually have the time to engage in such tasks? At the time of our fieldwork, these were open questions. The practices of filtering data show that they remain unspecified. The nurses whom we have interviewed express concerns that this type of decision (i.e., which data should be filtered) is critical and needs to be agreed on and standardized across the infrastructure.

The second aspect of filtering our data concerns matching aims with datasets to understand which data are relevant to which decisions. For instance, in remote care, nurses deal with a large amount of very specific data (e.g., body temperature, blood pressure). Part of their work is training patients in understanding which measurement values require a follow-up. For example, for patients with COPD, a rising temperature

is a sign of infection, and they need to take antibiotics before their condition is exacerbated. An interviewed nurse says:

> [A patient] has COPD. I can follow her graph; also, I can see that okay, now it goes down a bit, and so what does [she] tend to do then? The last time she had those values, she started with the [drug name] she had in the drawer, [...] but because she gets anxious because she's getting worse and feels unsafe alone, I can be the one who can, in a way, support her and say, do you remember last time? (nurse, interview)

In this instance, the nurse helps the patient pay attention to warning signs in the data produced by the patient. This is a form of filtering delegated to patients; they should learn to discern which data are meaningful and signal that they should start taking antibiotics.

In the city governance data infrastructure, data filtering involves balancing the tension between "getting the right data and getting the data right," as expressed by an environmental scientist in an interview. This means that a lot of human work is entailed in assessing data consistency, that is, understanding if the available data are all the 'right' data. For example, the output of the algorithm for tree modeling might produce a distribution that appears correct to an untrained eye, but it might contain data that do not refer to trees at all: "This [image] is by city districts… you might be wondering where in Oslo do we have 50 meter-tall trees. Well, that's a noise in the data here" (environmental scientist, excerpt from fieldnotes). This occurs in cities with plenty of ongoing construction work, where the satellite detects cranes and mistakes them for very tall trees. The researchers' work is thus fundamental in accurately interpreting the models produced by the algorithms and filtering out the unlikely trees. In turn, getting the data right involves combining datasets to derive useful information. In practice, this translates into having time to work with the data, that is, to sort them out and put the different pieces together for specific purposes. Ultimately, this is a problem of expertise and money:

> To extract useful information costs money because you need to have the right expertise to handle the data, but there are different sorts of data, and one cannot have expertise in everything (environmental scientist, interview).

In sum, filtering data involves several important yet often informal points of decision making, and several users' concerns are involved in making decisions about what data

are relevant and good enough. The practices that entail sorting out the data thus crucially rely on developing awareness of the trajectory across the data infrastructure, such as whether the data are supposed to be shared with pan-European databases or with an EMR system. Filtering practices shape the infrastructure by deciding which data would 'travel' and which would not, as well as by distributing the responsibility for making those decisions. These decisions shape the data collection practices and the way that the data will be used for further analysis.

6.3 Data curation as ensuring data protection

Data curation also includes the practices related to data protection. In particular, we find that data protection encompasses all the strategies enacted by different actors to identify, flag, and address potential threats related to the technical security of the data (e.g., against hacking), intellectual property rights management (also in connection with the wish to publish the results of the data work), and privacy aspects.

For instance, technical security is an issue in remote healthcare. An interviewed nurse says:

> When we visit them [patients] at the start, they ask if they can use their own personal devices or mobile phones to report the data, but this is not possible [...]. Diabetic patients might already have a device at home for measuring blood sugar levels and ask if they can keep using that.

It is not possible to link any personal device to the remote care solution, and patients have to switch to the devices provided by the municipal service. These devices are selected by the company that developed the software for remote care. The technical architecture of the infrastructure requires devices to be registered in the software and be certified as medical devices in order to be used. This ensures that the data are securely recorded in the software and are accumulated in the patient records. The software company's technical team selects the devices from the market and carefully tests them. Patients and nurses have unique identification numbers and accounts in the system. Thus, the way that the architecture is set up and the secure data-handling practices shape how the infrastructure is built up and used in practice.

Many of our informants voice concerns related to intellectual property rights. In environmental monitoring, they are often unsure whether sharing data in centralized pan-European data infrastructures would allow other researchers to use their data without properly citing them, thus not acknowledging the amount of work that they put

in generating, cleaning, and sharing the data. As an environmental researcher admits during a workshop that we observed:

> I am reluctant to share the data resulting from my environmental analyses in the open databases. They are important for my career advancement, and I am afraid that I will not be properly cited by those who will reuse my results.

Similarly, the data curation carried out by the scientists is not often formally acknowledged in the policies that regulate funding for environmental data infrastructures, which are issued by public authorities, such as the Norwegian Research Council or the EU. According to a frustrated environmental research station manager whom we interviewed:

> There is a lot of work that we do that is completely absent from the official policy documents that outline the European roadmap for research infrastructure.

Many of the environmental researchers with whom we interacted are aware of this lack of formal recognition, which involves academic prestige and intellectual property rights.

The third concern related to data protection is privacy. Our informants work to ensure that personal and sensitive information is protected. Patients need to consent to the processing of their data when they are enrolled in the remote care service. An interviewee from the municipal health services explains:

> This is based on the consent of the user [patient]. The user has to say for himself that 'I want you to look at it [the data] and follow it,' so [the patients] have to give their consent. We have to make sure that all patients in the project have given consent.

However, data privacy is also not so clear cut, and data protection practices are still loosely defined and poorly regulated in some areas. For instance, in city governance, satellite data are used to map and model the green areas in a municipality. These datasets are complemented with other GPS data acquired from personal tracking devices, such as smartphones and smartwatches that track people's movements. However, the data generated by smartphones contain personal information, in addition to the data about movement. The main stakeholder group involved in making decisions about data protection therefore (at least in principle) comprises the citizens whose privacy is

threatened by surveillance technologies. Several informants point to the fact that tracking technologies evolve much faster than the regulations that are supposed to control them. This means that it is currently unclear what decisions can actually be made by citizens and that the tension between the public good and private interests remains to be resolved.

In sum, data curation as ensuring data protection requires balancing between enabling fluid data sharing, on the one hand, and the need to ensure that data are generated and used in compliance with regulations, on the other hand. Data curation is performed to resolve this tension and emerges in different ways. In environmental monitoring, it crucially hinges on the often unacknowledged and invisible adaptations by environmental scientists. In contrast, remote healthcare concerns the more visible and explicit work of setting up the technical infrastructure. Data protection practices reveal a multifaceted issue. The practices that we have investigated involve resolving tensions (at least temporarily) with other concerns, such as those related to data quality or official regulations.

7 Discussion

Data governance consists of the strategies to harness the value of data throughout their lifecycle (Khatri and Brown 2010; Otto 2011). The challenge for IS researchers is to specify how practices of technology use (Orlikowski 2000) dynamically shape data infrastructures and can be captured by conceptualizations of data governance throughout the "Long Now" of infrastructure (Ribes and Finholt 2009). For this purpose, in this study, we have addressed this research question: *How can we characterize data governance in practice*? Our findings illustrate that the complexity of data governance in practice lies in the dependencies that emerge across users and systems, as well as the ad-hoc emerging practices that are put in place to handle the emerging dependencies. To capture this complexity, we have developed an analytical framework, as presented in the constructs and conceptual categories in Table 2. These constructs are: *achieving data quality, filtering the relevant data, and ensuring data protection*, and they are further specified into 9 concepts.

Our framework highlights that much data curation work is involved in managing the emerging dependencies while using the available technologies. An immediate example is cited in Table 2, under the 'Ensuring quality in the long term' category. Interpreting data quality is not a one-off but a recurrent concern. As a result, it depends on scientists' work to find additional devices in order to gather sufficient extra information about the context in which the data are generated, given that the available systems

do not allow recording such information. Additionally, we show that data governance encompasses centralized mechanisms for representing the data, such as metadata models. Crucially, these models in turn rely on work practices to filter the relevant data in order to make changes or additions in the local data repositories (refer to the 'Ensuring that the data are useful, both locally and globally' category in Table 2) or find ways to integrate the data generated by citizens (refer to the 'Integrating user-generated data' category in Table 2).

Our framework contributes to broadening the current understanding of data governance on two core aspects. The first entails including and specifying the data curation practices. Our analytical framework accordingly complements existing data governance structures by characterizing three facets of data curation, that is, bottom-up emerging usage patterns to handle and make sense of the data as users try to achieve data quality, filter the relevant data, and ensure data protection. The second aspect involves showing that data curation consists of the practices through which users make decisions about the data on a day-to-day basis—although often under the radar. Our findings show that in doing so, users fundamentally contribute to shaping the data infrastructure. For instance, we have described how users filter which data are included in the information flow and consequently affect the practices of those acting on the data farther down the information production chain (refer to the 'Learning how data choices affect resulting data values' category in Table 2).

Thus, we contribute to the data governance literature by specifying the unfolding of user involvement in data infrastructures-in-use. The constructs presented in Table 2 are forms of user involvement that emerge during daily work practices well beyond the traditional infrastructure development phase (cf. Iivari et al. 2010). Understanding data curation as comprising forms of user involvement in data infrastructures unveils two core aspects of data governance.

First, it shows how this type of work remains largely invisible and unaccounted for in frameworks for data governance. Thus, we suggest that by foregrounding data curation practices, researchers can problematize and uncover neglected users and patterns of work (Star and Strauss 1999). For example, our concept of 'Ensuring device calibration and granularity' unearths the invisible work of cleaning the datasets to iron out the inaccuracies from inconsistent visualizations generated by algorithms that have no way of distinguishing a tree from a crane based on satellite data. Often, these contributions remain visible to other workers on the same research site but invisible to managerial levels in the organization where data governance decisions are formally made. Based on a study of archived data processors, Plantin (2019) makes a similar argument by showing that the invisibility of this type of work has long-term consequences and per-

petuates the misleading conception of data as raw for outside observers, including the management. It is thus important to inform extant data governance frameworks and argue for expansions that consider the dynamic and emerging nature of workflows in order to curate data.

Second, it shows how user involvement is emergent rather than only organized. Our analysis reveals that the practices of contributing to data governance via data curation take heterogeneous forms and are context dependent. This is illustrated by the concept of 'Learning to produce good enough data with digital devices,' where patients become data producers. Patients gradually learn how to use digital devices so that the tracked data are good enough and relevant to the patients' conditions. Such emergent nature of data curation practices implies the impossibility of pinpointing upfront who the participants will be and how they will contribute to data infrastructure. This complicates the possibility of identifying clear-cut workflows to be injected in data governance frameworks. Nevertheless, our extended focus on user involvement in data infrastructure is meant to sensitize research on including users who come to the fore over time during the data lifecycle. A focus on the user as a data curator reveals a broad range of *users* of the data, i.e., people involved in producing, using, and re-using the data: not only nurses and environmental engineers, but also software developers, research station engineers, doctors, and patients.

At the methodological level, one possible avenue for tracing user involvement in data governance in practice is to follow the relational nature of the data. As observed in the theoretical background section, data are (co-)constructed and evolve by balancing heterogeneous agendas, specializations, and modes of working (Kallinikos et al. 2013). Researchers could trace how and why data become concerns and for whom (Ribes and Finholt 2009), how conflicting concerns might generate tensions, as well as how these tensions are addressed. For example, as we have illustrated in the environmental monitoring data infrastructure, data must often serve high-level political goals as part of the EU's constant efforts to integrate science and policy in the continent and simultaneously acquire meaning with reference to situated technical and ecological conditions at an environmental research station. Data have a Janus face; they are both policy instruments and the results of often very informal adaptations. Environmental researchers participate in making decisions about data infrastructure as they try to handle this tension in practice, such as by not only adapting the data format to a sanitized database but also developing informal, locally meaningful metadata that nonetheless remain invisible to the official governance strategies. Although official accounts only recall the former part of this work, the data infrastructure is also strongly constituted by and dependent on the latter emergent adaptations.

Two implications follow from our study. The first one is the need for *continuous training and education*. User involvement places an additional burden and responsibility on the users' shoulders because they need to constantly make time and learn new skills to curate data, not only to consume them. This should be matched by additional support for the users so that they can learn to produce well-formed and relevant data. As we have illustrated in the remote healthcare case, nurses need to learn data curation practices but currently lack the support to do so. Moreover, new work practices should be implemented as nurses' tasks are being transformed. For instance, nurses need to develop novel analytical skills to enable them to participate in the new remote care data infrastructure (see also Grisot et al. 2019). Other researchers have also shown that new occupations emerge, requiring specific competencies for data work (Bossen et al. 2019). For some of the users involved, such as environmental researchers, this is also a pragmatic issue of having their intellectual property rights recognized (refer to the 'Flagging intellectual property rights' category in Table 2). As expressed by an environmental researcher, openly sharing pristine datasets does not do justice to the amount of work she has put in preparing and analyzing them, which is not acknowledged properly when the data are reused by others, with consequences for the citations she receives and her career advancement.

A second implication of considering emergent user involvement in data governance in practice relates to the *ethics of data management*. In our analysis, the data curators often navigate the unclear scenario of the ongoing digitalization but are not supported with training in the ethics of data management, especially in connection to threats to intellectual, technical, and privacy rights. As a result, users develop their own ethical code as they are provided with new tracking devices, among others. This is visible in the practices of ensuring data protection (Table 2), in which users develop a specific understanding of protection, which might not be the same as that of the other users or organizations. Consequently, their involvement in infrastructure appears based on a self-developed attitude toward the ethics of data management. From a Scandinavian IS perspective, the users' role in the data infrastructures that we have studied resonates with the figure of the 'ethical system developer' described by Bjerknes and Bratteteig (1995), namely people who typically act morally and promote democracy informally by developing their own moral code and engaging in a specific work or life context.

8 Conclusions

Data infrastructures are characterized by unprecedented reach and scale that challenge existing management-oriented conceptualizations of data governance. Data are not

fixed assets but emerge and evolve dynamically as part of the situated work practices through which heterogeneous users engage with technology. Despite the focus in IS on data governance in light of the ongoing digitalization, there is still a gap in terms of understanding how data governance can embrace such a nature of the data. To fill this gap, in this paper we sought to characterize how data governance unfolds *in practice*. We proposed an analytical framework that extends current conceptualizations of data governance by taking actual patterns of data production, use, and reuse into account.

The constructs and conceptual categories of our framework (Table 2) can be used by practitioners and other scholars as a sensitizing lens to capture usage patterns and emerging dependencies and to problematize how the data that are used to make decisions are produced by whom in data infrastructures. It can also be adopted by organization managers as a basis for not only informing, but also monitoring strategies of data governance (Alhassan et al. 2016) in public or private organizations.

We highlighted the way that users become involved in data governance in practice by making decisions about the infrastructures during their daily *data curation* activities. Specifically, developing sensitivity to user involvement in data governance in practice is important for training users in forming the data and dealing with emerging ethical dilemmas. Considering data curation as part of data governance strategies also requires data infrastructures researchers, organizations, and policy makers to further analyze the consequent redistribution of the work, time, resources, authority, and responsibilities that follow suit.

Our conceptualization of data curation as data governance practice has empirical limitations. We are aware that different palettes of practices might emerge from studies conducted in various domains, sectors, and countries. However, our analysis of data governance in practice outlines a trend as both the public and the industry sectors are implementing data infrastructures and increasingly data-intensive work practices in several contexts. We therefore believe that the framework can be adapted and extended to other domains as well.

Notes

1. Recently, several organizations and scholars have begun to refer specifically to data infrastructure to make explicit the central role of data. In this paper, we refer to data infrastructure as "the institutional, physical, and digital means for storing, sharing and consuming data across networked technologies" (Kitchin 2014, p. 32).
2. https://www.lter-europe.net/
3. While practices consist of more established approaches and routines, heuristics comprise a set of practices that the actors typically self-learn to approach and solve a problem in a way

that is sufficient to achieve an immediate, short-term aim.

Acknowledgment

The research presented in this paper was part of the project 'InfraData' (2017-2018), funded by the IKTPLUSS program of the Norwegian Research Council (project nr: 270912). We gratefully aknowledge the support received by the unnamed organizations and interviwees who provided their insights. We are greatful for the valuable feedback provided on earlier draft of the paper by Eric Monteiro and by the Forskerfabrikken discussion group at the Department of Computer Science, NTNU, and by the Digital Innovation group at the Department of Informatics, University of Oslo. We also thank the Editor-in-Chief and the anonymous reviewers for helping improve the manuscript.

References

Aanestad, M., Grisot, M., Hanseth, O., and Vassilakopoulou, P., eds., (2017). *Information Infrastructures within European Health Care. Working with the Installed Base*, Springer.

Abbasi, A., Sarker, S., and Chiang, R., (2016). Big Data Research in Information Systems: Toward an Inclusive Research Agenda, *Journal of the Association for Information Systems* (17:2): Article 3.

Alaimo, C., Kallinikos, J., and Valderrama, E., (2020). Platforms as Service Ecosystems: Lessons from Social Media, *Journal of Information Technology* (35:1): 25-48.

Alhassan, I., Sammon, D., and Daly, M., (2016). Data Governance Activities: An Analysis of the Literature, *Journal of Decision Systems* (25:sup1): 64-75.

Baxter, P., and Jack, S., (2008). Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers, *The Qualitative Report* (13:4): 544-559.

Benfeldt, O., Persson, J. S., and Madsen, S., (2019). Data Governance as a Collective Action Problem, *Information Systems Frontiers*.

Bjerknes, G., and Bratteteig, T., (1995). User Participation and Democracy: A Discussion of Scandinavian Research on System Development, *Scandinavian Journal of Information Systems* (7:1): Article 1.

Boos, D., Guenter, H., Grote, G., and Kinder, K., (2013). Controllable Accountabilities: The Internet of Things and Its Challenges for Organisations, *Behaviour & Information Technology* (32:5): 449-467.

Borgman, C. L., Wallis, J. C., and Mayernik, M. S., (2012). Who's Got the Data? Interdependencies in Science and Technology Collaborations, *Computer Supported Cooperative Work (CSCW)* (21:6): 485-523.

Bossen, C., Chen, Y., and Pine, K. H., (2019). The Emergence of New Data Work Occupations in Healthcare: The Case of Medical Scribes, *International Journal of Medical Informatics* (123): 76-83.

Bratteteig, T., and Wagner, I., (2016). Unpacking the Notion of Participation in Participatory Design, *Computer Supported Cooperative Work (CSCW)* (25:6): 425-475.

Eisenhardt, K. M., (1989). Building Theories from Case Study Research, *Academy of Management Review* (14:4): 532-550.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V. R., Tsou, A., Weingart, S., and Sugimoto, C. R., (2015). Big Data, Bigger Dilemmas: A Critical Review, *Journal of the Association for Information Science and Technology* (66:8): 1523-1545.

Emerson, R. M., Fretz, R. I., and Shaw, L. L., (2011). *Writing Ethnographic Fieldnotes*, (2nd ed.), Chicago: University Of Chicago Press.

Fitzgerald, M., (2016). Data-Driven City Management: A Close Look at Amsterdam's Smart City Initiative, *MIT Sloan Management Review,* Cambridge (57:4): 3-13.

Gregor, S., (2006). The Nature of Theory in Information Systems, *MIS Quarterly* (30:3): 611-642.

Grisot, M., Hanseth, O., and Thorseng, A. A., (2014). Innovation of, in, on infrastructures: articulating the role of architecture in information infrastructure evolution. *Journal of the Association for Information Systems* (15:4): Article 2.

Grisot, M., Moltubakk Kempton, A., Hagen, L., and Aanestad, M., (2019). Data-Work for Personalized Care: Examining Nurses' Practices in Remote Monitoring of Chronic Patients, *Health Informatics Journal* (25:3): 608 -616.

Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M., and Feldberg, F. (2017). Debating Big Data: A Literature Review on Realizing Value from Big Data, *The Journal of Strategic Information Systems* (26:3): 191-209.

Hanseth, O., Monteiro, E., and Hatling, M., (1996). Developing Information Infrastructure: The Tension Between Standardization and Flexibility, *Science, Technology & Human Values* (21:4): 407-426.

Henfridsson, O., Nandhakumar, J., Scarbrough, H., and Panourgias, N., (2018). Recombination in the Open-Ended Value Landscape of Digital Innovation, *Information and Organization* (28:2): 89-100.

Iannacci, F. (2010). When Is an Information Infrastructure? Investigating the Emergence of Public Sector Information Infrastructures, *European Journal of Information Systems* (19:1): 35-48.

Iivari, J., Isomäki, H., and Pekkola, S., (2010). The User—the Great Unknown of Systems Development: Reasons, Forms, Challenges, Experiences and Intellectual Contributions of User Involvement, *Information Systems Journal* (20:2): 109-117.

Kallinikos, J., Aaltonen, A., and Marton, A., (2013). The Ambivalent Ontology of Digital Artifacts, *Management Information Systems Quarterly* (37:2): 357-370.

Karasti, H., Baker, K. S., and Halkola, E., (2006). Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network, *Computer Supported Cooperative Work (CSCW)* (15:4): 321-358.

Khatri, V., and Brown, C. V., (2010). Designing Data Governance, *Communications of the ACM* (53:1): 148-152.

Kitchin, R., (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, SAGE, London.

Kitchin, R., (2017). Thinking Critically about and Researching Algorithms, *Information, Communication & Society* (20:1): 14-29.

Klein, H. K., and Myers, M. D., (1999). A Set of Principles for Conducting and Evaluating Interpretive Studies in Information Systems, *MIS Quarterly* (23:1): 67-94.

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

Leonelli, S., (2019). Data—from Objects to Assets, *Nature* (574:7778): 317-320.

Lohr, S., (2014). For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights, *The New York Times*. (https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html). Accessed 02 May 2020.

Mikalef, P., Pappas, I. O., Krogstie, J., and Pavlou, P. A., (2020). Big Data and Business Analytics: A Research Agenda for Realizing Business Value, *Information & Management* (57:1): Article 103237.

Mikalsen, M., and Monteiro, E., (2018). Data Handling in Knowledge Infrastructures: A Case Study from Oil Exploration, *CSCW: Proceedings of the ACM on Human-Computer Interaction,* (2:CSCW): Article 123.

Millerand, F., and Baker, K. S., (2010). Who Are the Users? Who Are the Developers? Webs of Users and Developers in the Development Process of a Technical Standard, *Information Systems Journal* (20:2): 137-161.

Monteiro, E., and Parmiggiani, E., (2019). Synthetic Knowing: The Politics of Internet of Things, *MIS Quarterly* (43:1): 167-184.

Monteiro, E., Pollock, N., Hanseth, O., and Williams, R., (2013). From Artefacts to Infrastructures, *Computer Supported Cooperative Work (CSCW)* (22:4-6): 575-607.

Myers, M. D., (1999). Investigating Information Systems with Ethnographic Research, *Communications of the AIS* (2): Article 23.

Nicolini, D., Gherardi, S., and Yanow, D., (2003). *Knowing in Organizations: A Practice-Based Approach*, M.E. Sharpe, Armonk, New York, and London.

Ofner, M. H., Otto, B., and Österle, H., (2012). Integrating a Data Quality Perspective into Business Process Management, *Business Process Management Journal* (18:6): 1036-1067.

Orlikowski, W. J., (2000). Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations, *Organization Science* (11:4): 404-428.

Orlikowski, W. J., (2002). Knowing in Practice: Enacting a Collective Capability in Distributed Organizing, *Organization Science* (13:3): 249-273.

Otto, B., (2011). Organizing Data Governance: Findings from the Telecommunications Industry and Consequences for Large Service Providers., *Communication of the AIS* (29:3): 45-66.

Parmiggiani, E., and Grisot, M., (2019). Data Infrastructures in the Public Sector: A Critical Research Agenda Rooted in Scandinavian IS Research, In: *Proceedings of the 10th Scandinavian Conference on Information Systems*, Article 13.

Parmiggiani, E., Karasti, H., Baker, K. S., and Botero, A., (2018). Politics in Environmental Research Infrastructure Formation: When Top-down Policy-Making Meets Bottom-up Fragmentation | Platypus, Platypus – The CASTAC Blog, , June 13. (http://blog.castac.org/2018/06/research-infrastructure/). Accessed 14 February 2019.

Passi, S., and Jackson, S. J., (2018). Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects, *Proc. ACM Hum.-Comput. Interact.* (2:CSCW): Article 136.

Pettigrew, A. M., (1990). Longitudinal Field Research on Change: Theory and Practice, *Organization Science* (1:3): 267-292.

Pipek, V., and Wulf, V., (2009). Infrastructuring: Toward an Integrated Perspective on the Design and Use of Information Technology, *Journal of the Association for Information Systems* (10:5): Article 6.

Plantin, J.-C., (2019). Data Cleaners for Pristine Datasets: Visibility and Invisibility of Data Processors in Social Science, *Science, Technology, & Human Values* (44:1): 52-73.

Ribes, D., and Finholt, T. A., (2009). The Long Now of Technology Infrastructure: Articulating Tensions in Development, *Journal of the Association for Information Systems* (10:5): Article 5.

Ribes, D., and Polk, J., (2014). Flexibility Relative to What? Change to Research Infrastructure, *Journal of the Association for Information Systems* (15:5); Article 1.

Ribes, D., and Polk, J. B., (2015). Organizing for Ontological Change: The Kernel of an AIDS Research Infrastructure, *Social Studies of Science* (45:2): 214-241.

Sirmon, D. G., Hitt, M. A., and Ireland, R. D., (2007). Managing Firm Resources in Dynamic Environments to Create Value: Looking Inside the Black Box, *Academy of Management Review* (32:1): 273-292.

Star, S. L., and Strauss, A., (1999). Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work, *Computer Supported Cooperative Work (CSCW)* (8:1-2): 9-30.

Tallon, P. P., (2013). Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost, *Computer* (46:6): 32-38.

Ure, J., Procter, R., Lin, Y., Hartswood, M., Anderson, S., Lloyd, S., Wardlaw, J., Gonzalez-Velez, H., and Ho, K., (2009). The Development of Data Infrastructures for EHealth: A Socio-Technical Perspective, *Journal of the Association for Information Systems* (10:5): Article 3.

Vassilakopoulou, P., Grisot, M., Jensen, T., Sellberg, N., Eltes, J., Thorseng, A., and Aanestad, M., (2017). Building National EHealth Platforms: The Challenge of Inclusiveness, In: *ICIS 2017 Proceedings.*

Venters, W., Oborn, E., and Barrett, M., (2014). A Trichordal Temporal Approach to Digital Coordination: The Sociomaterial Mangling of the CERN Grid, *Management Information Systems Quarterly* (38:3): 927-949.

Walsham, G., (2006). Doing Interpretive Research, *European Journal of Information Systems* (15:3): 320-330.

Ylinen, M., and Pekkola, S., (2018). Enterprise Architecture as a Scapegoat for Difficulties in Public Sector Organizational Transformation, In: *ICIS 2018 Proceedings*.