Becoming Sustainable Together: How Data Curators Support Large-Scale Digitalisation Initiatives

Nana Kwame Amagyei, Jostein Engesmo and Niki Panteli

September 5, 2023

BECOMING SUSTAINABLE TOGETHER: HOW DATA CURATORS SUPPORT LARGE-SCALE DIGITALISATION INITIATIVES

Research full-length paper

Amagyei, Nana Kwame, Norwegian University of Science and Technology, Trondheim, Norway, nana.k.h.amagyei@ntnu.no

Engesmo, Jostein, Norwegian University of Science and Technology, Trondheim, Norway, jostein.engesmo@ntnu.no

Panteli, Niki, Lancaster University, Lancaster, UK, & Norwegian University of Science and Technology, Trondheim, Norway, niki.panteli@ntnu.no

Abstract

The paper is driven by an interest to study the work practices of data curators who clean and preserve scientific data and support large-scale digitalisation and data accumulation initiatives. We posit that a neglect of this work could lead to the potential impact of such initiatives being over- or under- estimated. In this paper, we draw on a qualitative case study that examined the work practices of data curators who share scientific data openly and over extended time periods. Drawing from the practice lens perspective, we identify three data curation practices – characterising, augmenting, and liaising – as important work practices that explains how data curators support distributed and long-term digitalisation initiatives. Implications for theory and practice are discussed.

Keywords: Data Curation, Large-scale Digitalisation, Data Sustainability, Data Governance.

1 Introduction

> *"Imagine an environmental scientist, Sam, and his team. They monitor weather patterns all over Norway. They monitor temperature, precipitation, solar radiation, oceans, winds, clouds, and substances in the atmosphere. When Sam's team talks about climate, they often refer to weather patterns over a 30-year period. Through their research, they hope to contribute to a good climate, a healthy environment, and a sustainable society. The data they produce as part of their research are used by government agencies and researchers around the world. So, they curate data and check to see if there are things that might have affected their data and measurements. For example, there could have been a fire nearby or a malfunction of the data collection instrument. After this check, the approved results are entered into a database as digital representations. These digital representations of data will be shared with numerous environmental scientists in the years, decades, and centuries to come to improve scientific knowledge and understanding of climate.*
>
> *Assume 200 years later. Thanks to modern data storage technology, Sam's data and the millions of data provided by other environmental scientists from Sam's generation continue to be actively used; they are being reanalysed using the latest technological advances to develop*

> *insightful models and new data representations. Sam's data are becoming unexpectedly valuable, providing some interesting readings about climate. Environmental scientists are asking difficult questions in new ways, testing new hypotheses, and conducting scientific studies that will contribute to a better understanding of climate and improve the lives of Sam's grandchildren and those in the next generation."* – (Inspired by Jarvenpaa and Essén, 2023)

The need for large-scale digitalisation and data accumulation initiatives for breakthrough discoveries has been acknowledged by several researchers within the Information Systems (IS) field (Aaltonen et al., 2021; Cristina Alaimo & Kallinikos, 2022; Jarvenpaa & Markus, 2020; Mikalsen & Monteiro, 2021). Such data infrastructures are necessary when no single organization has the expertise, capacity, and financial resources necessary to address major societal challenges in climate, crime, and health (Jarvenpaa & Essén, 2023). At the same time, IS researchers are reminded that data in such data infrastructures should not be viewed as givens that are out there in the world, waiting to be collected and released but as contingent representations produced through situated practices by the people who conceptualize, prepare and care for the data in these data infrastructures (Jones, 2019; Eric Monteiro & Parmiggiani, 2019; Parmiggiani et al., 2022).

Digitalisation initiatives to accumulate data from distributed sources, on a large-scale and over long time periods then require attention to the different interorganizational contexts including the primary contributors and different organizational actors who prepare and care for data for their public release (Mikalsen & Monteiro, 2021; Parmiggiani et al., 2022; Parmiggiani & Grisot, 2020). As in our introductory example, data collected at one point in time can become valuable in relation to scientific discoveries and policy several hundred years later. Yet, this possibility required efforts to prepare and care for the data at a much earlier point in time – with continued investment and practices to keep the data infrastructure flexible and maintain data across human and technological generations (Ribes & Finholt, 2009; Ribes & Polk, 2014; Winter & Davidson, 2019).

The work of the people in charge of preparing and caring for data, to whom we refer as "data curators", is defined by a need to maintain data and serve a wide variety of prospective users across the globe – each looking for data suitable for their own interests and methods. Successful data curation for long-term release is marked by the extent to which data are accessed and reused within new research contexts. Notably, making data available online do not automatically make them usable (Leonelli, 2016). Tensions arise from concerns about – transmission of data over human and technological generations, ownership of the data, control and access, and the need for long-term sustainability of data (Ribes & Finholt, 2009; Ribes & Polk, 2014). Whether data are fruitfully adopted across contexts is then the result of strategies developed by data curators including how they collaborate and use data management technologies and their expertise to prepare and care for long-term data.

We adopt Chua and colleagues' (2022) definition of data curation as a data management activity that involves the development of data infrastructures to collect, index, store, and facilitate data access for future reuse (Chua et al., 2022). In this study, we take a closer look at this work on facilitating data access for future reuse. We do so in the context of environmental science (environmental science is an interdisciplinary field that combines physical, chemical, and biological sciences with social, political, and economic understanding needed to study the environment and address environmental problems). Data curation in environmental science data infrastructures comprise the work to prepare and care for data to persist and meet the needs of the present generation without compromising data reuse in the future. Our proposed view of data curation is not only consistent with, but also extends existing views in the IS literature. These views often conceptualize data curation as an activity to make data more meaningful, accurate, interpretable, and useful for further organizational and interorganizational uses (Leidner et al., 2022; Link et al., 2017; S. Seidel et al., 2013; V. P. Seidel et al., 2020). However, a focus on how scientific data from the past and present are preserved for future reuse seems an under-discussed aspect in IS that has implications for data governance and data sustainability more broadly, and to which our current work aims to contribute.

Against this background, we ask: *what data curation practices ensure that data in data infrastructures are sustained over extended time periods?* The purpose of this paper, then, is twofold: (i) to show how data curators address practical implications of large-scale digitalisation and data accumulation initiatives through their daily work practices (ii) to demonstrate the relevance of such work in discussions about data governance and data sustainability literatures. In so doing, this article is novel in several respects. First it brings together the growing, but still independent research streams on data curation, data governance and data sustainability literatures together. As these phenomena are expected to grow in importance, this article highlights the implications of linking them. Second the article brings a long-term perspective of data curation to the body of work on data infrastructures in IS. We identify three data curation practices – characterising, augmenting, and liaising – as important daily work done by data curators, which we believe contribute to discussions on data sustainability in large-scale and distributed data infrastructures.

The remainder of this article is organized as follows: In the next section, we present our theoretical lens of work practices. In this section, we also discuss our overarching concepts: data curation, data infrastructures and data sustainability from existing research in IS and neighbouring fields. We identify gaps that provide opportunities for further examination into data curation as a practice critical to studies of data infrastructures and data sustainability. We then present our research setting, methods, and findings. We discuss our findings with reference to extant literature, present our contribution to IS studies, and conclude with implications and future research opportunities.

2 Data curation: A Practice to ensure Data Sustainability in Data Infrastructures

We draw on Nicolini's (2012) pragmatic approach of practice as care (Nicolini, 2012). In this sense, data curation is understood as the practice of *caring for data* (cf. (H. Karasti et al., 2018)), where actors are driven by practical concerns that orient their daily work (Nicolini, 2012, p. 20). We draw on a practice lens that refers to the specific structure routinely enacted as data curators use specific technologies, machines, techniques, appliances, devices or gadgets in recurrent ways in everyday situated activities (Orlikowski, 2007). Practices entail open-ended spatial-temporal actions (Nicolini, 2012; Schatzki, 2001). A practice perspective in IS employs technology-in-use as the unit of analysis (Orlikowski, 2007) and organisational phenomena as the effects of interconnected material, discursive and social practices (Nicolini & Monteiro, 2017). Through their regularised engagement with a particular technology in particular ways and conditions, data curators repeatedly pay close attention to activities in producing reliable data by enacting a set of rules and resources which structure their ongoing interactions with the technology (Orlikowski, 1992).

A practice lens allows us to see data curation as the work of situated practices where data curators inscribe organising visions of data governance frameworks into actual technology use and enact change and innovation as they improvise and adjust their daily routines over time (Orlikowski, 2000). The lens is useful to understand the work needed to make technology function *in-situ* while enacting organising visions (Bechky, 2021; Jackson et al., 2014; Passi & Jackson, 2017). Data curation thus comprises the practices that cultivate and channel expectations of future users and maintain these visions in the face of a changing data infrastructure (Parmiggiani et al., 2022).

Focussing on the longitudinal dimension of this practice, highlights how curating data for public release becomes embedded in data infrastructures through alignment of complex local and political requirements in daily routine work (Steinhardt & Jackson, 2015). Ribes and Finholt (2019) show how a data curation orientation helps to understand and address emerging tensions related to work organization, available technologies, and funding and to ensure the long-term development of data infrastructures (Ribes & Finholt, 2009). Such perspective allows data curators engage with data and make important daily decisions and influence data infrastructure design and development (Parmiggiani & Grisot, 2020). Research has demonstrated the unstable and bidirectional nature of data curation work and its role as a necessary precursor for data analytics (Parmiggiani et al., 2022). In a related study, Ribes and Polk (2014) show how data curators at distributed sites managed data from donor cohorts for over 40 years. Data curation kept this data infrastructure flexible and adaptable to an evolving understanding of the complex nature of HIV/AIDS during this period (Ribes & Polk, 2014). Other scholars focus on data curation as requiring significant articulation work (Star & Strauss, 1999), which tends to remain invisible in streamlined digital transformation accounts and data governance models (Pine et al., 2022). Recent studies in the oil and gas domain have shown that data are inherently always 'cooked' by data curators to facilitate their release, use and reuse (Mikalsen & Monteiro, 2021).

Scholars in related fields who have studied data curation in data infrastructures highlight its critical role in supporting large-scale digitalisation and data accumulation initiatives. These studies consider data sharing initiatives as a boundary condition for curation (Leonelli, 2014). Since such initiatives require significant curation work and human agency to prepare and steward data for public release (Karasti et al., 2006). These studies demonstrate that curating data for public release often clashes with data curators' existing research practices and cultures (H. Karasti et al., 2006). This leads to tensions among data curators and result in significant additional work – as data curators do not only have to share their primary data but must further ensure that these data are meaningful to prospective users several decades later. Despite these tensions, data curation activities that support large-scale digitalisation and data accumulation initiatives are rarely formally recognized in policies for governing such data sharing infrastructures (Vassilakopoulou et al., 2019; Winter & Davidson, 2019).

In today's era of digital transformation, where policy calls for making data openly accessible, this practice is becoming increasingly urgent (Bossen et al., 2016; Leonelli, 2016; Parmiggiani et al., 2022; Pine et al., 2022). In this paper, we define data infrastructures as multi-layered, long-term, and interconnecting heterogeneous agendas and systems with inputs from and outputs to many different independent actors (Ciborra and Hanseth 1998; Monteiro et al. 2013). From this perspective, data are relational objects – whose production, use and long-term sustainability depends on the interests, goals and motives of the people involved, and their institutional and financial context (Gitelman & Jackson, 2013; Leonelli, 2019).

We define data sustainability as the ability of data in data infrastructures to persist across human and technological generations. This capacity includes the potential of data to transcend technological and social arrangements and enable knowledge advancements beyond current issues (Jarvenpaa & Essén, 2023). Data sustainability draws from, but differs from, the definition of sustainability in the IS literature, where sustainability is often defined as "the capacity to endure" (Davidson, 2014) and as "development that meets the needs of the present without compromising the ability of future generations to meet their own needs" (Brundtland et al., 1987). Data sustainability is a concept that anchors ongoing data curation work in data infrastructures explicitly on expected futures and remembered pasts (Jarvenpaa & Essén, 2023). Suggesting that data curation in data infrastructures is an ongoing practice with a temporal horizon.

In distributed and longitudinal data infrastructures, prospective users ask questions about the context in which scientific data were created and processed (Leonelli, 2014). Data curators pay attention to strategies for releasing scientific data for public reuse, including approaches to document information about changes in the field, technological configurations and calibration of sensors (Borgman et al., 2020; H. Karasti et al., 2006; Leonelli, 2014). These data curation practices play an important role in ensuring that the data released are reliable to prospective users (Leonelli, 2016). Reliable data are thus not an intrinsic property of the data, but a collaborative and situational achievement in which data curators temporarily resolve emerging tensions in pragmatic ways (Passi & Jackson, 2018).

Data curation studies in data infrastructures unearth several challenges relating to governance of data in large-scale digital transformation; including but not limited to data ownership issues, inadequate incentives for researchers to release their data, changes to technologies and human personnel, technical hurdles related to incompatible hardware, software, and data structures, and costs associated with documenting, releasing, and storing data (Aaltonen et al., 2021; Loebbecke & Picot, 2015; Parmiggiani & Grisot, 2020; Ribes & Polk, 2014). The link between data curation and data governance continue to surge. This strand of literature presents data curation as an activity that is not defined in advance in data governance frameworks but emerges as data curators work with data daily (Parmiggiani & Grisot, 2020). This makes it more difficult for researchers to understand its full scope and for policy makers to fully consider this work up front. IS researchers are therefore encouraged to study such phenomena that evolve as people engage with it (Monteiro et al. 2022).

The data governance literature in IS largely builds on and extends the normative data governance framework of Khatri and Brown (2010).The influential framework outlines five areas where decisions should be made and for which people in the organization should be responsible, these include: (1) data principles, which clarify the role of data as an asset to the organization; (2) domain decisions, which relate to how data quality is assessed; (3) metadata, which clarifies how to define the semantics of data and how to define data consistently and continuously so that they are interpretable; (4) data access, which clarifies what standards and procedures apply to data access and how compliance is monitored; and (5) data lifecycle, which relates to determining how data are retained and retired (Khatri & Brown, 2010). Scholarly work in IS and related fields has extended the normative data governance framework by Khatri and Brown (2010). This strand of literature recognizes that data producers and data users are often dispersed across different organizational settings and have different power and control relationships (Abraham et al., 2019; Alhassan et al., 2016; Zuboff, 2015). Due to this, IS scholars acknowledge the need for adapting current data governance frameworks (Jarvenpaa & Essén, 2023).

An emerging line of research has begun to recognize not only the different interests in the processes of data collection and use, but also their different assumptions, tensions, and conflicts that are evident in daily data management to achieve quality, filter relevant data, and ensure privacy (Parmiggiani & Grisot, 2020). These studies view data not as a raw representation of reality and a precursor to knowledge production and use, but as actively constructed (C. Alaimo et al., 2020; Cristina Alaimo & Kallinikos, 2022; Jones, 2019; Leonelli & Tempini, 2020; Mikalsen & Monteiro, 2021).

Extant data governance studies have deepened our understanding of how data are created and used in practice, including the necessary physical and logical infrastructures, resources, and challenges in coordinating, and aligning the interests of many different stakeholders. Yet neither the data governance studies, nor the data curation studies have explicitly discussed data sustainability challenges that may arise in allowing data to endure across technological and human generations. Hence, we know little in IS about how to address the needs and interests of future data (re)users. Such discussions in IS are still relevant (Jarvenpaa & Essén, 2023; Eric Monteiro & Parmiggiani, 2019; Parmiggiani et al., 2022; Parmiggiani & Grisot, 2020). By focussing on the activities and work practices of data curators who prepare and care for long-term scientific data, we attempt to systematically articulate the commitments of data curators and their role in large-scale digitalisation and data accumulation initiatives.

In the next section, we introduce our research setting, which provided a case for this study and then present our methods.

3 Research setting

The European Long Term Ecological Research network in Norway (eLTER Norway) is the national node for long-term environmental science studies in Norway. It has many local research institutes that monitor different aspects of the environment, and thus produce and use different types of data, including data on water, air, temperature, tree diameter, animals and so on. Data are collected both manually through observations and recordings by scientists, and automatically, through satellites, drones, IoT sensor devices, etc. Local research institutes in eLTER-Norway are spread across the country and adhere to the European Commission open data policy. There are different professionals at a given institute, each of whom perform some form of data curation. For example, *data collection* may be performed by biologists, chemists, physicists, zoologists, and environmentalists; *data cleaning* may be performed by data scientists and engineers; *database administration* may be performed by data administrators and software engineers; *data updates and analysis* may be performed by students and ecologists; *data centre and equipment maintenance* may be performed by craftsmen. For this reason, we put all these professionals in our data as "data curators" to ensure consistency.

In 2002, the European Commission created a coordinated mechanism to strategically connect and integrate data from distributed local and national research institutes in European countries. This mechanism is called the European Strategy Forum for Research Infrastructures (ESFRI). ESFRI regularly publishes and updates a policy roadmap that reflects the strategic objectives of the European Commission's open data policy agenda. Despite the crucial role that data and data curation play in data infrastructures, the policy roadmap has little visibility of data curation work in the 'top-condensed' ESFRI open data model.

Data originating from distributed research institutes are recognized in the model when there is a need for data to flow into the central node(s) of the data infrastructure. This invisible data flow represents the expectation that data curation practices are in place at local and national research institutes and that data curators are able to produce and share good quality data for the ESFRI Open Data Infrastructure. This observation seems to be a particularly large assumption made. Existing local and national data infrastructure efforts are very heterogeneous and dispersed, and there is a lack of coordination both at the local research institute level and among national research institutes.

Therefore, we set out to interview and observe data curators in the eLTER Norway network. We chose this setting because, the eLTER network was included in the ESFRI roadmap in 2018. ESFRI thus provides a policy framework for eLTER, and these priorities are mirrored at national and local European eLTER research institutes. eLTER has several nodes spread across Europe and we observed data curators in three local institutes in the Norwegian node: two in Trondheim and one in Oslo (names omitted for anonymity). In what follows, we present our findings (from local data curators, to understand their data curation practices that ensure data are preserved in the eLTER data sharing infrastructure over time).

4 Findings

4.1 Characterising – Practices to avoid or reduce data loss.

Data curators utilize new sensing technologies to capture environmental phenomenon which are otherwise difficult or impossible to capture manually. This is because, "*it is unfeasible to do everything manually, and this is where remote sensing comes into play. We go to selected spots and collect field data at some places. We sometimes use comparison methods to fill the gaps in a more timely fashion in cases where there are missing data" (Data curator 1, interview).* In some cases, sensing technologies cannot be deployed due to high cost of sensors. This requires data curators to improvise because *"...with fresh water, the sensors that are available are expensive, and you don't necessarily put out a lot of those" (Data Curator 2, field notes).* In other cases where sensor-based techniques are largely employed, human interaction is still needed to distinguish, say, which type of plant species has been captured by the sensing technology, because the technologies are limited in providing such details. According to one of the participants, *"as soon as you need an interpretation of what type of plant species has been captured by the sensor, that is something that requires human interaction. At least for plants when you determine species, you need to have humans that collect the* *data to interpret the data through appropriate documentation" (Data curator 2, interview). "Sometimes acoustic devices cannot distinguish the species of bird by their sound. And in an environment where there are different species of bird making similar sounds human interpretation is required" (Data curator 17, interview).* Beyond these practices designing sensing solutions rely on highly efficient sensors: "*…research areas like wildlife tracking rely on sensors to be effective" (Data curator 5, field notes).*

To ensure improved sensor efficiency, then, data curators develop practices to determine physical properties of data collection technologies and ensure that data are not lost, or that captured data are successfully sent to appropriate data repositories: "*if none of the data must get lost, then people would use techniques [through which] they can send the data over" (Data curator 14, interview).* In other cases, manual effort is required to ensure that the sensors function as intended *"...you can insert a SIM-card, to make sure that it keeps sending the data but you have to check on it regularly" (Data curator 4, interview).* This practice is to ensure that data from sensing technologies continue to send reliable data to data repositories. Technologies may produce a significant volume of data however their use is further complicated by device malfunction, which require characterising the form of human intervention needed to address this. Neglecting this practice may result in missing values, quality issues, and discrepancies in the methods and ways in which data are collected.

The use of different approaches to capture environmental phenomenon also comes with it the need to preserve and store data. This poses the issue of developing appropriate routines necessary for storing and using some of the data for immediate research purposes while ensuring that as much collected data are cleaned for public dissemination. This concern is raised by one data curator who asserts that *"[With] large amounts of data, storage is a challenge. So you have to build good routines to avoid storing data that you don't need. [...] Some people have routines where they fetch data from specific times they are interested in, process it, and delete it afterward. Then they keep the results to minimize the storage requirements" (Data Curator 3, interview).* Not having the right tools to process and share the significant variety in data can limit data curators in sharing as much meaningful data publicly as possible *"...when you start using sensor-based data, then you need to have the technology to store the data and analyze the data, that can prevent you from sharing all the data" (Data Curator 6, field notes).*

These different data collection techniques suggest that environmental science data are heterogeneous. Data curators learn to characterize the different forms of management requirements for data collection technologies to ensure that data are reliable and accurate for further reuses. Such practices include performing regular sensor checks, replicating sensors, scheduling repair and maintenance, following up on alerts by sensors or other automated alert systems. Characterising as a data curation practice then highlights that data curators consider the physical properties of data collection technologies and characterise their present and future management needs. This characterization is done in terms of the scientific data being collected and aims to avoid or reduce data loss and improve data veracity for secondary users. It often requires appropriate documentation, awareness of different tools and methods for data collection, and adopting on-the-go data management routines.

4.2 Augmenting – Practices to identify data problems.

Data curators often use well-defined data management models which allows for curation techniques to be automated. For example, to identify inaccurate data, Espen a data curator tells us, *"we have an automated procedure that detects missing values in the database and flag them or notify us. We go into the database and determine why there is a missing value in there and correct them if it is possible" (Data Curator 13, field notes).* We understood from our interaction with Espen that human labour is unavoidable in automated processes, which often require decisions on whether to remove, adjust, or replace data with an estimated value. There exist instances when a data value that is real, but extreme, may automatically be flagged as problematic because it lies outside a programmed or expected range. To ensure that this does not happen, data that are automatically flagged as problematic are reviewed carefully by an expert. Eva informs us that *"sometimes the cameras show certain blurry images that our machine learning program flags as not lynx, but when you review the image, you know that this is a lynx and the program got it wrong." (Data Curation 12, field notes)*. This suggests that the risk of ignoring potentially important data to be shared reliably for public audiences is high if human agency to curate the data is neglected.

Automated techniques are usually programmed to detect valid data and reject invalid data. Emma highlights how such machine learning algorithms support cleaning and organizing data obtained from sensors for monitoring lynx.: *"we use this machine learning algorithm (pointing to her computer screen) which detects the lynx". (Data Curation 10, field notes).*

: *""we use the data for our own publications, and we are also doing data wrangling because we have to share the data" (Data Curator 13, field notes)*. When asked about the importance of this practice, Emma mentioned to us that repeating the practice of verify data errors (where correct data are marked as errors or other data representations besides their true nature) are useful to validate outcomes of automated approaches and minimized human error: *"after students have verified output of the machine learning algorithm, we also validate the work of the students, we document all these, and then we communicate it to the data science guys who improve the algorithm" (Data Curator 10, field notes*). Judy, who monitors air temperature in the Arctic also mentioned how her team performs range checks to ensure that data fall within a known upper and lower bound: *"sometimes we know that the data should fall within a range for example relative humidity is always between 0% and 100%, any value outside this range shows that there is an error in the data, sometimes also, the bounds are not absolute, so we have to check historical or long-term data that usually show data on extreme values which are useful to set appropriate bounds." (Data curator, 7, field notes).* In cases where obtaining data require two or more methods or devices, data curators can compare the different data sets to identify and resolve missing data values: *"…after data from the cameras are downloaded to the computer, we use the date and time from the cameras, which helps us to check errors like sometimes we see the lynx better on the image captured by the other camera…"*. She continued to tell us: *"…so we compare the date and time from the cameras, which helps us to check and correct errors made by the machine learning algorithm" (Data curator, 10, field notes).*

When the same data value repeats continuously in a database this may indicate a sensor failure and hence problematic data. Espen tells us *"…the speed of wind changes continuously so when the same value is recorded repeatedly then we know a problem has occurred" (Data Curator 13, field notes).* In some cases, automated approaches may fail to determine inconsistent data, in such cases, consistency checks are important to clean and improve the data. Consistency checks are performed by data curators to evaluate the differences between related parameters, such as ensuring that the minimum water depth is less than the maximum water depth: "*sometimes the depth of the water measurements creates suspicions, maybe the sensors are not submerged deep enough" (Data curator 2, interview).*

Secondary or potential users of shared data often expect to determine whether data obtained from online databases are reliable or of good quality for their purposes. In such cases, they require metadata – or data that provides information about details of the data including the source of the online data, the sensors used, and changes to the methods that were used to collect the data. *"We describe the data with metadata. The purpose of metadata is that somebody else should be able to take that data set understand enough about how that data have been collected the kind of sensor used, for instance, if it's sensor data, the kind of settings used in the sensor, so that they are able to use it again without the special knowledge of the person who actually collected it." (Data curator 3, interview)*.

When digital data – data stored on computers as bits – contain issues such as inconsistent or missing values, these issues challenge their further reuses such. Through augmenting practices data curators make it possible to identify and resolve problems in shared digital data. These practices are numerous and may include providing metadata or adding contextual information to data, correcting or filling in missing data, verifying automated techniques, comparing data from related sensors, performing range checks on numerical data, and performing persistence checks on continuous data. Augmenting as a data curation practice then highlights that data curators identify data issues in digital data to complement immediate scientific work and future data release with accurate and interpretable primary data.

4.3 Liaising – Practices to learn from different data management cultures.

Based on the realities and needs of local institutes, eLTER has created network-level data curation groups to ensure, organize, and enforce effective implementation of eLTER's Open Data infrastructure. One of such groups is the Data Managers Assembly (DMA). Its main role is to decide on the basic configuration of the eLTER data infrastructure, including networking with external partners, integrating the infrastructure into other national and international programs, and the strategic planning of future data activities. In this sense, the DMA forms a group of people who liaise to share a common concern related to data curation and who come together to fulfil both local-level and network-level data management goals.

Periodically, at least once a year, a meeting of data curators is held to discuss current eLTER data curation issues and to discuss related topics. Data curators bring local experiences to their network-level activities and learn from each other on issues of local-level data sharing, provision, standards, discovery, and use. The group plans outreach activities related to knowledge transfer and information product development. *"A critical point for the success of the data managers group was the recognition that there were legitimate reasons for some differences between the systems at each site. [...] As far as the group is concerned, you also accept views that are different from your own. There are a variety of approaches between sites, and there is strength in diversity. The goal is to implement and develop the eLTER data infrastructure together." (Data curator 9, interview)*.

Data sharing issues through community-wide work on the eLTER data infrastructure provides an opportunity to develop a flexible and enduring eLTER RI: *"You are always connected, affiliated, associated with the community, and that gives the eLTER open data infrastructure that continuity." (Data Curator 8, interview)*. A flexible and enduring eLTER data infrastructure builds the trust needed within the data management community to interact regularly and maintain reciprocity. *"We do not have to rebuild every time, we have built trust. It's good to see how other sites are doing, either as a contrast or as a suggestion for improvement." (Data curator 10, interview)*. Such collaborative data infrastructure effort provides a reliable avenue for learning and sharing ideas related to data curation. This helps data curators to acquire knowledge collaboratively. *"It's safe to say things that show examples where you have not been as successful, or disappointments, even how they plan funding of data management activities. Once you can do that in a group, there is a bond. The group has taken the time to promote an inclusive, sustainable approach to technology and to make sure that we learn together."(Data curator 10, interview).*

The network-level working group provides a space for collaborative work on the data infrastructure. The community welcomes and is willing to consider any potential discoveries of technologies appropriate to the field of environmental monitoring research and long-term data management. It also demonstrates how technological heterogeneity at the local-level has not only been accommodated but transformed into a shared resource of proven technology experience through network-wide efforts in which each local institutes becomes a "workshop" with its own local characteristics. Most importantly, it demonstrates a principle widely recognized in the eLTER network that *"many of the good ideas that work also come from below, not just from above" (Data Curator 11, interview)*. The collaborative work of data curators demonstrates an enduring, ongoing, and open relationship between the local practices of data curators and the eLTER network to collectively address ongoing and long-term data sharing issues.

Data curators participate in network-level forums and awareness-raising, training, and education workshops. Such liaisons are based on the realities and needs of each local institute and reflect the history and specificities of local science conduct and data management. This practice provides the opportunity to develop an eLTER data sharing community with continuity. It seeks to provide a reliable place for exchange and reciprocity, to see how other local institutes are managing organizational data, either as a contrast or as a stimulus for improvement. Liaising as a data curation practice then highlights that data curators' network with colleagues who are involved in different forms of data management and science conduct cultures. The practice aims to align different data management cultures and science research with the growing eLTER data infrastructure.

5 Discussion and conclusion

This paper advances the discourse on data curation as a practice that strives to ensure data sustainability in data infrastructures – an important, though relatively under-discussed, aspect of data curation in IS literature. We define data curation as a data management activity involving the development of both physical and logical infrastructures that facilitates data access for subsequent analysis and enables data to be collected, indexed, and stored (Chua et al., 2022). We follow Jarvenpaa and Essén's (2023) definition of data sustainability as the ability of data to endure across technological and human generations (Jarvenpaa & Essén, 2023). In our study, we propose that distributed and longitudinal digital transformation initiatives prompt for a need to articulate the data curation work that is done in the present with the goal of preserving the representational capacity of data for prospective users several decades and centuries in the future. We suggest that well-articulated data curation practices are critical to data sustainability as they aim to produce accurate and interpretable data that can be easily understood and reused by prospective users.

To understand the extent of the data curation work needed to ensure data sustainability in large-scale digitalisation and data accumulation initiatives, we adopted the theoretical lens of practices (Jones, 2019; Nicolini & Monteiro, 2017; Orlikowski & Scott, 2016; Schatzki, 2001). For theoretical purposes, IS researchers are encouraged to study phenomena that mutate and evolve over time (Bailey et al., 2022; Eric; Monteiro et al., 2022). For practical purposes, a major concern for the European Commission's governing agencies on large-scale digitalisation and data accumulation initiatives, is the realisation that "*it is vital that the common elements of local research institute governance, their funding and management guarantee long-term sustainability*" (ESFRI, 2021). Against this background, we set out to understand how data curators at three environmental science research institutes in Norway handle data to make them accurate and interpretable for public release. We asked *what data curation practices ensure that data in data infrastructures are sustained over extended time periods?*

We found that most data curators take responsibility for releasing their primary data into public repositories for other researchers and the public to download and use freely. Yet, there were several other practical considerations that made this public release possible. First, the properties of data production technologies: including what materials the technologies are made of, where they are positioned for data collection, and for how long, require a characterization of the different practices for maintaining these devices. Data curators therefore assess the physical properties of data collection technologies in order to characterize their different management needs. Data curators may replicate sensors, perform regular checks, schedule repair and maintenance, follow-up on automated alerts or replace parts. These practices aim to avoid or reduce data loss and improve data veracity.

Further the study provides insights into how new digital transformation initiatives to accumulate scientific data on a distributed and large-scale have created new and contradictory forms of work for data curators. Such initiatives have led to changes relating to how data curators conduct science research – which is often recognized as part of their job descriptions and determines their scientific career successes; while simultaneously managing primary data – which is rarely recognized as an important part of their scientific careers. This represents a significant augmentation of responsibilities – moving from the need to understand and synthesize data within scientists' short-term career timeframes, to requirements for contextualization and preservation of primary data for reuse over much longer timeframes. We refer to this practice of data curation as augmentation because data curators complement immediate scientific work and future data release with data management efforts to produce accurate and interpretable primary data ledge. These augmentation practices aim to identify problems in digital data and include, but not limited, to providing metadata, correcting or filling in missing data, verifying outcomes of automated techniques, comparing data from related sensors, and performing several checks on digital data including range checks on numerical data and persistence checks on continuous data.

We also found that another crucial practice for data curators was participating in network-level activities to learn together and share data curation lessons. This practice is intended for data curators to understand the different data curation cultures. It is an important practice because it particularly supports agendas that are trying to find common approaches to distributed and large-scale data governance and ensure the sustainability of data. We term this data curation practice as liaising, because data curators within the network have access to a community of whose interest is to share data management lessons and align data sharing and science research with the growing data infrastructure.

Our findings are in line with existing IS studies on data as a relational phenomenon shaped by their provenance – that is, by the methods, procedures, and technologies used to generate, clean, and disseminate the data (Barley & Bechky, 1994; Mikalsen & Monteiro, 2021; Parmiggiani et al., 2022; Porter, 1996). Yet, in our case issues of data sustainability and continuity repeatedly come to the fore in our analysis. We suggest that ongoing digital transformation is likely to face challenging issues, especially in terms of situated data curation practices. Data curation practices such as characterizing, augmenting, and liaising introduces a long-term perspective in conversations about data governance in data infrastructures. Our work therefore contributes a temporal perspective of data curation as a governance practice in data infrastructures. This study points to the need to look beyond new technologies, data curation models, and data governance frameworks and discuss data curation as not only important for short-term purposes, but equally important is its role in the long-term sustainability of data in data infrastructures. Including how data are made accurate and interpretable for secondary users in the distant future of technologies and people. Our perspective sensitizes us to the need to ask questions about these distant futures and consider them, or at least remind ourselves not to ignore them.

5.1 Practical Implications and Future Research Directions

Our study has implications for data governance in distributed data infrastructures. Data governance in environmental science is undergoing significant changes as new technologies for data collection and processing become available and funding agencies push for standardized and large-scale data accumulation initiatives. Data curators at research institutions must deal with these new technologies and policies as part of their daily data curation practices. Our case demonstrates that data governance in environmental science is always an ongoing process, and at the same time our understanding of it is nascent and incomplete. The metaphor of 'researching on a ship that is being built while it is in operation' (Helena Karasti et al., 2010; Ribes & Polk, 2014) remains an appropriate image for thinking about the data curation activities that lie ahead for governing agencies and other relevant stakeholders of such initiatives.

At the local level – in distributed data infrastructures – there are many interesting directions to continue this kind of work to understand common elements of local data curation activities. This should lead to more clarity in assigning roles and responsibilities regarding data management. This is necessary to make visible what exists, give it a label and recognition, and more accurately identify missing components.

Building a data sharing community that is focused on this nature of data curation can allow people to consolidate a common vocabulary, share their collective and distributed expertise in more explicit ways, and maintain the documentation and development work that needs to be done on the ground within the community. Interdisciplinary work would be needed to address the complexities and dynamics of establishing a functional, and stable environment for data curation and sustaining data in the context of a larger data infrastructure. Therefore, it would be important for stakeholders of data sharing communities to take explicit steps to raise awareness of the importance of data curation in open data infrastructure development. The message should clarify that not only is it important, but that data curation requires resources, investments, and funding at many levels and that this is not currently supported by research funding or infrastructure sources.

In addition, gaps need to be identified and bridged. The gap between funding agencies and data curation needs of national and local research stations needs to be bridged. Our context in Norway is certainly not the only member country that does not yet have data management practices capable of producing high quality data for sharing and preservation. Member nations involved in large-scale digitalisation and data accumulation initiatives must figure out how to address and support local data management on their own. Devising processes to coordinate data activities and ensure interoperability of data would be imperative.

Given data sharing policies, organizations are encouraged to not only develop frameworks and strategies to invest in technologies or data management plans, but also to focus on the actual data handling practices of employees. These may include the methods, capabilities, and knowledge that are developed now and, in the future to handle data (Leonelli, 2019). This may provide organizations the chance to understand employee's perspective on such policies and adopt an approach to organizational data governance that is targeted more toward collaborative data governance, as everyone's responsibility. It can also lead to faithfully capturing and representing the complex, diverse, and evolving structures, behaviours, and cultures within the organizations, which can support ambitions toward managing common elements of different data management cultures.

Organizations are encouraged to schedule time and identify the local context in which data management activities occur so they can learn from mistakes and help shape data-related projects and outcomes. Organizations are also encouraged to be aware of diverse stakeholder groups, recognize the importance of mutual respect for these roles, and find opportunities to learn about and empower marginalized groups.

Future work will explore the anticipatory nature of this data curation work, including what data curation practices can anticipate and ensure that data in data infrastructures are ready for known and unknown future uses. Such known future uses may include academic publications, research and climate studies, and unknown future uses may include secondary uses that cannot be predicted in advance such as complex technological analysis and new programmatic advertising and other intelligent public service.

References

Aaltonen, A., Alaimo, C., & Kallinikos, J. (2021). The Making of Data Commodities: Data Analytics as an Embedded Process. *Journal of Management Information Systems*, *38*(2), 401–429. https://doi.org/10.1080/07421222.2021.1912928

Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, *49*, 424–438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

Alaimo, C., Kallinikos, J., & Aaltonen, A. (2020). Data and value. In K. Lyytinen, S. Nambisan, & Y. Yoo (Eds.), *Handbook of Digital Innovation*. Edward Elgar Publishing.

Alaimo, Cristina, & Kallinikos, J. (2022). Organizations Decentered: Data Objects, Technology and Knowledge. *Organization Science*, *33*(1), 19–37. https://doi.org/10.1287/orsc.2021.1552

Alhassan, I., Sammon, D., & Daly, M. (2016). Data governance activities: an analysis of the literature. *Journal of Decision Systems*. https://doi.org/10.1080/12460125.2016.1187397

Bailey, D. E., Faraj, S., Hinds, P. J., Leonardi, P. M., & von Krogh, G. (2022). We Are All Theorists of Technology Now: A Relational Perspective on Emerging Technology and Organizing. *Organization Science*, *33*(1), 1–18. https://doi.org/10.1287/orsc.2021.1562

Barley, S. R., & Bechky, B. A. (1994). In the Backrooms of Science. *Work and Occupations*, *21*(1), 85–126. https://doi.org/10.1177/0730888494021001004

Bechky, B. A. (2021). *Blood, Powder, and Residue: How Crime Labs Translate Evidence into Proof*. Princeton University Press.

Borgman, C. L., Wofford, M. F., Golshan, M. S., Darch, P. T., & Scroggins, M. J. (2020). *Collaborative ethnography at scale: reflections on 20 years of data integration*.

Bossen, C., Pine, K. H., Elllingsen, G., & Cabitza, F. (2016). Data-work in healthcare: The new work ecologies of healthcare infrastructures. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, *26-Februar*(1), 509–514. https://doi.org/10.1145/2818052.2855505

Brundtland, G. H., Khalid, M., Agnelli, S., Al-Athel, S. A., Chidzero, B. J. N. Y., Fadika, L. M., & Al., V. H. et. (1987). *Our common future; by world commission on environment and development*. Oxford Univesity Press.

Ciborra, C. U., & Hanseth, O. (1998). From tool to Gestell: Agendas for managing the information infrastructure. *Information Technology & People*, *11*(4), 305–327. https://doi.org/10.1108/09593849810246129

Davidson, K. (2014). A Typology to Categorize the Ideologies of Actors in the Sustainable Development Debate. *Sustainable Development*, *22*(1), 1–14. https://doi.org/10.1002/sd.520

ESFRI. (2021). *Roadmap 2021 Strategy Report on Research Infrastructures*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://roadmap2021.esfri.eu/media/1295/esfri-roadmap-2021.pdf

Gitelman, L., & Jackson, V. (2013). Introduction. In *“Raw data” is an oxymoron*.

Jackson, S. J., Gillespie, T., & Payette, S. (2014). The Policy Knot: Re-integrating Policy, Practice and Design in CSCW Studies of Social Computing. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW ’14*, 588–602. https://doi.org/10.1145/2531602.2531674

Jarvenpaa, S. L., & Essén, A. (2023). Data sustainability: Data governance in data infrastructures across technological and human generations. *Information and Organization*, *33*(1), 100449. https://doi.org/10.1016/j.infoandorg.2023.100449

Jarvenpaa, S. L., & Markus, M. L. (2020). *Data Sourcing and Data Partnerships: Opportunities for IS Sourcing Research* (pp. 61–79). https://doi.org/10.1007/978-3-030-45819-5_4

Jones, M. (2019). What we talk about when we talk about (big) data. *Journal of Strategic Information Systems*. https://doi.org/10.1016/j.jsis.2018.10.005

Karasti, H., Baker, K. S., & 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*, *15*(4), 321–358. https://doi.org/10.1007/s10606-006-9023-2

Karasti, H., Botero, A., Baker, K. S., & Parmiggiani, E. (2018). *Little Data, Big Data, No Data? Data Management in the Era of Research Infrastructures.*

Karasti, Helena, Baker, K. S., & Millerand, F. (2010). Infrastructure Time: Long-term Matters in Collaborative Development. *Computer Supported Cooperative Work (CSCW)*, *19*(3–4), 377–415. https://doi.org/10.1007/s10606-010-9113-z

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*. https://doi.org/10.1145/1629175.1629210

Leidner, D., Sutanto, J., & Goutas, L. (2022). Multifarious Roles and Conflicts on an Interorganizational Green IS. *MIS Quarterly*, *46*(1), 591–608. https://doi.org/10.25300/MISQ/2022/15116

Leonelli, S. (2014). What difference does quantity make? On the epistemology of Big Data in biology. *Big Data and Society*. https://doi.org/10.1177/2053951714534395

Leonelli, S. (2016). *Data-Centri Biology: A philosophical study*. University of Chicago Press. https://doi.org/https://doi.org/10.7208/9780226416502

Leonelli, S. (2019). The challenges of big data biology. *ELife*, *8*. https://doi.org/10.7554/eLife.47381

Leonelli, S., & Tempini, N. (2020). *Data Journeys in the Sciences* (S. Leonelli & N. Tempini (eds.)). Springer International Publishing. https://doi.org/10.1007/978-3-030-37177-7

Link, G., Lumbard, K., Germonprez, M., Conboy, K., & Feller, J. (2017). Contemporary Issues of Open Data in

Information Systems Research: Considerations and Recommendations. *Communications of the Association for Information Systems*, *41*, 587–610. https://doi.org/10.17705/1CAIS.04125

Loebbecke, C., & Picot, A. (2015). Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda. *Journal of Strategic Information Systems*. https://doi.org/10.1016/j.jsis.2015.08.002

Mikalsen, M., & Monteiro, E. (2021). Acting with inherently uncertain data: practices of data-centric knowing. *Journal of the AIS*, *22*(6), 1–21. https://doi.org/https://doi.org/10.17705/1jais.00722

Monteiro, Eric;, Constantinides, P., Scott, S., Shaikh, M., & Burton-Jones, A. (2022). Editor's Comments: Qualitative Research Methods in Information Systems: A Call for Phenomenon-Focused Problematization. *MIS Quarterly*, *46*(4), iii-xix.

Monteiro, Eric, & Parmiggiani, E. (2019). Synthetic Knowing: The Politics of the Internet of Things. *MIS Quarterly*, *43*(1), 167–184. https://doi.org/10.25300/MISQ/2019/13799

Monteiro, Eric, Pollock, N., Hanseth, O., & Williams, R. (2013). From artefacts to infrastructures. *Computer Supported Cooperative Work: CSCW: An International Journal*. https://doi.org/10.1007/s10606-012-9167-1

Nicolini, D. (2012). *Practice theory, work, and organization: An introduction*. OUP Oxford.

Nicolini, D., & Monteiro, P. (2017). The Practice Approach: For a Praxeology of Organisational and Management Studies. In A. Langley & H. Tsoukas (Eds.), *The SAGE Handbook of Process Organization Studies* (1st ed., pp. 110–126). SAGE Publications.

Orlikowski, W. J. (2000). Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations. *Organization Science*, *11*(4), 404–428. https://doi.org/10.1287/orsc.11.4.404.14600

Orlikowski, W. J. (2007). Sociomaterial Practices: Exploring Technology at Work. *Organization Studies*, *28*(9), 1435–1448. https://doi.org/10.1177/0170840607081138

Orlikowski, W. J. (1992). Learning from Notes. *Proceedings of the 1992 ACM Conference on Computer-Supported Cooperative Work - CSCW '92*, 362–369. https://doi.org/10.1145/143457.143549

Orlikowski, W. J., & Scott, S. V. (2016). Digital Work: A Research Agenda. In B. Czarniawska (Ed.), *A Research Agenda for Management and Organization Studies* (pp. 88–96). Edward Elgar Publishing. http://hdl.handle.net/1721.1/108411

Parmiggiani, E., & Grisot, M. (2020). Data Curation as Governance Practice. *Scandinavian Journal of Information Systems*, *32*(1).

Parmiggiani, E., Østerlie, T., & Almklov, P. G. (2022). In the Backrooms of Data Science. *Journal of the Association for Information Systems*, *23*(1), 139–164. https://doi.org/10.17705/1jais.00718

Passi, S., & Jackson, S. J. (2017). Data vision: Learning to see through algorithmic abstraction. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. https://doi.org/10.1145/2998181.2998331

Passi, S., & Jackson, S. J. (2018). Trust in Data Science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–28. https://doi.org/10.1145/3274405

Pine, K., Bossen, C., Holten Møller, N., Miceli, M., Lu, A. J., Chen, Y., Horgan, L., Su, Z., Neff, G., & Mazmanian, M. (2022). Investigating Data Work Across Domains. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–6. https://doi.org/10.1145/3491101.3503724

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

Ribes, D., & Finholt, T. (2009). The Long Now of Technology Infrastructure: Articulating Tensions in Development. *Journal of the Association for Information Systems*, *10*(5), 375–398. https://doi.org/10.17705/1jais.00199

Ribes, D., & Polk, J. (2014). Flexibility Relative to What? Change to Research Infrastructure. *Journal of the Association for Information Systems*, *15*(5), 287–305. https://doi.org/10.17705/1jais.00360

Schatzki, T. (2001). *Introduction: practice theory. The practice turn in contemporary theory.*

Seidel, S., Recker, J., & vom Brocke, J. (2013). Sensemaking and Sustainable Practicing: Functional Affordances of Information Systems in Green Transformations. *MIS Quarterly*, *37*(4), 1275–1299. https://doi.org/10.25300/MISQ/2013/37.4.13

Seidel, V. P., Hannigan, T. R., & Phillips, N. (2020). Rumor Communities, Social Media, and Forthcoming Innovations: The Shaping of Technological Frames in Product Market Evolution. *Academy of Management Review*, *45*(2), 304–324. https://doi.org/10.5465/amr.2015.0425

Steinhardt, S. B., & Jackson, S. J. (2015). Anticipation Work. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 443–453. https://doi.org/10.1145/2675133.2675298

Vassilakopoulou, P., Skorve, E., & Aanestad, M. (2019). Enabling openness of valuable information resources: Curbing data subtractability and exclusion. *Information Systems Journal*, *29*(4), 768–786. https://doi.org/10.1111/isj.12191

Winter, J. S., & Davidson, E. (2019). Big data governance of personal health information and challenges to contextual integrity. *The Information Society*, *35*(1), 36–51. https://doi.org/10.1080/01972243.2018.1542648

Zuboff, S. (2015). Big other: Surveillance Capitalism and the Prospects of an Information Civilization. *Journal of Information Technology*, *30*(1), 75–89. https://doi.org/10.1057/jit.2015.5

Appendix here: 10.5281/zenodo.8070432

Includes more on Methods: data sources, analysis, and emerging constructs.