

## Data curation as anticipatory generification in data infrastructure

Elena Parmiggiani, Nana Kwame Amagyei & Steinar Kornelius Selebø Kollerud

To cite this article: Elena Parmiggiani, Nana Kwame Amagyei & Steinar Kornelius Selebø Kollerud (2023): Data curation as anticipatory generification in data infrastructure, European Journal of Information Systems, DOI: [10.1080/0960085X.2023.2232333](https://doi.org/10.1080/0960085X.2023.2232333)

To link to this article: <https://doi.org/10.1080/0960085X.2023.2232333>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 05 Jul 2023.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)

## Data curation as anticipatory generification in data infrastructure

Elena Parmiggiani, Nana Kwame Amagyei and Steinar Kornelius Selebø Kollerud

Norwegian University of Science and Technology, Department of Computer Science, Trondheim, Norway

### ABSTRACT

Data curation is crucial for data reusability. New possibilities for digital data sharing are an urgent concern for data curators, who must keep historical datasets and present data collections always ready to meet unknown future data needs. This calls for a more nuanced understanding of the temporal horizons of data curation in Information Systems research. Based on a qualitative interpretive case study of data management in an environmental monitoring infrastructure, we characterise three data curation practices to support data reuse. These practices follow three interleaving temporal perspectives: retrospective (by upgrading historical datasets), present-oriented (by monitoring ongoing data collections), and future-looking (by disseminating data). We conceptualise this work as *anticipatory generification*, involving continuous and temporally oriented data curation to maintain data sufficiently open-ended to anticipate future data reusability. Anticipatory generification is essential for the sustainable evolution of environmental data infrastructures. Our study contributes to the Information Systems literature by further theorising the temporal perspectives of data infrastructures and providing additional insight into how the future is anticipated in practice.

### ARTICLE HISTORY

Received 9 April 2021  
Accepted 26 June 2023

### KEYWORDS

data infrastructure; data curation; anticipatory generification; data sharing; data reusability; sustainability


## 1. Introduction

*All participants are signed on to an online workshop for scientists and data curators to discuss challenges, opportunities, and trends for data management and education in environmental monitoring. Lisa, an ecologist, gives the first presentation. Her first slide is titled “The Importance of Long-Term Datasets”. (...) Immediately following the presentation, Annie goes deeper into the discussion, “What can we tell about the functioning of an ecosystem based on 100 years of data?” She is excited about her wording. She adds that it stressed the importance of building our understanding of how an ecosystem works over time based on data and people’s role in ensuring that data are still good enough 100 years later. (April 2021, online workshop, excerpt from fieldnotes)*

Data infrastructures depend on data curation. Only apparently a mundane task, data curation is a combination of routine and ad-hoc practices that are performed by different professionals “behind the scenes” of data infrastructures to find, prepare, and maintain data for further use and reuse (Jones, 2019; Parmiggiani et al., 2022; Vassilakopoulou et al., 2019). Data sharing aimed at future (re)use is an increasingly urgent concern for data curators because new possibilities are emerging for making digital data available to other users, systems, or organisations. These data can travel far (Leonelli, 2019), as they are further packaged and reused to develop new services across organisations or domains (Alaimo et al., 2020; de Reuver et al., 2018). Currently, such processes are at the heart of business models based, for example, on data sourcing (Jarvenpaa & Markus, 2020),

crowdsourcing (Lukyanenko et al., 2019), predictive analytics (Waardenburg et al., 2021), and social media (Alaimo et al., 2020). Data sharing within and across data infrastructures is also surging in the sciences to address complex research questions (Ribes & Polk, 2014; Tempini & Leonelli, 2018), meet open data policies (Link et al., 2017), and allow interoperability (Knowledge, n.d.).

Research on data work in Information Systems (IS) has demonstrated that digital data are not generated, shared, and then travel in a vacuum but are entangled with the means and conditions of their production and sharing and require continuous curation efforts to shape and prepare them (Jones, 2019; Monteiro & Parmiggiani, 2019). These studies have surfaced the long-term nature of data lifecycles, notably in the case of oil and gas exploration (Mikalsen & Monteiro, 2021), healthcare delivery (Aanestad et al., 2017), medical research (Ribes & Polk, 2014; Vassilakopoulou et al., 2019), and in firm- or business-related contexts such as telecommunication (Aaltonen et al., 2021). The consequence of this evidence is that the work aimed at ensuring the future (re)usability of shared data is an inherently longitudinal endeavour: since data are supposed to endure over time (Jarvenpaa & Essén, 2023), data curators are constantly caught with one foot in the needs of the present and with the other in a desired future (Ribes & Finholt, 2009). As a result, different temporal horizons characterise the work to let data endure from the past and present and be shared for

**CONTACT** Elena Parmiggiani  [parmiggi@ntnu.no](mailto:parmiggi@ntnu.no)

future reuse: “Historical and present data also must remain available and accessible in near and distant futures, for going back in time and seeing new data linkages and combinations” (Jarvenpaa & Essén, 2023, p. 1). However, we lack empirical insight and theorising about how curators learn to cope with these different temporal perspectives to ensure reusability (ibid.). To address this, we ask: *How does data curation strive to ensure data reusability in data infrastructures?*

This paper addresses data infrastructures in the environmental sciences. This case highlights that data sharing is not a one-time achievement but is maintained as part of continuous data curation in which data must be prepared for going “into the wild”; that is, they must be made ready for both expected and unexpected use and reuse by different people and institutions over time (cf. Lukyanenko et al., 2019). Thus, a crucial challenge for data curators in environmental research is to make digital data sufficiently *open-ended* to meet knowledge needs across human and technological generations without fully knowing what they will be needed for, when, and in what context.

We draw on a qualitative interpretive case study of data management at environmental monitoring research stations in Norway connected to the European Long-Term Ecological Research (eLTER) network. This is a compelling case of the continuous work to collect data locally and share them openly across research stations and countries over time. Earlier research on anticipatory work unearthed scientists’ data curation efforts to envision and frame assumptions of future data needs (Bechky, 2021; Flyverbom & Garsten, 2021; Steinhardt & Jackson, 2015). Whereas anticipatory work typically has a guiding character aimed at reducing complexity by constraining possibilities for reuse, we propose the concept of *anticipatory generification* to stress the importance of increasing, as opposed to constraining, possibilities of data reusability because of the uncertain conditions of future data use in environmental data infrastructures. This notion draws on the concept of generification originally presented by Pollock et al. (2007), who studied practices of keeping generic software packages malleable enough so that they can be exported outside the settings in which they were initially produced. We define anticipatory generification as the *ongoing, heterogeneous, and temporally oriented data curation to make data sufficiently open-ended to enable future data (re)usability over time*. We flesh out three sets of recurrent and interconnected practices of anticipatory generification along three temporal perspectives: (i) Looking backwards: Upgrading data; (ii) Looking to the present: Monitoring data; (iii) Looking forward: Disseminating data. These three sets of practices contribute to keeping data updated, accurate, meaningful, and accessible so that data are sufficiently

open-ended and, thus, generic<sup>1</sup> to answer uncertain future scientific research questions.

Our characterisation of anticipatory generification extends theories of data curation in data infrastructures by characterising and theorising how data reusability is addressed in practice along three temporal horizons to ensure the continuity between the past, the present, and the future. Our findings have theoretical implications for discussions of anticipatory work and the sustainability of data infrastructures.

## 2. Theoretical background

The recent surge of data infrastructures has been fuelled by the availability of interconnected off-the-shelf sensing devices, and thus digital data, to support a range of new services, including intelligent public services (Velsberg et al., 2020), programmatic advertising (Alaimo et al., 2020), long-term environmental monitoring in industrial activities (Monteiro & Parmiggiani, 2019), and scientific research (Ribes & Finholt, 2009). We define data infrastructures as sociotechnical arrangements encompassing a variety of users, stakeholders, agendas, workflows, and technologies such as sensors, repositories for data storage and retrieval, data analytics systems, and interfaces for data access and sharing (Monteiro et al., 2013).

Data infrastructures rest on data curation practices to find and prepare data (Jones, 2019). Data curation has been recently broadly defined as “a data management activity, [which] involves the development of physical and logical infrastructures that make it feasible to collect, index, and store data, and facilitate data access for subsequent analysis” (Chua et al., 2022). It is a unique form of work that consists of ongoing, convoluted, and longitudinal practices of balancing heterogeneous concerns while finding and preparing data (Parmiggiani et al., 2022). Research on data infrastructures is adamant that data curation practices are long-term and future-oriented, in the sense that their ultimate goal is to ensure the longevity of data infrastructure over time (Ribes & Finholt, 2009), adapting it to new phenomena (Monteiro & Parmiggiani, 2019), new user groups (Ribes & Polk, 2014) or new analytical methods (Waardenburg et al., 2021). This longitudinal perspective implies that the use value of data is not pre-given but contingent on continuous knowledge work (Vassilakopoulou et al., 2019). Data (re)users gradually learn to ask practical questions about the context in which data were crafted, including information about who has previously processed the data, sensor configuration and calibration, and methods used, which in turn impacts the design of the data format and collection (Mikalsen & Monteiro, 2021; Ribes & Polk, 2014).

Digitalisation has enabled innovative forms of data sharing, resulting in new possibilities for data (re)use and innovation. As a result of this process, data stretch and evolve over time. Recent studies have observed this phenomenon in various contexts. For example, telecommunication data are assembled, packaged, and traded into commodities, later repurposed and aggregated to create innovative products for programmatic advertising (Aaltonen et al., 2021). In social media, diverse data types are recombined into updatable data bundles widely circulated across firms and business sectors to develop new services (Alaimo et al., 2020). In the offshore oil and gas industry, data about physically inaccessible subsurface geology are collected through faulty sensor networks, massaged, accumulated, traded, and reused across companies to produce new interpretations of hydrocarbon reservoirs over decades (Mikalsen & Monteiro, 2021). Data sharing towards unprompted data reuse also surges outside business-related contexts such as citizen science (Lukyanenko et al., 2019). In scientific research, data accumulate and are circulated and reinterpreted across projects (Tempini & Leonelli, 2018). Simultaneously, uncertainty about future (re)use plays a prominent role due to the need to use data to analyse slowly evolving phenomena, the understanding of which progressively changes as new data are collected and interpreted (Ribes & Polk, 2014).

Evidence from these studies shows that data curators must cope with the difficulty of imagining future contexts of data (re)use because it depends on evolving business or scientific insights on past and present data. Jarvenpaa and Essén (2023; see Garud & Gehman, 2012) describe this as an issue of data sustainability, intended as the work to ensure the durability of data infrastructures: “as continuity and links between the past, present, and future, noting the possibility and importance of going ‘back to the future’.” (p. 5). They call for more research to unpack the temporal orientations of data work to understand how future-oriented attention to data depends on choices about data from the past and the present.

### 3. Towards data curation as anticipatory generification

Data sharing is the effect of interconnected material and social practices driven by practical concerns that guide daily work (cf. Nicolini, 2012). We are inspired by a technology-in-practice lens (Orlikowski, 2000) to unpack how users learn to engage with the longitudinal nature of digital data and maintain and share them as resources that could be reused over long periods of time. The concept of *anticipatory work* has been proposed in IS and the neighbouring fields to shed light on the situated data management practices to make data ready for assumed future reuse in different

contexts (Barley, 2015; Bechky, 2021; Flyverbom & Garsten, 2021; Johansen et al., 2016; Steinhardt & Jackson, 2015). Anticipatory work is “the practices that cultivate and channel expectations of the future, design pathways into those imaginations and maintain those visions in the face of a dynamic world” (Steinhardt & Jackson, 2015, p. 433). Visions about the future become embedded in data infrastructures by adapting complex local and institutional requirements in routine infrastructural work (Ribes & Finholt, 2009) more than in thought (Johansen et al., 2016). For example, Barley (2015) demonstrates how weather scientists organise their work to produce data representations that can be shared as commons with other knowledge communities by shaping their practices in anticipation of possible future data representation needs of other teams. Bechky (2021) provides a very vivid picture of the daily laboratory practices of forensic scientists who “anticipate the concerns of attorneys, who, in turn, anticipate jurors and the public at large” (p. 75) in preparing the data by documenting, reporting and setting up instrumentation. Often, this daily work is substantially altered and augmented by the anticipatory practices developed by workers to cope with the additional burden caused by curating data (Waardenburg et al., 2021). This work is highly consequential for the way data are constructed (ibid.) as resources in anticipation of future needs, for example, in crime investigation (Bechky, 2021) and climate science (Leonelli, 2019).

The extant literature has to date primarily focused on examples in which anticipatory work is used to define, frame, or limit possibilities for future reuse. Environmental monitoring data infrastructures are an extreme case in which data must endure over technological and human generations to be (re)analysed by future (re)users to answer evolving research questions about long-term phenomena. This calls for conceptualising anticipatory work that accounts for future data reusability’s highly open-ended and uncertain conditions. Pollock and colleagues’ concept of *generification* (Pollock et al., 2007; see also Pollock & Williams, 2009; Monteiro et al., 2013) is helpful to open the work of anticipation to encompass uncertain conditions of further data (re)use. In their work, Pollock and colleagues unearth the longitudinal practices to make enterprise resource planning software so generic that it can be reused and adapted across several business contexts. Generification work solves infrastructural challenges by relying on local practices and carefully adapting infrastructures over time (Hanseth & Bygstad, 2015; Silsand & Ellingsen, 2014). Our empirically driven theoretical interest is in a reality in which the generification work of keeping ready for the unexpected involves digital data that are always in the making: they must be kept open-ended, that is, adaptable to local needs while also allowing for

unprompted interoperability across different settings over time such as research stations, universities, cross-border databases and data platforms, and policy-makers. We dub this form of work *anticipatory generification*. Building on research on anticipatory work and generification, our empirical case represents a relevant testbed for shedding further light on the temporal orientation of data curation to ensure reusability.

#### 4. Empirical context and case

The work of environmental monitoring varies greatly depending on how a monitoring research station is organised and funded, what technologies are available and what natural objects are being monitored. A recurring form of organisation is research stations close to the measured environmental phenomenon, such as fresh or oceanic waters, forests, bird communities, air quality, or Arctic environments. Research stations usually combine manual approaches with several digital technologies to measure different environmental parameters and to clean, analyse and store data.

In Europe, the European Union (EU) promotes sharing and reuse of open data (Link et al., 2017), motivated by its “policy goal of integrating national research systems in institutional and epistemic terms” (Kaltenbrunner, 2017, p. 276) through instruments that translate political objectives into concrete advice for research infrastructures in Europe. Despite this focus by policy and funding agencies, data-sharing requirements are underspecified. This is due to the long-term nature of environmental monitoring work and is reflected in the fact that policy relies largely on standardised representations of data management that tend to have little regard for actual data management practices. On the contrary, environmental policy researchers assert that environmental behaviour needs to be examined in the context of longitudinal, day-to-day monitoring routines (Blanchard et al., 2014).

We draw on a study of the Norwegian node of the eLTER network. eLTER aims to develop a holistic understanding of the integrated impacts of climate and environmental change on a wide range of European ecosystems. The Norwegian node includes several stations spread across the country. Norway prioritises monitoring air quality, fauna, and flora in freshwater, saltwater, forests, and Arctic tundra. As a member of the European Economic Area, Norway actively participates in receiving and implementing EU research funds. The Norwegian Research Council (NRC) provides initial funding to research stations to help them modernise their facilities and meet the requirements of EU programmes.

Different professionals work at environmental research stations, all contributing to data curation (Karasti et al., 2018); for example, environmental

researchers, biologists, zoologists, and physicists are usually employed by research stations or financed by grants. Interns and bachelor’s and master’s students also often spend a few months performing environmental monitoring at a research station as part of their training. Environmental research stations also often employ (either permanently or temporarily as part of NRC- or EU-funded projects) software developers and systems engineers who programme sensor networks and develop data storage and analysis software. Finally, the maintenance of a research station and its equipment depends on several engineers, craftsmen, and volunteers’ work. The time these people spend at a research station depends on their tasks. Some research stations are closed during the coldest months.

#### 5. Research methods

This paper is based on an interpretive qualitative case study (Walsham, 2006) of data management activities in environmental data infrastructures in Norway. The case was theoretically sampled (Eisenhardt & Graebner, 2007) from our long-established interest in data work in heterogeneous domains (Monteiro & Parmiggiani, 2019; Parmiggiani & Grisot, 2020). This case is particularly suitable to illuminate the relationship between data curation and sharing in data management. Environmental data management is revelatory (Yin, 2009) of the challenges associated with data sharing because of the nature of scientific work and the requirements for publicly funded research stations to make data openly available. This case also provided a pragmatic opportunity for unusual access (ibid), as we conducted our study against the background of a long-term engagement with environmental research in Scandinavia and Europe. The first author has a decade of experience studying data work in research-based and industrial settings in different Nordic countries (Karasti et al., 2018; Parmiggiani et al., 2022). Access was initially facilitated by the first author’s participation in a project funded by the NRC (2017–2018 “InfraData: infrastructuring Internet of Things for public governance”) to study data work in infrastructures. The data collection and analysis were performed in line with the guidelines and approved by the Norwegian Centre for Research Data (currently Norwegian Agency for Shared Services in Education and Research).<sup>2</sup>

#### 6. Data collection

The findings of this paper are based on qualitative data: semi-structured interviews, observations, and documentation. A detailed overview of our data sources can be found in Table 1. Data were collected from 2017 through 2023 by each author in three

**Table 1.** Overview of the empirical data sources.

Data source	Number of informants per role/site
Observations (Approx. 70 hours, 1 day equals approximately 10 hours)	4 days at 4 environmental research stations 1 day at a conference for environmental researchers 1 day at an online workshop on environmental data management in eLTER 2 days at online workshops for eLTER data curators
30 semi-structured interviews (45 to 90 minutes each)	2 environmental station managers 13 environmental researchers 2 environmental station engineers 6 project managers at research environmental institution 3 system engineers 1 software developer 3 database administrator and data curator
5 structured interviews (written)	1 database administrator 1 environmental station manager 3 environmental researchers
Documentation	Strategy reports by EU White papers, calls for applications, guidelines for establishing research infrastructures and strategic reports by the NRC Official descriptions, data sharing policy, standards, and guidelines for data collection by the eLTER network Other repositories of regulations, white papers, and policy papers by Norway and the European Commission

intervals: author 1 in 2017–2019, author 2 in 2020–2023, and author 3 in autumn 2020–spring 2021.

Our primary data source consisted of qualitative interviews (Myers & Newman, 2007), separately conducted by all the authors (see Table 1 for an overview of the roles of each interviewee). We conducted 30 semi-structured in-person interviews in 2018–2019 and spring 2023 and digitally via Zoom, MS Skype or Teams during the COVID-19 outbreak in spring 2020–autumn 2021. The interviews were necessary to explore and elaborate on the informants' practical concerns and opinions and their experiences working with data in environmental monitoring, including concrete examples. To supplement the data gathered via semi-structured interviews, we conducted five structured written interviews online to investigate participants' perspectives further. Although interviews are not the ideal primary source to investigate work and practice (Nicolini, 2009), they allow us to reveal invisible work that might be invisible during ordinary activities (Myers & Newman, 2007; Rubin & Rubin, 2011), thus identifying and discussing patterns in users' actions over time, in addition to actions at the moment (Schultze & Avital, 2011).

The second data source was observations, totaling approximately 70 hours. First, we obtained a more detailed understanding of the day-to-day data work activities and decisions at research stations. Online events and seminars organised by the eLTER network were interesting in investigating emerging concerns across research stations and scientific domains. Second, observations

helped us to witness how participants talked about and made sense of their work as they interacted with their colleagues and peers during joint events, such as conferences and seminars. Third, observations assisted us in developing the focus and questions to ask for the subsequent rounds of data collection, eliciting reflections on data work practices.

Finally, we collected policy documents and other documentation issued by Norwegian and European regulators on data infrastructure funding. This documentation was necessary for shedding light on existing regulations on data-sharing requirements and funding for environmental monitoring.

## 7. Data analysis

Data analysis followed an interpretive research paradigm based on a hermeneutic approach to make sense of environmental monitoring work practices as a complex whole by alternating between our perspectives and those of the informants (Klein & Myers, 1999). The analysis was performed by all three authors, first separately and then together. It happened in three phases and partly overlapped with data collection through a deductive-inductive strategy in three phases, iterating between theory and data (Eisenhardt, 1989; Tjora, 2018).

In the first phase, each author coded the collected data using an open coding strategy. We were initially interested in uncovering the work to manage data to track environmental phenomena over time, in line with our established engagement with research problems related to data management in different

contexts (Monteiro & Parmiggiani, 2019; Parmiggiani & Grisot, 2020). In line with interpretivism's stakeholder-centric perspective, we labelled the practical concerns (Nicolini, 2012) voiced by our informants. We followed Emerson et al. (2011) guidelines for coding and making sense of ethnographic data. The first author initially used coloured pens, highlighters, and sticky notes. The second and third authors used computer-based qualitative data analysis software.

In the second phase, we identified several concerns related to data sharing, particularly the lack of knowledge about the context of future reuse of those data. This triggered our curiosity. We particularly examined how researchers solved these concerns in their daily routines. To address this, we defined the final version of our research question. To answer it, we merged all our codes into one spreadsheet and clustered them into overarching groups of practices. We then compared the emerging clusters in incremental feedback loops in which the codes were grouped into conceptual categories. This round was largely inductive but with deductive imports inspired by theories on the temporality of data infrastructure (Ribes & Finholt, 2009; see also Jarvenpaa & Essén, 2023). We realised that environmental researchers spoke unprompted about the temporal orientation of their work when describing practices to upgrade historical datasets, continuously monitoring current data collections, and keeping data ready to be disseminated (see quotes in Table 1, "Excerpt" column). This led us to refine our conceptual categories by surfacing three main patterns of interleaving ongoing curation practices to prepare past and present data for future (re)use.

In the third phase of our data analysis, the first and second authors worked together to further group all the categories into broader constructs and approach theory. Our experience researching data work in different contexts influenced our analysis (Suddaby, 2006). Through collective deliberation, we engaged with theoretical imports to conceptualise this form of work and initially agreed to conceptualise it as anticipatory work. This part of the work was led by the first author with more extended experience in IS. We validated our analysis through discussions with colleagues and collaborators. In doing so, we became aware that anticipatory work in the literature is usually associated with narrowing down or closing the possibility space for guiding future activities. In dialogue with our colleagues, we realised that a striking element of our case, however, was the ongoing concern to make data as open as possible so that they could be ready to anticipate scientific questions that might emerge in the future. It was still anticipation but of a different, open-ended quality. At this stage, we were inspired by the work of Pollock and colleagues (2007) on generification. We finally developed the concept of anticipatory generification to highlight data curation as a

practice to maintain environmental data as open as possible for further sharing by upgrading past accumulated data, addressing present needs, and allowing future possibilities for reuse. Table A1 in Appendix A presents the interpretive template resulting from our data analysis.

Figure 1 summarises our conceptualisation of anticipatory generification practices for illustrative purposes. It shows the temporal perspectives of how workers at a typical eLTER research station curate the data to be ready for sharing by being up-to-date, accurate, and meaningful to future (re)users. As we shall present in the next section, they do this by looking *backwards* to retrieve and update historical datasets collected locally at the research station, focusing on the *present* to monitor ongoing data collections; and taking a *forward* perspective by disseminating data. Whereas these practices sometimes overlap as they are performed back-and-forth and are interdependent, we distinguish analytically between them to highlight the essential elements that support long-term data sharing and future reuse. The following section presents these three anticipatory generification practices summarising multiple concerns, technologies, interactions, and coordination efforts that characterise data curation.

## 8. Findings

### 8.1. Upgrading data: looking backwards

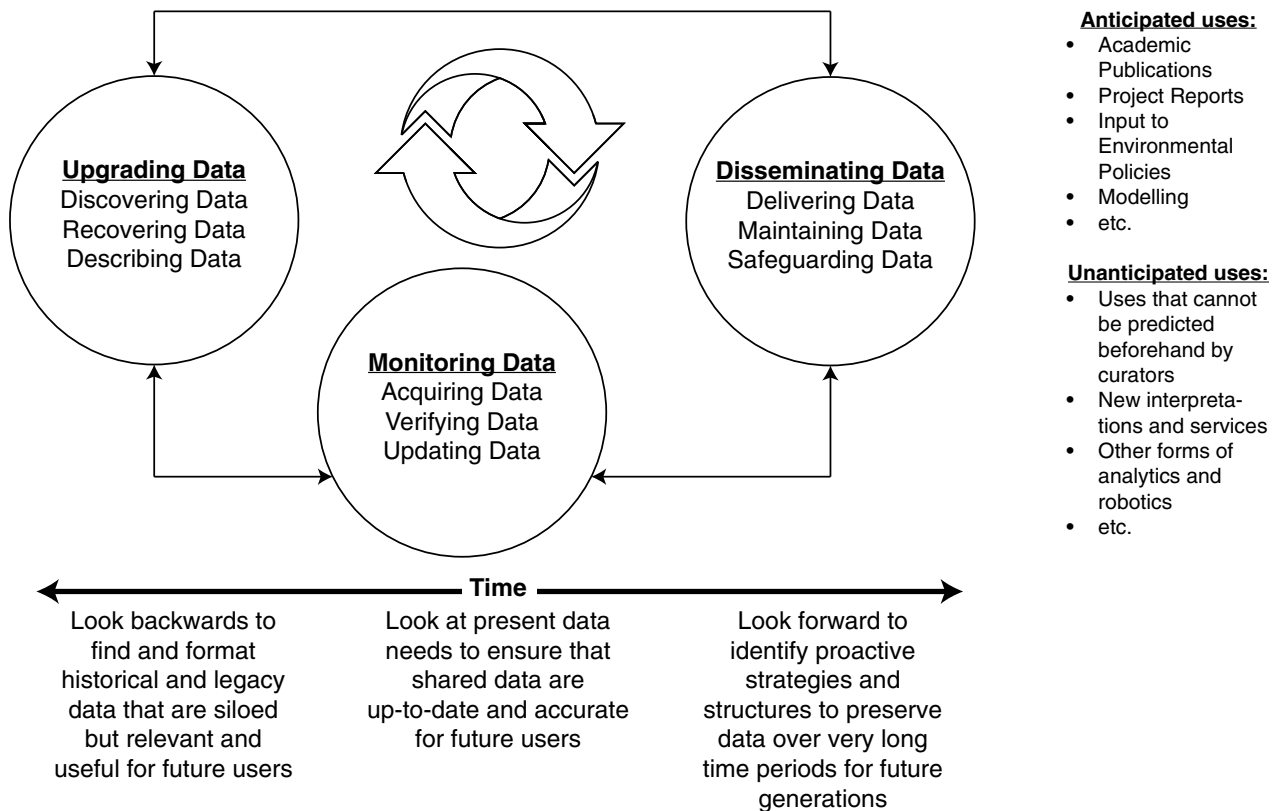
#### 8.1.1. Discovering data

It is common for scientists to have useful data stored in filing cabinets and paper sources. Requirements by funders and regulators to share environmental data is forcing scientists to adopt approaches to make such data sources available to the public:

[U]ntil now we store it in our paper archive, it follows a project. The Council is starting to set more demands on the availability of data, and we are constantly working on new solutions as to how we do it. (Tone,<sup>3</sup> scientist, interview)

In general, eLTER environmental scientists collect different types and forms of data:

We are analysing nutrient chemistry in the water. And then we're also looking at phytoplankton, that's microalgae. And then we're looking at zooplankton and then we're also looking at fish larvae and eggs of fish because they are the primary consumers of this micro and mesozooplankton. So, we have four different categories, and they all have different details here about how to do this ... I've been working for five years now to get the different data that are basically sitting on different Excel sheets to become part of a bigger Data Archive, and to make it accessible to others and the public. And that's been a long road, because we've had nutrient chemistry data, for instance, we have more than 30 years of data. (Eva, scientist, interview)



**Figure 1.** Anticipatory generification practices to ensure data reusability in data infrastructures along the three identified temporal perspectives.

Each scientist decides on methods, analysis tools, and how data are processed and stored. Ecological data are stored in different sources, including data on Excel spreadsheets, paper files (typically scattered in scientists' offices and local research station archives), on scientists' handheld devices and computer hard drives:

“The current situation at [our local research station] is very fragmented, where individual researchers and research projects decide on how they want to store data, where they want to store the data, and how it is shared. [...] An essential part of our work is dealing with these challenges to try and improve and standardise the situation”. (Pieter, program coordinator, interview)

### 8.1.2. Recovering data

In one of the research stations, data curator Ragnar and his team are retrieving old data and recovering digital data from legacy systems, i.e., old or outdated software and applications including databases. They aim to find as much relevant data as possible and organise them meaningfully in an open data portal. Ragnar and his team of curators intend to find and copy data from sources such as older versions of Excel, local data analysis tools, and laboratory information management systems:

If we have to share the data openly, then we cannot choose which data to share. We must share as much as we can. We had data in our old laboratory information management system that we had to copy onto

our newer system ... and we are talking about big terabytes of data. (Johannes, data curator, fieldnotes)

The curators in Ragnar's team have developed several computer scripts for recovering and formatting legacy data. However, these approaches still require manual workarounds:

The data stewards have made a script so that we can extract the data from our older database to a format that can work with the new one [...] but the main task, [...] is [ensuring that the retrieved data] has all the right tags. So, it's not like a [completely] automated procedure. (Tomas, Environmental Researcher, field notes)

This work is necessary to ensure that digital data that are inaccessible to the public and have different formats than existing data sharing systems can be recovered and adapted for use in a different future setting.

### 8.1.3. Describing data

Curators also strive to enrich data records with contextual information. A common approach is to consult primary data producers, usually through informal interactions, to ensure that the contextual data are meaningful to secondary users:

Sometimes we interview scientists, especially those who are retiring to get as much (meta)data as possible. We can also go to [scientists'] offices to copy files directly. We sometimes get data from older paper files in our archives (Jude, data curator, interview)



Research stations in eLTER typically use metadata (i. e., data with properties about data) standards approved by the eLTER network to describe data. This is supposed to ensure that data and metadata are understood when used in a different future setting. In general, curators describe data with several contextual information:

Metadata need to include information on the observation location and the research context. [A standard metadata model] defines metadata elements about the organisation (e.g., contact information and networks), the location, the observation characteristics (e.g., climate, habitats) or available equipment. Additionally, there are fields about the focus and design of a site, network affiliation and information about data policies and data management. (Document, DEIMS)

Fulfilling such metadata requirements is done with the assumption that future data users will want to know more about how and who collected the data and under what conditions, in order to assess whether data obtained on an open portal on the web is reliable for their purposes:

We describe data with the relevant information to help another user in the future to understand the settings and tools used for collecting the data. This way, they trust the data (Anna, senior researcher, interview)

Anna, a senior researcher, described in detail the importance and complexities involved in metadata work:

So, if a complex [eLTER] metadata standard has 50 fields, we might only fill in 20, but they are comparable [to other standards used within the network], so they can be translated. So that's how you make sure that the data can be exported [or shared within the network] in a standardised way without necessarily having to meet all the metadata requirements [...] But it's something that should certainly always be done, [...] it's something that tells you that it takes a while to move an entire discipline to Open Data. So, it's a process that's underway but it's by no means perfect. (Anna, senior researcher)

When describing data with metadata, it is usually assumed that those who know the most about the data, their nature, and the actual data input should also provide information about the context. However, this is not always the norm because most scientists prefer to use their time to write journal publications rather than curate the data meaningfully for open sharing.

In sum, curators and scientists upgrade existing data so that future users can find them for disparate needs, such as academic research, project report or as input for environmental policymaking, and other unexpected uses that curators or scientists cannot

predict. Practices for upgrading data are backwards-looking because they consist of *digitisation work* to discover typically old but relevant data in paper-based sources, *repurposing work* to recover old datasets and resolve their different historical formats into contemporary ones resulting from changing technologies and versions, and *metadata work* often done after data have been collected to describe them in a meaningful way for future users. This requires deciding today what “data about data” will be needed by future users. This work is a practice with a forward-looking dimension, but it requires going back to data that has already been collected to describe them. The practices for upgrading data are varied and may require curators to interview researchers who know more about the data collection processes, methods, and environments in which the primary data were collected, to determine what paper-based sources or legacy data are relevant, and to describe them adequately for future users. Curators may also develop scripts to retrieve legacy data and put them into contemporary forms. These practices, which appear just clerical, are anticipatory because they strive to ensure that historical datasets will be relevant and useful for future research questions.

## 8.2. Monitoring data: looking to the present

### 8.2.1. Acquiring data

The predictable and immediate work of eLTER scientists includes collecting regular seasonal and annual data cycles. From a long-term perspective, scientists are further required to monitor the progress of recurring data collections and check that they remain accurate over an extended period. Environmental scientist Freya and her team monitor lynx species in Norway. They have monitored data on lynx species to ensure that they are accurate. Freya's team solves ongoing data collection problems, such as placing cameras in the field, obtaining consent for camera placement, and determining the number of each lynx species. As project manager, Freya is responsible for coordinating all activities in the project. She has a data management plan describing what data will be collected, how data will be cleaned, and in what shared repositories data will be available to team members and the public. However, this plan only provides a scaffold for Freya and her team, who constantly work to make data collection possible. The project aims to determine the exact number of lynxes in an area and report that number to the authorities issuing hunting permits to hunters in the area. Digital cameras fitted with infrared sensors are to be installed within an area in the forest to detect and capture warm moving objects such as lynxes in this case. This method is termed camera trapping. Freya interacted with Ragnar, a data curator,

on the possibility of installing camera traps within the new area. Ragnar was concerned because this area is owned by locals who do not trust the government. However, the expertise of these locals was required not only because they owned the lands in the area and could grant access to Freya and her team, but most importantly because these locals knew more about the movement of the lynx species and could advise on best locations to instal the cameras to collect data adequately. Freya was faced with the challenge of obtaining approval from the owners of the lands, as well; she needed local guides, usually hunters, to advise on where best to position cameras:

“These local people know more about where these animals go. And they would also be able to get the consensus from those who own that land. There are many people that are very restricted about monitoring, especially the large carnivores (...) [I]t’s almost impossible for us to go to ask the people out on the ground [if we] can put up a camera. But if a local person asks, they will say yes more often than if we asked, because they trust [other] locals more, and they don’t trust the government. Even though we are not working for the government, they look at us in the same way”. (Freya, Project Manager, field notes)

With the concern of recruiting local guides as allies for data collection, Freya contacts the National Hunting and Fishing Agency about a possible collaboration to recruit local guides on her behalf:

“We have these people who are working at the Hunting and Fishing Agency. They ask on [our] behalf. So, they are the ones who do the hiring of the locals. And we are hiring the organisation and the organisation ... gets the local people. (...) We have approximately a little bit more than 100 persons [local volunteers] around in Norway who help us with collecting data from cameras”. (Freya, Project Manager, field notes)

### 8.2.2. Verifying data

Involving local guides helped Freya and her team to address another challenge with data collection – checking images on the camera for duplicates and reporting back to Freya and her team. Local guides are also precious in resolving trust issues with the government. According to Freya, local guides will trust government policies on hunting activities within the area if they see the data that led to such policies themselves. This may reduce conflicts between her team and the locals in further cycles of data collection:

“They [the local guides] will be able to look at the data [on the camera] before they send the memory card to us. That also helps with the trust issues, the conflict, they will trust the data more and they trust the government more when they’re seeing these data by themselves. So, we can see that it reduces the conflicts a lot when we are using these local guides sometimes it’s a little bit more work for us, but it makes it

possible to monitor these large carnivores”. (Freya, Project Manager field notes)

After the local guides check the images, the memory cards in the digital cameras are sent to Freya and her team. The cards are connected to a computer, and data are transferred. However, due to the volume and variety of data, local guides cannot check all the data for duplicates. As a result, once data reach their servers, Freya and her team use algorithmic techniques to clean the data:

“We have a machine learning program that use the information about the time. It will put all these pictures from the same time series into one ID and will come with a suggestion of the species that is on this picture. All human pictures are censored and removed. And they are stored in a database”. (Ragnar, Data curator interview)

The machine learning algorithm helps curate the large volume of data. However, the variety of data exposes its limitations. Freya does not have a big enough team to validate the outcome of the machine-learning algorithm. She thus employs students to verify the data:

“It can be difficult to see in the dark. All pictures there aren’t erh... a lot of pictures have quite bad quality because most are out at night. And it’s forest and it’s windy and it’s raining and it’s a lot of bushes. It can be hard [for the machine learning program] to be sure. So often we are using a lot of students to go over all these pictures to verify each time series”. (Freya, Project Manager field notes)

### 8.2.3. Updating data

After the work done by students, data are stored in another database where the experts, i.e., scientists and curators, return to validate the data cleaning work of students:

Yeah, because of how it [points to a blurred picture of a stone] is shaped and so, and if those who have verified the pictures [did] a little bit too fast. It can be stored as a lynx, so we [scientists and curators] have, from time to time, been going through to just look. Recently one student added some of those rarer species. We can’t do that. (Freya, Project Manager interview)

Nils gave us an even more vivid description of the long-term nature of sorting out names of seasonal data cycles and its role in ensuring that data remain up to date for future uses:

We do data wrangling also in terms of biological data which often involves matching scientific names, Latin names of species with Norwegian and English names. If I’m going to make a report for Popular Science report and a report for the Norwegian government, I need to translate the species from Latin names to Norwegian names ... It’s often a matter of not just matching the species, but ensuring that the species that were first reported are the correct scientific name

that's currently in use for that species because they change over time as species are reclassified to make sure that you're actually using the most up to date, species name. So those things are typical data-wrangling issues. Then you have of course rooting out errors ... there are several quality control checks before the data are shared. But still, there are some errors that pop up every now and then. (Nils, Data curator interview)

This point further highlights that curating scientific data for sharing publicly is rarely a wholly automated process but requires a significant amount of human agency if accurate data are to be shared.

In sum, experts resolve several data issues by monitoring seasonal and annual data collections and their preservation using diverse sociotechnical practices. These practices range from the *data generation work* to include local participants in data collection, the *data validation work* to supplement automated data cleaning processes with additional expertise, and the *data resolution work* to resolve and sort out labels for data continually. This work is done, for example, by obtaining permission from landowners to position cameras, recruiting local guides for data collection, and checking cameras for duplicate and repeating images. Looking at the present, therefore, encompasses ongoing practices to monitor data to keep them ready for reuse by continuously controlling data collection and ensuring that future users can find accurate and up-to-date shared data. Combined with maintaining historical data records, this work is essential to contribute reliable data in long-term open data sharing.

### 8.3. Disseminating data: looking forward

#### 8.3.1. Delivering data

eLTER scientists and curators disseminate data via several ad-hoc approaches, such as physical drives and other peer-to-peer solutions. For immediate uses, these approaches help them collaborate and produce research results often defined by short-term time frames, such as monitoring the concentration of carbon monoxide in an air sample. However, such approaches may not suffice in studying long-term phenomena such as climate change, which often requires seasonal data to be adequately accumulated over decades.

We observed during fieldwork that techniques for publishing data among team members in the lynx monitoring project, for example, comprised using online databases, text files, file transfer protocols or portable disks:

The kind of data that we share with master and PhD students who verify data quality and other research stations affiliated with the project depends on what we are looking out for. It could be only a text file with

information about absence and no absence [of lynx species] in the dataset. We are working on a new system to make it easier to export data files directly from the [machine learning] system. Sometimes we send the data in a text file or in a file with information on location and species, time, and so on. We use a MySQL database just to get the data that is stored up [...]. And we could also just send a bunch of pictures. There is [also] an online solution where we can upload large files that will just stay there for a couple of days. And then they [team members] can download it. We use FTP, a File Transfer Protocol. Other times we have used [hard] disks. They [The processes of publishing data] are not standardised. It depends on whom we are dealing with and what they prefer most and how much data is to be shared. (Freya project leader, interview)

When we asked whether the same tools were used for publishing data openly to the public, we were told, "No, but we share pictures of animal observations with the public [on an open public portal]". (ibid)

#### 8.3.2. Maintaining data

Due to the distributed nature of eLTER sites across geographical boundaries, eLTER data curators and scientists have created a network-level forum that forms an informal community. The curators and scientists bring their understanding of long-term data management, including technologies and practices for maintaining data to be disseminated in their local context, to these network-level activities:

There are legitimate reasons for differences in science work and data management among sites and we must appreciate the heterogeneities. (Pieter, program coordinator, interview)

The network-level community provides an arena for curators and scientists within the eLTER network to collaborate and maintain the infrastructure for data sharing. Such site-based community-level strategies provide a reliable venue for learning about the different scientific data management practices within the network:

It is good to see how other sites are doing things, either as a contrast or as an idea to improve. eLTER sites have taken the time to create a network-level forum that fosters an integrative, sustainable approach with technology, ensuring that we learn together. (Ann, scientist, interview)

A critical practice to promote data sharing in these networks includes investing time and resources to raise awareness on open data sharing and educate all relevant stakeholders about the inextricable link between data curation and data sharing:

Part of my work has been to make them [data curators, engineers, technicians, scientists, students] aware that we are also benefiting from sharing the data. We're not just giving it [data] away. And also, we are getting public funds to do this work. So, in

principle, it's not our personal data either. It's a common good that should be shared. [...] It has been quite a bit of a long road to get people to that point, but we're almost there now. [...]. We need all hands on deck [as different domain experts coordinate their efforts] to solve problems more efficiently. (Tone project manager, field notes)

### 8.3.3. Safeguarding data

Despite these efforts, data dissemination is not yet established in practice, although eLTER scientists agree that open data sharing is essential for the future of science. Our participants offered several explanations for why openly disseminating scientific data is inconsistent. These include insufficient incentives for scientists to publish data, a lack of platforms for managing and publicly disseminating scientific data, and the extra work required. A recurring problem related to the dissemination of scientific data described by our informants was the lack of a system to recognise the work of the researchers who produced, processed, and published the data. One source of frustration in this regard is that the scientist who collected the data also put considerable additional work into curating the data and creating publicly usable datasets that persist on the web for decades. This lack of compensation for the additional work and incentives for the other professionals who maintain the scientific data were voiced as a recurring problem. Some of the informants described this as a cultural problem:

Data sharing requires a cultural and technical shift. I will work hard to make the data available today, but what are the structures to ensure that when I am not here today this shared data persists? (Ragnar, curator, interview)

Some informants expressed concern about the negative consequences of making open data sharing a key performance indicator for local sites to obtain funding:

You are required to share project data as part of the funding terms. If you don't share the data, you will lose future funding opportunities. There is also a trade-off that if you share the data, others may take your data and publish the paper you want to publish based on that data. So, there is some anxiety among some colleagues that they are afraid that they spend much time collecting data, and they only want to share it with other people once they are done analysing it. (Amy, environmental scientist, interview)

For long-term purposes, ad-hoc approaches may be insufficient because publishing data for public use requires curators and scientists to decide not only about which database tools, technologies, or web portals to use but also how to make these tools available to future secondary users, as well as the organisational, security, and privacy concerns associated with the public dissemination of scientific data. To manage

the complexities of publishing long-term ecological data, Nestore and his team at the eLTER-Norway coordinating centre highlighted concerns related to the tools to use to publish data in a way that organisational data are not compromised:

We consider a long-term data-sharing solution based on the project and its privacy needs. With projects that use a local research station's general-purpose portal to publish data publicly, the data curation team must ensure that the portal has the required functionality for managing and sharing the different domain-agnostic data. With projects that have a custom data portal [to publish data with team members outside a research station, but not the public], the data curation team must ensure that data are both internally and publicly available, without compromising private organisational data to unauthorised users. With projects that share data using externally available open data portals, the data curation team must ensure that these external sources are also easily accessible internally. (Nestore, Database Administrator, fieldnotes)

This shows how curators anticipate addressing privacy and intellectual property rights concerns by gatekeeping data to ensure they are published to the right audience.

In sum, curators and scientists disseminate data so future users can access them over decades or centuries. Data dissemination practices consist of the data *publishing work* of ensuring that data are readily available to future users, the *gatekeeping work* of ensuring that only legitimate or intended users have access to shared data, and the *networking work* of ensuring that common elements of governance of local sites (including funding) can guarantee open data over the long term. These practices are forward-looking because they favour a gradual future development of the eLTER open data-sharing infrastructure. This work is done through a combination of sociotechnical practices, such as deciding which emerging technologies to use for public dissemination of data, which data to safeguard and how to control or limit public access to data, training and participating in network-level activities on data management and open data and establishing structures for long-term preservation of data and the associated data infrastructure. Combined with upgrading historical data records and monitoring ongoing data, this work is essential to preserve environmental data over time.

## 9. Discussion: the temporal orientations of anticipatory generification

Data are not pre-given but come to matter for further analysis in the context of situated data curation practices in data infrastructures (Jones, 2019; Parmiggiani et al., 2022; Ribes & Polk, 2014; Waardenburg et al., 2021). However, the challenge is to shed light on how data must be prepared and kept ready for sharing in

uncertain future contexts of reuse. This study complements existing work on data curation in data infrastructures (Jones, 2019; Leonelli, 2019; Parmiggiani et al., 2022; Vassilakopoulou et al., 2019) by further theorising how the longitudinal character of data reusability is addressed in practice. We do this through the concept of anticipatory generification. Anticipatory generification is a form of data curation that ensures that data are sufficiently open-ended for uncertain future (re)use. In our study, environmental researchers achieve open-endedness through practices to keep data relevant, updated, accurate, and accessible over time. We find that anticipatory generification is not only ongoing but also temporally oriented: it follows three partly overlapping temporal perspectives focusing on the past by upgrading historical datasets, the present through ongoing monitoring of data collection streams, and the future through strategies to coordinate data dissemination (Table A1, Figure 1). These three sets of recurrent curation practices are sociotechnical and vary significantly in nature, involving, among others, technical work to update databases and repositories, social connections to retrieve old datasets, liaising with local inhabitants and authorities in addition to installing sensors, and organising and standardising metadata.

The concept of *generification* has been primarily used in the context of software infrastructure in the IS literature. However, we maintain that it is also applicable to the case of digital data in data infrastructures. Pollock et al. (2007) use generification to unearth the work of exporting standard software from production settings and making it sufficiently adaptable to different needs and contexts of use. We use generification to capture the work to make the data collected in a specific setting just about malleable enough to be exported to future unknown analytical contexts. This aspect accentuates the *anticipatory* nature of generification work in data infrastructures, i.e., the need to produce and maintain data so that they will be useful to communicate and generate knowledge in the future despite uncertainty. We find that this immediate concern with the future translates into a practical concern with not only future conditions of use but also past and present data collections. Based on this observation, we further characterise the anticipatory nature of data curators' generification work by fleshing out its three interleaving temporal orientations (Jarvenpaa & Essén, 2023; cf.; Flyverbom & Garsten, 2021).

First, *retrospective* or backwards-looking data curation keeps historical data series relevant and useful. Environmental researchers look at the work that has already been done on the datasets and evaluate if valuable records are available or if additional historical contextual information is required. For data sharing, it can be understood with reference to the traceability of

the data journey (Leonelli, 2016) or pedigree which consists of gathering sufficient information about how and by whom the data were collected and analysed. It can also be seen as a form of repair of the data pedigree, in which data curators enhance the future sustainability of the data infrastructure (Mikalsen & Monteiro, 2021) by ensuring the backward compatibility of datasets. A by-product of this process is improved traceability, as it has been demonstrated that data (re)users, particularly in scientific domains, seldom take data at face value but tend to investigate them based on the chain of information available about who performed the data collection or analysis and how (ibid.). Because the potential value of a dataset in the future can be challenging to establish ex ante, recovering historical data requires a special commitment and a degree of guesswork in which environmental researchers estimate what historical data should be retrieved and prepared. Importantly, retrospective data curation also has a future-looking dimension, as is the case for metadata work: this is a type of work that requires deciding what data about data will be needed tomorrow, but one that relies on the practice of going back to data that have already been collected to describe them. Consequently, future research should shed further light into the dynamic nature of ensuring data traceability.

Second, *present-oriented* curation practices focus on regularly and continuously monitoring data collection projects. This includes handling data as part of a more extensive long-term collection in the now, thus with an eye to retrospective and forward-looking perspectives (cf. Venters et al., 2014). This often requires careful documentation of all changes made. While apparently a form of technical work to track and document data streams, this work is also social. It relies on several under-the-radar activities, such as involving various stakeholders and resolving immediate errors in the data collection flow. The IS literature generally refers to this type of work as related to data quality and agrees that a lack of control over the data flow leads to a loss of data quality in data sharing (Abbasi et al., 2016). However, our findings demonstrate that data quality is not an abstract property of data and hence antecedent for anticipatory generification, but it is accomplished as part of ongoing data curation routines. In our analysis, data of good enough quality are data that are sufficiently accurate and updated with respect to the involved scientists' current understanding of the phenomenon studied. In addition, almost paradoxically, anticipating the generification of data depends on ad-hoc local data work. Our analysis reveals the situated and collective nature of anticipatory generification: data accuracy depends on the data acquisition and verification work conducted by local stakeholders in a field, such as volunteers and guides with whom researchers liaise informally. Albeit

invisible in formal data collection and analysis accounts, these actors have a very influential role in shaping what is to be considered accurate and up-to-date (Waardenburg et al., 2021). Therefore, our study emphasises the need for future research to further study the role of such actors in shaping anticipations of future use as part of data curation processes in the now.

Third, *forward-looking* data curation involves practices associated with disseminating data, ranging from more traditional dissemination practices, such as deciding how to share data, to practices of gatekeeping data to protect scientists' intellectual work and maintaining a research station up-to-speed with funding and technological opportunities. This set of practices relates to traditional accounts of anticipation as work involving envisioning a desired future (Flyverbom & Garsten, 2021). In this sense, anticipation has been described as performative because it has concrete organisational consequences (ibid). Our study demonstrates how anticipatory generification shapes research work's organisation through mundane day-to-day decisions about data and the associated publications. Within the sciences, future-oriented generification work is often associated with the aim of increasing openness, for example, through ideals of increased open data sharing as mandated by policymakers or required to solve complex research problems (Leonelli, 2019). As Monteiro and Parmiggiani (2019; see also Shaikh & Vaast, 2016) note, the push for open data sharing is often rhetorical and is enacted through semi-closed processes of deciding what and how data should be shared in practice. Our study shows how environmental researchers enact informal, local gatekeeping practices to control data release, mindful of intellectual property rights. Therefore, open data sharing goals are met by carefully deciding which data to release and how. Importantly, we also demonstrate that these forward-looking efforts are sustained by the constant work of providing data curators with educational, financial, and training opportunities. Our findings thus invite future research to develop a more nuanced understanding of the actual practices of supporting open data dissemination as a longitudinal and multisided effort.

## 10. Theoretical implications

Our analysis of the temporal horizons that characterise data infrastructures has further theoretical implications for research on anticipatory work and the sustainability of data infrastructures.

First, we extend earlier research on anticipation work (Barley, 2015; Steinhardt & Jackson, 2015; Waardenburg et al., 2021) to enable data reusability by providing further insight into how the future is anticipated in practice. The concept of anticipation is

usually associated with an orientation towards the future. However, we show how three temporal horizons (past, present, and future) are all enfolded into data practices' orientation towards the future. Albeit heterogeneous, all three sets of practices are carried out to create the conditions for future (re)use by simultaneously ensuring that data are ready for present use, (un)expected future use, and reuse and reinterpretation of historical data sets. Although it might be intuitive to think about anticipatory generification as a linear progression in which today's choices will affect tomorrow's choices, we show how it instead unfolds as a dynamic back-and-forth movement (see also Parmiggiani et al., 2022; Venters et al., 2014), i.e., an effort of constantly repairing the past, controlling the present, and preparing for the future (Table 1). While unknown future data reuse is a concern for scientists (as software deployability would be in Pollock and colleagues' (2007) work), this conceptualisation makes clear that "future data reuse" is a moving target that is also shaped by the possibility of handling past and present datasets.

Of course, the salient temporal character of anticipatory generification depends on the epistemological and multidisciplinary nature of the phenomenon under study. The uncertainty associated with retrospective and forward-looking data curation is significant in the environmental domain, as we also illustrate in the findings. This resonates with the situation of geoscientists in the oil and gas domain (Mikalsen & Monteiro, 2021; Parmiggiani et al., 2022). Nonetheless, environmental monitoring is illustrative of a transition towards knowledge work that is increasingly – although not exclusively – grounded on digital data, which by nature have the potential of being more mobile and thus circulate in novel ways compared to traditional archives. The concept of anticipatory generification might be useful to conceptualise practices of making data sufficiently open-ended for uncertain further reuse in other domains. One example is social media (Alaimo et al., 2020), in which data are packaged and sold as commodities across myriad contexts, such as political and commercial profiling related to programmatic advertising (Zuboff, 2019) through circuitous processes to curate, share, and (re)package data into interchangeable goods that are assigned a monetary value (Aaltonen et al., 2021).

Second, anticipatory generification contributes to theorising the sustainable evolution of data infrastructures (Jarvenpaa & Essén, 2023). Instead of a specific phase of implementation or use, anticipatory generification is a form of extended design that shapes data infrastructures through daily use and ongoing upgrade and repair of the datasets over extended periods (Monteiro et al., 2013; Parmiggiani & Grisot, 2020). Our characterisation

of anticipatory generification resembles earlier work on generification work as standardisation in software infrastructures. For example, Hanseth & Bygstad (2015) speak of anticipatory standardisation to mean generification as an official and traditional process. The process we describe, however, is highly informal as it is a bottom-up, emergent process. In this sense, it might be closer to what Bygstad and Hanseth call flexible generification intended as dynamic working solutions that enable the innovation of services that adjust and stabilise over time and achieve legitimacy as a result. In our case, too, the focus of data curators is to devise working solutions that enable future uncertain innovation based on the collected data. However, this work is part of mundane routines that are ongoing and longitudinal (Ribes & Finholt, 2009) and not punctuated along design and implementation projects towards a stable situation. This extended perspective is important to meet the in-the-making character of environmental data, which evolve both locally at a research station and as they are shared with the scientific community and the public. Our data curation-centric and temporally oriented lens on anticipation generification aligns with “a process view [on data] ... driven by the view that data are temporal, co-dependent, indeterminant, and pervasively editable ... This view is less focused on the data as a resource to a delivered service or good and more focused on the value of entanglement of data and operations on data that could take place at any point” (Jarvenpaa & Markus, 2020, p. 72). Therefore, anticipatory generification enables sustainability in data infrastructures by focusing on the maintenance of environmental data as commons (cf. Vassilakopoulou et al., 2019) so that they can be (re)used to analyse and model ecological behaviour over the long term despite persisting uncertainty in data infrastructure.

## 11. Conclusions and practical implications

Environmental data infrastructures are incomplete by design (Garud et al., 2008). They need to ensure open data sharing and simultaneously allow answering evolving research questions. This paper extends existing IS literature on data curation by shedding light on how data are prepared against open-ended future use in environmental monitoring data infrastructure. We describe this work of harnessing data infrastructures’ incompleteness as anticipatory generification. In doing so, we provide a temporally aware account of how data are kept ready for further sharing (Jarvenpaa & Essén, 2023) aimed at better understanding the work required to improve the sustainability of data infrastructures.

This study’s limitations are empirical because our findings relate primarily to the environmental research domain in Norway within a European context. However, we expect concerns related to handling the temporal horizon of data infrastructures to emerge in other domains such as the industry where data are increasingly shared and exchanged via automated interfaces and predefined secured architectures in such settings. This is less the case in environmental monitoring, where data-sharing approaches are typically emergent, often based on projects, and tend to vary locally. In the industrial domain, we would expect additional issues to emerge in relation to cybersecurity and data gatekeeping (Norwegian Ministry of Local Government and Regional Development, 2022).

Our study has practical implications. Due to its temporal orientation, data curation is not a predefined capability ready to be modelled in top-down governance strategies and deployed in data infrastructures. Instead, it is performed ad hoc, punctuated by micro-level, daily bricolage interventions in which researchers and their colleagues learn to ensure that data are good enough for uncertain futures, trying to reconstruct the history of data to a sufficient degree. As a result, data curation is also very time-consuming. This complicates matters for organisations because incentives do not reflect well the temporal orientation of data curation. Misaligned from incentives in a community, it often holds little personal benefit for curators. Research on the use of software for collaboration and information sharing has highlighted that an uneven balance between work and benefits causes much resistance among users, motivating the need for incentives for all users to increase the likelihood of user approval (Plantin, 2021). Future research could look at making data curation explicit in governance frameworks (Parmiggiani & Grisot, 2020). Approaches in the scientific domain could include, for example, promoting more systematic publicly supported data curation training programmes to raise awareness of open data sharing concerns and better aligning career-related and scientific concerns, for example through mechanisms that recognise and reward data sharing, and that align with established systems, such as citations and the H-index. In addition, policymakers at national and EU levels could consider funding more person-hours to take advantage of the “window of opportunity” (Tyre & Orlikowski, 1991) to test solutions to integrate better the practices we have presented into routines in data infrastructures, for example by providing low-threshold funding to hire additional technicians, engineers, and data curators to manage long-term datasets.

## Notes

1. By generic data, we mean that different groups of researchers with different interests are potentially able to relate to the datasets without having been involved in their collection.
2. Approval numbers: 54477 (Parmiggiani), 193948 (Amagyei), and 508,681 (Kollerud).
3. All real names are anonymised.

## Acknowledgements

The argument and perspectives presented in this paper benefited from several conversations with colleagues and collaborators. We are deeply grateful to David Ribes for providing fundamental feedback that helped us focus our conceptualisation of anticipatory generification and to Eric Monteiro for our discussions on the framing and motivation of the study. We warmly acknowledge the environmental organizations and interviewees who provided access and precious insights. We also thank the Senior Editor and the anonymous reviewers for their valuable comments and guidance.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## 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>
- Aanestad, M., Grisot, M., Hanseth, O., & Vassilakopoulou, P., (Eds.). (2017). *Information Infrastructures within European Health Care. Working with the Installed Base*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-51020-0>
- Abbasi, A., Sarker, S., & Chiang, R. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems*, 17(2), I–XXXII. <https://doi.org/10.17705/1jais.00423>
- Alaimo, C., Kallinikos, J., & Valderrama, E. (2020). Platforms as service ecosystems: Lessons from social media. *Journal of Information Technology*, 35(1), 25–48. <https://doi.org/10.1177/0268396219881462>
- Barley, W. C. (2015). Anticipatory Work: How the need to represent knowledge across boundaries shapes work practices within them. *Organization Science*, 26(6), 1612–1628. <https://doi.org/10.1287/orsc.2015.1012>
- Bechky, B. A. (2021). *Blood, Powder, and Residue: How Crime Labs Translate Evidence into Proof*. Princeton University Press. <https://doi.org/10.1515/9780691205854>
- Blanchard, A., Hauge, K. H., Andersen, G., Fosså, J. H., Grøsvik, B. E., Handegard, N. O., Kaiser, M., Meier, S., Olsen, E., & Vikebø, F. (2014, January). Harmful routines? Uncertainty in science and conflicting views on routine petroleum operations in Norway. *Marine Policy*, 43, 313–320. 2014 <https://doi.org/10.1016/j.marpol.2013.07.001>
- Chua, C., Indulska, M., Lukyanenko, R., Maass, W., & Storey, V. C. (2022, February 14). *Data Management*. MIS Quarterly Research Curations. <https://www.misqresearchcurations.org/blog/2022/2/11/data-management>
- de Reuver, M., Sørensen, C., & Basole, R. C. (2018). The digital platform: A research agenda. *Journal of Information Technology*, 33(2), 124–135. <https://doi.org/10.1057/s41265-016-0033-3>
- Eisenhardt, K. M. (1989). Building theories from case study research. *The Academy of Management Review*, 14(4), 532–550. <https://doi.org/10.2307/258557>
- Eisenhardt, K. M., & Graebner, M. E. (2007). Theory building from cases: Opportunities and challenges. *Academy of Management Journal*, 50(1), 25–32. <https://doi.org/10.5465/amj.2007.24160888>
- Emerson, R. M., Fretz, R. I., & Shaw, L. L. (2011). *Writing Ethnographic Fieldnotes* (2nd ed.). University Of Chicago Press. <https://doi.org/10.7208/chicago/9780226206868.001.0001>
- Flyverbom, M., & Garsten, C. (2021). Anticipation and organization: Seeing, knowing and governing futures. *Organization Theory*, 2(3), 26317877211020324. <https://doi.org/10.1177/26317877211020325>
- Garud, R., & Gehman, J. (2012). Metatheoretical perspectives on sustainability journeys: Evolutionary, relational and durational. *Research Policy*, 41(6), 980–995. <https://doi.org/10.1016/j.respol.2011.07.009>
- Garud, R., Jain, S., & Tuertscher, P. (2008). Incomplete by Design and Designing for Incompleteness. *Organization Studies*, 29(3), 351–371. <https://doi.org/10.1177/0170840607088018>
- Hanseth, O., & Bygstad, B. (2015). Flexible generification: ICT standardization strategies and service innovation in health care. *European Journal of Information Systems*, 24(6), 645–663. <https://doi.org/10.1057/ejis.2015.1>
- 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. In R. Hirschheim, A. Heinzl, & J. Dibbern (Eds.), *Information Systems Outsourcing: The Era of Digital Transformation* (pp. 61–79). Springer International Publishing. [https://doi.org/10.1007/978-3-030-45819-5\\_4](https://doi.org/10.1007/978-3-030-45819-5_4)
- Johansen, J. P., Almklov, P. G., & Mohammad, A. B. (2016). What can possibly go wrong? Anticipatory work in space operations. *Cognition, Technology & Work*, 18(2), 333–350. <https://doi.org/10.1007/s10111-015-0357-8>
- Jones, M. (2019). What we talk about when we talk about (big) data. *Journal of Strategic Information Systems*, 28(1), 3–16. <https://doi.org/10.1016/j.jsis.2018.10.005>
- Kaltenbrunner, W. (2017). Digital Infrastructure for the Humanities in Europe and the US: governing scholarship through coordinated tool development. *Computer Supported Cooperative Work (CSCW)*, 26(3), 275–308. <https://doi.org/10.1007/s10606-017-9272-2>
- Karasti, H., Botero, A., Baker, K. S., & Parmiggiani, E. (2018). *Little data, big data, no data? data management in the era of research infrastructures*. Oulu, Finland: University of Oulu, Finland. <https://www.ideals.illinois.edu/handle/2142/100870>
- Klein, H. K., & Myers, M. D. (1999). A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems. *MIS Quarterly*, 23(1), 67–94. <https://doi.org/10.2307/249410>
- Knowledge, O. (n.d.). *What is Open Data?* Open Data Handbook. Retrieved April 17, 2023, from <https://opendatahandbook.org/guide/en/what-is-open-data/>
- Leonelli, S. (2016). *Data-Centric Biology: A Philosophical Study*. University of Chicago Press. <https://chicago.universitypressscholarship.com/view/10.7208/chicago/9780226416502.001.0001/upso-9780226416335>



- Leonelli, S. (2019). Data—From objects to assets. *Nature*, 574 (7778), 317–320. <https://doi.org/10.1038/d41586-019-03062-w>
- Link, G., Lombard, K., Feldman, K., Conboy, M., Feller, J., George, J., Germonprez, M., Goggins, S., Jeske, D., Kiely, G., Schuster, K., & Willis, M. (2017). Contemporary issues of open data in information systems research: considerations and recommendations. *Communications of the Association for Information Systems*, 41(1), 587–610. <https://doi.org/10.17705/1CAIS.04125>
- Lukyanenko, R., Parsons, J., Wiersma, Y. F., Maddah, M., & Suffolk University. (2019). Expecting the unexpected: effects of data collection design choices on the quality of crowdsourced user-generated content. *MIS Quarterly*, 43 (2), 623–647. <https://doi.org/10.25300/MISQ/2019/14439>
- Mikalsen, M., & Monteiro, E. (2021). Acting with inherently uncertain data: Practices of data-centric knowing. *Journal of the Association for Information Systems*, 22(6), 1715–1735. <https://doi.org/10.17705/1jais.00722>
- Monteiro, E., & 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, E., Pollock, N., Hanseth, O., & Williams, R. (2013). From Artefacts to Infrastructures. *Computer Supported Cooperative Work (CSCW)*, 22(4–6), 575–607. <https://doi.org/10.1007/s10606-012-9167-1>
- Myers, M. D., & Newman, M. (2007). The qualitative interview in is research: Examining the craft. *Information and Organization*, 17(1), 2–26. <https://doi.org/10.1016/j.infoandorg.2006.11.001>
- Nicolini, D. (2009). Articulating Practice through the Interview to the Double. *Management Learning*, 40(2), 195–212. <https://doi.org/10.1177/1350507608101230>
- Nicolini, D. (2012). *Practice Theory, Work, and Organization: An Introduction*. Oxford University Press.
- Norwegian Ministry of Local Government and Regional Development. (2022, May 31). *Deling av industridata [Sharing of industrial data]* [Rapport]. Regjeringen.no; regjeringen.no. <https://www.regjeringen.no/no/dokumenter/deling-av-industridata/id2916456/>
- 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>
- Parmiggiani, E., & Grisot, M. (2020). Data Curation as Governance Practice. *Scandinavian Journal of Information Systems*, 32(1), 1–36, Article 1.
- Parmiggiani, E., Østerlie, T., & Almklov, P. (2022). In the Backrooms of Data Science. *Journal of the Association for Information Science and Technology*, 23(1), 139–164. <https://doi.org/10.17705/1jais.00718>
- Plantin, J.-C. (2021). The data archive as factory: Alienation and resistance of data processors. *Big Data & Society*, 8 (1), 1–12. <https://doi.org/10.1177/205395172111007510>
- Pollock, N., & Williams, R. (2009). *Software and organisations: the biography of the enterprise-wide system or how sap conquered the world*. Routledge.
- Pollock, N., Williams, R., & D’Adderio, L. (2007). Global software and its provenance: generification work in the production of organizational software packages. *Social Studies of Science*, 37(2), 254–280. <https://doi.org/10.1177/0306312706066022>
- Ribes, D., & Finholt, T. A. (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>
- Rubin, H. J., & Rubin, I. S. (2011). *Qualitative Interviewing: The art of hearing data*. SAGE.
- Schultze, U., & Avital, M. (2011). Designing interviews to generate rich data for information systems research. *Information and Organization*, 21(1), 1–16. <https://doi.org/10.1016/j.infoandorg.2010.11.001>
- Shaikh, M., & Vaast, E. (2016). Folding and Unfolding: Balancing openness and transparency in open source communities. *Information Systems Research*, 27(4), 813–833. <https://doi.org/10.1287/isre.2016.0646>
- Silsand, L., & Ellingsen, G. (2014). Generification by Translation: Designing Generic Systems in Context of the Local. *Journal of the Association for Information Systems*, 15 (4), 177–196. <https://doi.org/10.17705/1jais.00358>
- Steinhardt, S. B., & Jackson, S. J. (2015). Anticipation Work: Cultivating vision in collective practice. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 443–453). Vancouver, BC, Canada.
- Suddaby, R. (2006). From the Editors: What Grounded Theory is Not. *Academy of Management Journal*, 49(4), 633–642. <https://doi.org/10.5465/amj.2006.22083020>
- Tempini, N., & Leonelli, S. (2018). Concealment and discovery: The role of information security in biomedical data re-use. *Social Studies of Science*, 48(5), 663–690. <https://doi.org/10.1177/0306312718804875>
- Tjora, A. (2018). *Qualitative Research as Stepwise-Deductive Induction* (1st ed.). Routledge. <https://doi.org/10.4324/9780203730072-1>
- Tyre, M. J., & Orlikowski, W. J. (1991). *Windows of opportunity —Creating occasions for technological adaptation in organizations* [Working Paper]. Alfred P. Sloan School of Management, Massachusetts Institute of Technology. <https://doi.org/https://dspace.mit.edu/handle/1721.1/49397>
- 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>
- Velsberg, O., Westergren, U. H., & Jonsson, K. (2020). Exploring smartness in public sector innovation—Creating smart public services with the Internet of Things. *European Journal of Information Systems*, 29(4), 350–368. <https://doi.org/10.1080/0960085X.2020.1761272>
- Venters, W., Oborn, E., & Barrett, M. (2014). A trichordal temporal approach to digital coordination: the sociomaterial mangling of the cern grid. *MIS Quarterly*, 38(3), 927–949. <https://doi.org/10.25300/MISQ/2014/38.3.13>
- Waardenburg, L., Huysman, M., & Sergeeva, A. V. (2021). In the Land of the blind, the one-eyed man is king: knowledge brokerage in the age of learning algorithms. *Organization Science*, 33(1), 59–82. <https://doi.org/10.1287/orsc.2021.1544>
- Walsham, G. (2006). Doing interpretive research. *European Journal of Information Systems*, 15(3), 320–330. <https://doi.org/10.1057/palgrave.ejis.3000589>
- Yin, R. K. (2009). *Case Study Research: Design and Methods*. SAGE Publications.
- Zuboff, S. (2019). *The age of surveillance capitalism: the fight for a human future at the new frontier of power* (1 ed.). PublicAffairs.

**Appendix  
Appendix A**

**Table A1.** The interpretive template resulting from our analysis.

Categories (Anticipatory generation practice with temporal perspective)	Constructs (data curation practice)	What People Do	Excerpts
<b>Upgrading Data - Looking Backward</b>	Discovering Data	Find expected and unexpected data in paper records	<p>"No, until now we store it in our paper archive, it follows a project. The Council is starting to set more demands on the availability of data, and we are constantly working on new solutions as to how we do it". (Tone, scientist, interview)</p> <p>"We are analysing nutrient chemistry in the water. And then we're also looking at phytoplankton, that's microalgae. And then we're looking at zooplankton and then we're also looking at fish larvae and eggs of fish because they are the primary consumers of this micro and mesozooplankton. So, we have four different categories, and they all have different details here about how to do this ... I've been working for five years now to get the different data that are basically sitting on different Excel sheets to become part of a bigger Data Archive, and to make it accessible to others and the public. And that's been a long road, because we've had nutrient chemistry data, for instance, we have more than 30 years of data". (Eva, scientist, interview)</p>
	Recovering Data	How anticipatory practices create prerequisites for the future	<p>Relevant data needed to do long-term science are "hidden" in many different sources. This data <i>digitising work</i> creates opportunities for future users to find useful but siloed data.</p> <p>Data in legacy systems that are inaccessible to the public and have different formats from existing systems for sharing public data, require curators to consider which technologies and data formats to use in a way that historical data are restored and adaptable for use in a different future setting. This data <i>repurposing work</i> ensures that shared data are easily interoperable with other datasets in the future.</p>
<b>Upgrading Data - Looking Forward</b>	Recovering Data	Obtain legacy data and avoid technological obsolescence	<p>"Sometimes we interview scientists, especially those who are retiring to get as much (meta)data as possible. We can also go to [scientists'] offices to copy files directly. We sometimes get data from older paper files in our archives" (Jude, data curator, interview)</p> <p>"We describe data with the relevant information to help another user in the future to understand the settings and tools used for collecting the data. This way, they trust the data" (Anna, senior researcher, interview)</p>
	Describing Data	Describe digitised records with contextual information usually through informal interactions with primary data producers	<p>Data curators strive to retrieve information about the context of data collection, also by relying on the knowledge of the scientists who collected the primary data. This <i>metadata work</i> creates opportunities for future users to understand the context of shared data.</p> <p>"The data stewards have made a script so that we can extract the data from our older database to a format that can work with the new one [...] but the main task [...] is [ensuring that the retrieved data] has all the right tags. So, it's not like a [completely] automated procedure". (Tomas, Environmental Researcher, field notes)</p> <p>"Sometimes we interview scientists, especially those who are retiring to get as much (meta)data as possible. We can also go to [scientists'] offices to copy files directly. We sometimes get data from older paper files in our archives" (Jude, data curator, interview)</p> <p>"We describe data with the relevant information to help another user in the future to understand the settings and tools used for collecting the data. This way, they trust the data" (Anna, senior researcher, interview)</p>

(Continued)

Table A1. (Continued).

Categories (Anticipatory generation practice with temporal perspective)	Constructs (data curation practice)	What People Do	How anticipatory practices create prerequisites for the future	Excerpts
<b>Monitoring Data</b> - Looking to the present	Acquiring Data	Liaise with local stakeholders in the physical environment where data are collected	Curators generate data collection criteria and partner with institutions and landowners to enact them. This data <i>generation work</i> ensures that relevant data are collected and available to future users.	<p>“These local people know more about where these animals go. And they would also be able to get the consensus from those who own that land. There are many people that are very restricted about monitoring, especially the large carnivores. [...], it's almost impossible for us to go to ask the people out on the ground [if we] can put up a camera. But if a local person asks, they will say more often than if we asked, because they trust [other] locals more, and they don't trust the government. Even though we are not working for the government, they look at us in the same way”. (Freya, Project Manager, field notes).</p> <p>“We have these people who are working at the Hunting and Fishing Agency. They ask on behalf of us. So, they are the ones who do the hiring of the locals. And we are hiring the organisation and the organisation ... gets the local people (...). We have approximately a little bit more than 100 persons [local volunteers] around in Norway who help us with collecting data from cameras. (Freya, Project Manager field notes).</p>
Verifying data	Enrol local expertise to perform data collection	To ensure that collected data are accurate for future users, curators involve local experts and other professionals to verify ongoing data collection. This data <i>validation work</i> ensures that shared data are reliable for use by future users.		<p>“They [the local guides] will be able to look at the data [on the camera] before they send the memory card to us. That also helps with the trust issues, the conflict, they will trust the data more and they trust the government more when they're seeing this data by themselves. So, we can see that it reduces the conflicts a lot when we are using these local guides sometimes it's a little bit more work for us, but it makes it possible to monitor these large carnivores”. (Freya, Project Manager field notes).</p> <p>“It can be difficult to see in the dark. All pictures there isn't erh... a lot of pictures have quite bad quality because All pictures there aren't erh... a lot of pictures have quite bad quality because most are out at night. And it's forest and it's windy and it's raining and it's a lot of bushes. It can be hard [for the machine learning program] to be sure. So often we are using a lot of students to go over all these pictures to verify each time series”. (Freya, Project Manager field notes).</p>

(Continued)

**Table A1.** (Continued).

Categories (Anticipatory generation practice with temporal perspective)	Constructs (data curation practice)	What People Do	How anticipatory practices create prerequisites for the future	Excerpts
Updating Data	Resolve seasonal cycles of data collections	Fix missing values and address immediate concerns related to sharing recurring long-term data – this data <i>resolution work</i> ensures that shared data are up to date.		<p>"We do data wrangling also in terms of biological data which often involves matching scientific names, Latin names of species with Norwegian and English names. If I'm going to make a report for Popular Science report and a report for the Norwegian government, I need to translate the species from Latin names to Norwegian names ... it's often a matter of not just matching the species, but ensuring that the species that were first reported are the correct scientific name that's currently in use for that species because they change over time as species are reclassified to make sure that you're actually using the most up to date, species name. So those things are typical data-wrangling issues. Then you have of course rooting out errors ... there are several quality control checks before the data are shared. But still, there are some errors that pop up every now and then". (Nils, Data curator interview)</p> <p>"Yeah, because of how it (the stone is) shaped and so, and if those who have verified the pictures [...] a little bit too fast. It can be stored as a lynx, so we [scientists and curators] have, from time to time, been going through to just look. Recently one student added some of those rarer species. We can't do that". (Freya, Project Manager interview)</p>
<b>Disseminating Data - Looking Forward</b>	Delivering Data	Decide on which database tools, standards, web-portals, etc. to distribute data	Curators ensure that data are accessible to both designated scientists and future (re)users, on a day-to-day basis – this data <i>publishing work</i> ensures that enough data are easily accessible to the others.	<p>"The kind of data that we share with Master and PhD students who verify data quality and other research stations affiliated with the project depends on what we are looking out for. It could be only a text file with information about absence and no absence [of lynx species in the dataset]. (...) Sometimes we send the data in a text file or in a file with information on location and species, time, and so on. We use a MySQL database just to get the data that is stored up [...] And we could also just send a bunch of pictures. There is [also] an online solution where we can upload large files that will just stay there for a couple of days. And then they (team members) can download it. We use <i>FTP or File Transfer Protocol</i>. Other times we have used (hard) disks. They [The processes of publishing data] are not standardised. It depends on whom we are dealing with and what they prefer most and how much data is to be shared". (Freya project leader, interview)</p>

(Continued)

Table A1. (Continued).

Categories (Anticipatory generification practice with temporal perspective)	Constructs (data curation practice)	What People Do	How anticipatory practices create prerequisites for the future	Excerpts
Maintaining Data	Determine strategies for personnel, data management and funding structures to preserve long-term data and maintain associated data sharing infrastructure	Keep abreast with local data management strategies and technologies, organise training, and raise awareness on the role of data curation in data sharing through network-level activities- this <i>networking work</i> ensures that data can be shared and evolve in the long term in a way that also supports local science work.	<p>"Part of my work has been to make them [data curators, engineers, technicians, scientists, students] aware that we are also benefiting from sharing the data. We're not just giving it [data] away. And also, we are getting public funds to do this work. So, in principle, it's not our personal data either. It's a common good that should be shared. [...] It has been quite a bit of a long road to get people to that point, but we're almost there now. [...] We need all hands on deck [as different domain experts coordinate their efforts] to solve problems more efficiently". (Tone project manager, field notes)</p> <p>"It is good to see how other sites are doing things, either as a contrast or as an idea to improve. eLTER sites have taken the time to create a network-level forum that fosters an integrative, sustainable approach with technology, ensuring that we learn together". (Ann, scientist, interview)</p>	
Safeguarding Data	Control or limit general access to data	Adhere to guidance or legal requirements for sharing data – this <i>gatekeeping work</i> ensures that intellectual property and general data protection rights requirements are met, and that data are accessible to only legitimate or intended users.	<p>"We consider a long-term data-sharing solution based on the project and its privacy needs. With projects that use a local research station's general-purpose portal to publish data publicly, the data curation team must ensure that the portal has the required functionality for managing and sharing the different domain-agnostic data. With projects that have a custom data portal [to publish data with team members outside a research station, but not the public], the data curation team must ensure that data are both internally and publicly available, without compromising private organisational data to unauthorised users. With projects that share data using externally available open data portals, the data curation team must ensure that these external sources are also easily accessible internally". (Nestore, Database Administrator, fieldnotes)</p> <p>"You are required to share project data as part of the funding terms. If you don't share the data you will lose future funding opportunities. There is also a trade-off that if you share the data, others may take your data and publish the paper you want to publish based on that data. So, there is some anxiety among some colleagues that they are afraid that they spend much time collecting the data, and they only want to share it with other people once they are done analysing it". (Army, environmental scientist, interview)</p>	