Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Maria Ivanova

Data Preparation and Sharing Practices: A Case of Environmental Monitoring in Norway

An interpretive case study

Master's thesis in Computer Science
Supervisor: Elena Parmiggiani
Co-supervisor: Nana Kwame Henebeng Amagyei

June 2022

NTNU
Norwegian University of
Science and Technology

Maria Ivanova

Data Preparation and Sharing Practices: A Case of Environmental Monitoring in Norway

An interpretive case study

Master’s thesis in Computer Science
Supervisor: Elena Parmiggiani
Co-supervisor: Nana Kwame Henebeng Amagyei
June 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Abstract

Environmental data provides us with knowledge about the state of our environment, and is used as a foundation for informing, establishing, and revising environmental policy. Thus, the quality of this data, and the ability of this data to reflect the real world are crucial for policy decision-makers. Still, there seems to be a gap in existing literature when it comes to how this quality is achieved through relations researchers have with one another and their data, with most studies focusing on technical and practical measures.

Previous literature has identified challenges in preparing and sharing environmental data, with the focus being on external influences such as resources or policy. With existing studies documenting these interrelationships, it seems fitting to look at this from a different perspective, exploring the internal processes of environmental data work, how they influence each other, and how external factors play into the preparation and sharing of data.

This project consists of an explanatory case study, with an interpretive approach. The aim of the study is to contribute with empirical data on the coordinative and cooperative practices of researchers preparing and sharing data in Norwegian environmental research stations, and how these practices influence data quality. The case study draws on qualitative data in the form of interviews, observations and document studies. The findings identify key roles and dependencies in environmental data preparation. Furthermore, the findings suggest that data preparation is a cooperative activity that relies on multiple dependencies to function, and that the quality of the resulting data is a product of these dependencies. Additionally, the findings suggest that data sharing activities in Norwegian non-profit organizations are often much less pressured by regulatory measures in contrast to previous literature on the topic, suggesting that there is substantial variance in matters of local policy implementation.

Sammendrag

Miljødata blir brukt til å informere, begrunne, og evaluere miljøpolitikk. Dermed er det avgjørende at denne dataen både er av høy kvalitet, og at den er presis i sin gjenspeiling av den virkelige verden. Hvordan forskere oppnår denne kvaliteten på data har blitt undersøkt fra tekniske og praktiske perspektiv, med lite fokus på innvirkning fra forskere og deres interaksjon.

Tidligere litteratur har identifisert utfordringer forskere opplever i å forberede og dele data, der fokuset er eksterne innflytelser slik som mangel på ressurser og utfordrende retningslinjer. Det er derimot lite fokus på interne prosesser og hvordan disse påvirker hverandre, samt hvilken rolle eksterne innflytelser har på forberedning og deling av data.

Dette prosjektet består av en fortolkende case-studie. Prosjektet har som mål å bidra med ny empirisk innsikt i hvordan samarbeid og koordinativ praksis i dataforberedelse og datadeling foregår på norske miljøforskningsinstitutt, og hvordan denne praksisen påvirker datakvalitet. Case-studien er basert på kvalitative data i form av intervjuer, observasjoner, og dokumentasjon. Funnene i studien identifiserer viktige roller og avhengigheter i dataforberedelsesprosesser. Videre antyder funnene at dataforberedelse blir påvirket av flere typer avhengigheter, og at disse videre påvirker datakvalitet. Funnene viser også at forskere fra norske forskningsinstitusjoner ikke alltid opplever press i form av retningslinjer og annen politikk i like stor grad som tidligere forskning viser, som kan være tegn på at innføring og håndheving av disse varierer i stor grad, samt er avhengig av lokal politikk.

Acknowledgements

Firstly, I would like to give a big thank you to my supervisor, Elena Parmiggiani, for always giving great feedback and overall being a source of wisdom when needed.

Additionally, I want to thank my co-supervisor Nana Kwame Henebeng Amagyei, for being a needed discussion partner, and accompanying me on my interviews and subsequent travels.

I would also like to thank all my wonderful informants, without whom this thesis would not be possible. Thank you for taking time out of your already busy schedules in order to enrich my understanding of your field, and for those that invited me to your workplace - an extra special thank you for providing me with unique insight I otherwise would never have had the chance to acquire.

I want to thank my partner, Bjørn Kåre S., for the endless support and understanding, which carried on even when things got tough.

Additionally, I want to thank my dear friend Henriette A. for the outpour of motivational support, study room snack provisions, and overall great company while writing our theses together.

Lastly, I want to thank my parents for always having my back.

Trondheim, June 2022

Maria Ivanova

Contents

Figures

Tables

Acronyms

CSCW - Computer Supported Cooperative Work

eLTER - European Long-term Ecosystem Research

ESFRI - European Strategy Forum on Research Infrastructures

II - Information Infrastructure

PD - Practical Dependency

QA - Quality Assurance

QC - Quality Control

RI - Research Infrastructure

SDI - Stepwise Deductive Induction

TD - Trust Dependency

Chapter 1

Introduction

This chapter introduces the purpose of this project, as well as the resulting research questions. Furthermore, the contributions, limitations of the scope, and the structure of this thesis will be outlined. Some of the material in the next section extends the work from my literature review [1], done in the fall of 2021.

1.1 Project purpose

With the rising focus on environmental concerns in the last decades, the EU has taken charge, introducing regulative policies with environmental research data as a vital asset in their decision-making [2–4]. Environmental research is an interdisciplinary academic field that seeks to increase our understanding of the environment, and give insight into natural and human-caused changes and how they affect us. Often, this research is done through long-term monitoring projects that span decades, and monitor the changes in different environmental aspects such as air, water, soil, and wildlife. These projects depend on great amounts of resources in the form of funding, tools (e.g. sensors and analytical instruments), information systems, and researchers.

Data preparation is the act of transforming data from the raw state of retrieval into a finalized version that creates meaning and coherence. The data produced as a result of environmental research projects often influences and determines new environmental laws and regulations. Consequently, it is imperative that this data accurately reflects the environment it models. Thus, the influences on the *quality* of this data, and what quality means to the researchers preparing and sharing this data is essential to understand.

Furthermore, environmental research projects are often situated in larger research infrastructures, or RIs for short. RIs have the promise of '...community wide and cross-disciplinary collaboration, computationally driven collection, representation and analysis of data, and end-to-end integration' [5, p. 232]. They are highly heterogeneous in nature, as they comprise environmental research sites, tools, users, practices and communities which are all interconnected and dynamically evolving [6]. The complexity and relative novelty of RI development brings along a plethora of tensions and challenges, some of which have been addressed in literature before [7–9]. As preparing and sharing environmental data are central functions of RIs, one can expect that tensions and challenges will affect, and will be affected by their realization in practice [10].

Long-term Ecosystem Research in Europe, commonly referred to as eLTER, is a network recognized as '...a key component of global efforts to better understand the structure and functions of ecosystems and their response to environmental, societal and economic drivers' [11, p. 632]. In the efforts to pursue its main mission of facilitating high-impact research, eLTER plans to develop a common research infrastructure, eLTER RI [12], with the goal to make data and training a shared commodity for all participating research sites, thereby minimizing fragmentation of scientific resources [8]. Additionally, the goal of eLTER RI is to minimize policy fragmentation by enforcing agreed policies across all participating sites. Policy, in this project, is seen as the requirements and guidelines set by policymakers, defining, affecting and restricting research practices and objectives.

Several literary works have addressed the challenges researchers experience as a result of policy and resources not adequately supporting the retrieval, preparation, sharing, and maintenance work that goes into research data, as well as the importance of these processes [10, 13–15]. Furthermore, ensuring quality data has traditionally been handled with technical and practical measures such as establishing methods, developing and implementing tools, and providing training for researchers [16, 17]. However, few studies have actually looked at how the practices of preparing and sharing data *are handled* at research sites[18], and how interactions between researchers influence this. There seems to be a gap in literature when it comes to how researchers prepare data, and ensure quality of this data, as well as how influences such as policy and the potential sharing of this data acts on this preparation. The lack of insight into local practices means that there are potentially two unmanaged influences on data quality simultaneously working in a top-down and a bottom-up manner; one coming from the policy-setters and ultimately restricting researchers' work with data, and the other coming from the researchers, affecting the infrastructure which they are a part of [7, 9, 10, 19]. Thus, my aim is to firstly look at how researchers cooperate to prepare *quality* data, and how this is coordinated on a local level. Secondly, my aim is to explore the practices of data sharing and how these, in addition to data sharing policy act as potential influences on data preparation activities.

1.2 Research questions

This master thesis is a continuation of a literature review conducted in the fall of 2021. In it, literature on environmental researchers in the context of infrastructures, as well as studies and theory on coordinative and cooperative practices were analysed. This thesis consists of a case study where the objective is to examine the coordinative and cooperative processes of researchers preparing data, as well as their data sharing practices and the policies that influence them. Thus, the research questions are the following:

- RQ1: What are the drivers of quality in data preparation?
    - RQ1.1: What role do coordinators have in data preparation, and what expertise is necessary to realize this role in practice?
    - RQ1.2: What dependencies do data workers face when preparing data?
- RQ2: What (dis)incentivizes researchers to share data?
    - RQ2.1: How does policy fragmentation influence data sharing?

1.3 Contributions

The aim of this project is to contribute with new empirical knowledge into the coordinative and cooperative processes of researchers preparing data in the context of environmental research sites in Norway. Furthermore, the aim is to provide insight into their data sharing activities, and how these activities, as well as the policies governing them, might influence data preparation. The findings and subsequent discussion can be used to further examine the new insights, themes, and analytical framework proposed in this thesis [20].

1.4 Limitations of the scope

This thesis adapts a socio-technical perspective, and does thus not focus on the technological (such as information systems, sensors, or technical tools) or scientific (such as elaborate descriptions of methods of analysis or data collection) details that are part of the studied environment. Furthermore, the data collected in this study was exclusively from a Norwegian environmental research infrastructure setting, with some insights into the handling of infrastructures at the EU level. Thus, one can expect that substantial differences can be observed outside this scope.

1.5 Structure of thesis

The structure of this thesis is as follows:

Chapter 2 Literature background. Presents and discusses key concepts and existing literature pertaining to these. The chapter is divided into four main themes that lay the theoretical foundation needed to carry out my case study, and support my discussion in order to answer my research questions.

Chapter 3 Case. Presents a backdrop to the premises and environment in which the study takes place.

Chapter 4. Presents the chosen research strategy and paradigm, the steps of participant recruitment, as well as the chosen method for data analysis and how this was implemented. Additionally, you will find the analytical framework of this thesis at the end of this chapter.

Chapter 5. Presents the findings of the case study. Follows the structure of the analytical framework presented in chapter 4.

Chapter 6. Discusses the findings in relation to the theoretical background presented in chapter 2.

Chapter 7. Reviews the key points of the findings and discussion. Additionally, the limitations to the work of this thesis, as well as suggestions for further work can be found here.

Appendix A. Contains information letters, consent forms and interview guides used for interviews and observations.

Appendix B. Contains Norwegian Centre for Research Data (NSD) application and acceptance letter.

Chapter 2

Literature Background

In this chapter, concepts and challenges I have found especially relevant to my research will be looked into. I have chosen to divide the literature background into four primary sections. Before diving into these sections, a brief background on computer supported cooperative work (CSCW) will be given. Then, the first section introduces infrastructures as a backdrop to the remaining theory. In the section on infrastructures, research and information infrastructures will be introduced. Furthermore, fragmentation and the role of infrastructure users will be looked into. The next two sections follow the chronology of the data lifecycle: First, data curation and its challenges will be introduced. Then, (open) data sharing will be discussed. This section is divided into three main parts: policy on data sharing, challenges of sharing data, and challenges of reusing data. Lastly, cooperation and its coordination in the context of research infrastructures, data curation, and data sharing will be presented, followed by a theoretical background on the role of trust in cooperative work.

Since there is an overall lack of existing literature on environmental data preparation and sharing in the context of cooperative and coordinative practices, I will sometimes draw on studies from other areas of research which I find to be relevant, and where I believe the observations are transferable to my work. Note that a substantial amount of the work presented here is taken from my literature review [1], which was conducted in the fall of 2021.

2.0.1 Computer Supported Cooperative Work (CSCW)

Before diving into the rest of the theory presented in this chapter, we will first take a look at the term computer supported cooperative work(CSCW). CSCW is an interdisciplinary research area first introduced in a 1984 workshop where interested parties collectively looked at how one can use technology more effectively to support people in their work [21]. Later, CSCW has been described as a field concerned with '...support requirements of cooperative work arrangements' [22, p. 5], but the general idea remains the same. As the field of CSCW has evolved, CSCW researchers have adopted two main viewpoints: technology-centric and work-centric [21]. The former is concerned with *design practices* supporting the development of technology to support cooperation, while the latter focuses on the *understanding of work practices* with the aim to use this understanding for technology development [21]. The literature presented in this chapter adheres to the latter category, as this is relevant for my research questions.

One issue addressed in CSCW is the existence of a social-technical gap, described as '...the divide between what we know we must support socially and what we can support technically' [23, p. 1]. Developers are aware that there is a need for nuance, flexibility, and contextualization [23] in CSCW, and yet the issue pertains several decades after it was first addressed [24]. Not only does the issue pertain, but it grows more complex when adding policy to the preexisting social and technical aspects [24]. Jackson, Gillespie, and Payette present the notion that '...three-way intersections between design, practice, and policy show up with particular complexity and importance during periods of formation and emergence...' [24, p. 5] which is arguably relevant in the context of research infrastructure emergence and its political constraints.

2.1 Infrastructures

One of the main incentives for infrastructure development, and one of the main reasons for infrastructure evolution is open science [25]. The opening of science drives the development of shared systems and standards [26, 27], and these consequently lead to the conception and evolution of infrastructures. It is thus interesting to take a closer look at what the term *infrastructure* encompasses, and how it materializes.

Infrastructures are big, multilayered and complex [28]. Their definitions and descriptions are consequently varied and plenty. It is important to note that there are two[1] relevant types of infrastructures for my research; information infrastructures

[1] I recognize that the EU and literature pertaining to research within the EU additionally uses

(IIs) and research infrastructures (RIs)[2]. These are both studied in the context of research communities. IIs are characterized by '...openness to number and types of users (no fixed notion of 'user'), interconnections of numerous modules/systems (i.e. multiplicity of purposes, agendas, strategies), dynamically evolving portfolios of (an ecosystem of) systems and shaped by an installed base of existing systems and practices (thus restricting the scope of design, as traditionally conceived)' [6, p. 577]. Using this definition, it is clear that IIs are comprised of technical systems and related data, as well as the people and policies that constitute the communities present in the infrastructure [29]. RIs encompass all these qualities, but are also 'knowledge infrastructures' [30, 31], in that they additionally are especially concerned with *research being distributed and collaborative* in the long term, as these are seen as key elements of promoting sustainable research development[29, 32]. This is not to say that RIs and IIs are two independent and separate entities; one can argue that in the presence of a research infrastructure there will always be an information infrastructure. The goal of eLTER RI to create a shared eLTER service portal providing access to data and sites might very well be one example of this [12]. The idea of intertwined infrastructures is supported by Star's [28] notion that an infrastructure '...takes on transparency by plugging into other infrastructures and tools in a standardized fashion' [28, p. 381].

2.1.1 Fragmentation of infrastructures

One of the ideas behind research infrastructures, and infrastructures in general, is that providing a common infrastructure for research minimizes fragmentation of scientific resources[8]. However, existing infrastructure efforts often have poor coordination '...both at the level of research sites and across research sites...' [14, p. 5], and existing differences between top-down and bottom-up approaches to research (with policy imposing unification and standardization onto inherently heterogeneous work practices, which unintentionally creates new exclusions and gaps in existing infrastructures) [9] means that solving the issues of fragmentation is not as simple of a task as originally intended. As the foundation on which data is created and curated varies greatly in different research sites [26], one is left with the problem of combining supranational research agendas with heterogeneous research that is highly dependent on local environment. In the EU, this is further complicated as research policy is fragmented itself, with '...consensus for

other highly relevant 'infrastructure terms' such as e-Science and e-Infrastructure. I have chosen two arguably intuitively understandable terms to work with, and do not see the need of introducing more variability into the definitions and terms of infrastructure.

[2] Specifically environmental RIs (ERIs), in the context of eLTER RI [12].

the creation of the necessary supra-national budgets ... unlikely to emerge' [8, p. 5]. However, policy structures in the EU clearly demonstrate a goal to overcome fragmentation, which leaves hope that the development of infrastructures in eLTER and similar networks might be able to tackle fragmentation in the future [8].

2.1.2 The role of infrastructure users

Infrastructures evolve in 'modular increments' in the sense that they emerge gradually (and, ideally, organically [27]) [28]. This change is not one that anyone is specifically in charge of; rather, users of infrastructure are inherent developers of this infrastructure through co-construction and participation [25, 28]. This has been observed in US LTER[3], where change occurs through '...a continuing mix of activities, reviews, workshops and meetings involving participation from a diversity of members from each site' [25, p. 23].

However, even though the development of infrastructure has been identified as an activity conducted by individual users of this infrastructure, academic career rewards are often dependent on individualistic accomplishments (such as publications) that do not necessary contribute to developing infrastructure (such as designing information systems) [7]. Incentives to participate in infrastructure development therefore need to come from a communal or individualistic attitude that contributing to bettering infrastructure will positively influence individuals and their community in the long run. Additionally, despite the vital role of users in infrastructure development, researchers such as domain experts are often seen as mere 'recipients' of its services when it comes to the development of the data practices and systems that are part of this infrastructure [10]. Thus, they are often not involved in design processes. Additionally, this restriction to design participants runs the risk of creating 'constructed' infrastructures that fail to meet users' needs as they fail to adapt to existing and changing environments [7, 27]. Finally, it is important to point out that elaborate design efforts have been recognized as problematic in regards to emergent change, as the planned nature of it restrains and interferes with the accreting and emergent qualities of infrastructure [27]. This restrainment is especially unfortunate, as one misses out on a reflexive evolution of infrastructure where '...researchers, developers, and policy makers gradually learn to learn together...' [9].

[3] Long Term Ecological Research Network comprising sites in the United States, Puerto Rico, and Antarctica [33]

A note on the following sections
Some of the issues discussed in the following sections are either studied in the conditions of infrastructure, or with the purpose of enlightening an area of research that is especially relevant in the light of research and/or information infrastructure plans. Thus, the discussion that follows addresses many issues that are intertwined with infrastructures and the challenges and prospects they present, without explicitly stating so. This is done to modularize the concepts discussed in this chapter.

2.2 Data curation

Data curation is concerned with the continuous *care* of data throughout it's lifecycle, and '...enable[s] data discovery and retrieval, maintain[s] quality, add[s] value, and provide[s] for re-use over time' [34]. In this thesis, curation is viewed as a four-step process of *retrieval, preparation (including cleaning*[4] *the data), sharing,* and *maintenance*. Insight into data curation issues and challenges is limited, however needed to achieve effective and sustainable data curation practices and procedures [14]. Note that the main objective of my thesis is concerned with preparation of data, which is only part of data curation. As there is more relevant literature on curation, and data preparation is shown to be a tedious and vital part of the curation process [36], it was decided to focus on data curation literature.

Useful and reusable data requires a resource-heavy process of curation[13, 15, 18, 26, 37]. In section 2.3, a misalignment between funding and data requirements will be addressed. This misalignment is a hindrance for data sharing and reusability of this data, and is largely concentrated in the curation practices that are necessary for serviceable[5] data sharing. The challenges it is comprised of have been documented in several academic works [13–15, 18], and one major issue is that curation work is often not acknowledged by funding bodies. Consequently, the responsibility of curation is left to individual researchers with little to no recognition [18]. Without acknowledgement, the challenge of acquiring sufficient funding and other resources will remain undiminished. This notion is supported

[4] Data cleaning is the repeated process of '...detecting and removing errors and inconsistencies from data in order to improve the quality of data' [35, p. 3].

[5] I choose to use the word *serviceable* to emphasize that the goal of data sharing is to fulfill its function and intent, which is to enable data reuse and transparency, and promote research.

by surveys on data curation in FinLTSER[6] in 2007/8 and a decade later, showing that issues with funding and resources prevent researchers from doing the work they aspire to do [14]. Evidently, there exists a gap in communicating this need to policy-makers, and this gap arguably needs to be explored and articulated in further research.

The issue of data curation and acknowledgement in research policy might also be a slow reaction to change, as the EU addressed the lack of reward systems for data curation in research networks in the 2018 FAIR action plan [38], and urged large research facilities to include FAIR data as a criterion for funding. The idea that slow reaction is to blame was also supported in EU's 2016 report and recommendations on open science, with the report claiming '... scholarly communication, data management methodologies, reward systems and training curricula do not adapt quickly enough *if at all* [emphasis added] to this revolution' [39, p. 5]. Case studies, workshop reflections, and other literature on data curation in environmental research seems to show that this is mainly a structural and political issue, with researchers constantly working to make the best of a situation they often experience as constricting and out of their control [13, 14, 18]. This is not to say that EU's claim of slow reaction is untrue, but is rather a suggestion that top-down influences such as policy have a major responsibility in regards to the facilitation and evolution of data sharing, curation, and the environment surrounding it.

As already mentioned, EU's action plan of 'turning FAIR into reality' [38], urged large research facilities to make FAIR data a criterion for research projects. What seems to be the case, is that FAIR data has been included into new funding criteria, yet funding and resources for the work going into accomplishing FAIR data has been largely overlooked[18]. Interestingly, in the case of EU's FAIR action plan [38], what seems to be the identified problem for accomplishing FAIR data is the lack of incentives for curating data, and not the tremendous amount of effort that goes into this curation. This might be one of the reasons why policies are not yet adjusted to the reality of data curation needs.

A big part of data curation is filtering *relevant data* [18], implying that data curation practices need custom adjustments. A need for custom adjustments to data in the form of adjusting, analyzing and reanalyzing data *by different people, instru-*

[6] Finnish Long-Term Socio-Ecological Research Network is a network for ecological and environmental research '...with longitudinal empirical engagements' in Finland [14, p. 4].

ments and even computers is what Edwards and colleagues refer to as *data friction*, which '...describes what happens at the interfaces between data 'surfaces': the points where data move between people, substrates, organizations, or machines – from one lab to another, from one discipline to another, from a sensor to a computer' [40, p. 669] et cetera, often creating '...conflicts, disagreements, and inexact, unruly processes' [40, p. 669]. On a more local scale, one researcher might invoke data friction when trying to understand or identify anomalies in data, by involving colleagues and other data. In cases where researchers deal with a high degree of systematic uncertainty[7] in data, this friction can take on the form of '...supporting and triangulating[8] one kind of data by connecting and hence grounding it relative to other supporting data' [43, p. 1729].

As data curation practices need custom adjustments, complete standardization as a way to success is likely not an option [25, 44]. Furthermore, what 'good data' means is not a question that researchers themselves have a definite answer to [26], which validates the notion that data quality is to a certain degree subjective [45]. The term quality is defined in ISO 9000 as '...the degree to which a set of inherent characteristics of an object fulfils requirements' [46], meaning that quality is context dependent. As suppliers of data for reuse may have a different set of requirements for quality data from the potential users of this data, it further complicates the notion of quality data in the context of data sharing, as what is deemed as quality can vary depending on local research context and perceived importance.

When examining quality data, one arguably also needs to take a look at what components need to be in place in order for this quality to be achieved. Two terms often used when talking about systematic approaches to data quality are *quality assurance (QA)* and *quality control (QC)*. QA is often used to describe the planned and systematic steps taken to provide confidence in data fulfilling quality requirements [17]. Sometimes, it is more specifically defined as the '...activities undertaken *prior to data collection* [emphasis added] to ensure that the data are of the highest possible quality at the time they are collected' [16, p. 72], including development of data management systems, training and certification of data workers, testing of data collection procedures, and obtaining relevant information

[7] Uncertainty due to possible errors that are not determined by chance but rather introduced by inaccuracy (such as of an observation or measurement, as a result of limited data, or as a result of a lack of knowledge)[41].

[8] 'Triangulation refers to the use of multiple methods or data sources in qualitative research to develop a comprehensive understanding of phenomena' [42, p. 545]

about the sampling site(s) where research is to be conducted, among others [16, 17]. In other words, QA is about *preventative measures*.

Quality control on the other hand, involves process monitoring with the goal being '...to identify and correct sources of either bias or excessive noise in the data both during and after data collection' [16, p. 76]. Thus, QC is concerned with *defect identification*. A lot of QA and QC efforts have historically been concentrated on laboratory analysis [16, 17, 47, 48], however, considering other parts of the data life-cycle is equally important to assure quality data. These parts include sampling, transport, analysis, and storage among others[17, p. 51]. Consequently, this often involves multiple data workers with different skill sets, and the need to see data in the context of its environment when doing both QA and QC [17, pp. 56–57].

EU's focus on a change in reward systems in their FAIR action plan [38] tells us that there is an apparent need to change not just policy, but also research culture and what work is valued in the research community. What this focus also tells us, is that data curation is a multifaceted issue, that needs to be addressed at local, infrastructural, and policy levels.

2.3 (Open) data sharing

The era of open data sharing has brought along a new ethos of openness and collaboration in research, and with it new challenges [49]. Open data in research can be defined as data that is '...accessible to relevant users, on equal terms, and at the lowest possible cost' [50], and is for many environmental research projects a funding criterion set by the EU and enforced by agencies such as the research council of Norway (Forskningsrådet)[51]. Many researchers now need to change their practices of data collection, naming and archiving to fit potential new users of this data, and questions of when data should be shared and how to inform both suppliers and users of open data standards and consequences need to be answered [49]. *Open* data has arguably become a popular term. Thus, most recent literature on data sharing focuses specifically on *open* data, meaning data that is available through openly accessible portals for researchers to use and reuse. However, open data sharing and data sharing in more controlled settings (such as between selected institutes) often deals with the same concerns, albeit sometimes in varying degrees. Thus, it was seen fitting to talk about both from the same perspective and in the same section. When concerns and references are especially relevant to only *open data*, this will be specified. Otherwise, the terms open data sharing and data sharing in controlled settings will both be simply addressed as ’data sharing’.

To gain understanding of ’the good and the bad’ of data sharing, this section will look at three perspectives. Firstly, policies on data sharing will be briefly discussed. Second, challenges of sharing data will be examined. Lastly, challenges reusing data will be looked into.

2.3.1 Policy on data sharing

Making data (openly) available for other researchers is a funding criterion for most research projects in the EU [26, 52], and the increased prevalence of large interdisciplinary projects means there is a need for facilitation for data sharing. The criterion is further specified in policy documents from funding bodies such as the European Strategy Forum on Research Infrastructures (ESFRI) and the research council of Norway (Forskningsrådet) [51, 53]. Especially environmental research is a ’...special target of data sharing efforts’ [54, p. 9] and open data has been especially encouraged for environmental problem solving [13].

One central addition to research policy in recent years has been the FAIR principles, which are guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets [55]. Following FAIR principles in research has been part of EU’s guidelines since 2016, when the action plan ’Turning FAIR into reality’ was published [38]. The principles are meant to improve both usability and reusability of data, and have descriptions of standards and general procedures researchers should follow. The goal is to make data not simply reusable, but also minimize the amount of heavy manual processing needed for this reuse to be realized in practice, by making sure that data is structured such that it is suitable for automated processing.

Despite the introduction of policies that promote open data sharing, there seems to be a misalignment between funding and the requirements for FAIR data in environmental research [18]. This misalignment existed before a lot of the policies and incentives present today [13, 56], and seeing how the EU has addressed the need to decrease the existing discrepancies [39], we are seemingly moving in the right direction. However, as discussed in section 2.2, a lot of the issues come down to neglected curation practices that are not acknowledged in EU’s research policy.

Challenges of data sharing

Next, challenges of data sharing and its reuse will be examined from two perspectives: the perspective of researchers supplying data, and the perspective of researchers on the receiving end of this data.

2.3.2 Supplying data

Researchers that retrieve, prepare, and share data can be viewed as *suppliers* of data. As discussed in section 2.2, the total process of this retrieval, preparation, and sharing, in addition to maintaining this data, is referred to as curation. Arguably the biggest challenge in regards to supplying reusable data in environmental research is the necessary curation practices that take this data from retrieval to finalization. As this has already been discussed in section 2.2, we can turn our focus towards the other data sharing hurdles identified in literature: lack of acknowledgement, lack of incentives, fear of (mis)interpretation and technical challenges. These hurdles will be discussed below.

Lack of acknowledgement

One concern that has been addressed in several academic works is that of handling intellectual property rights in data sharing [18, 37, 57]. Researchers put a tremendous amount of effort into generating, cleaning, and sharing data, yet they feel unsure of if, or to what degree, this effort is acknowledged by receiving parties [18]. Sometimes people do not want to share what they have worked at so hard as it feels unfair. Yet, more often than not it comes down to the security felt in social relationships and communities, and the absence of this security when faced with the prospect of sharing data openly; researchers do not mind sharing data with people they know and trust, but sharing their hard work with anonymous researchers and external parties feels like a risk they would rather not take [57].

Lack of incentives

All eLTER(ESFRI) research projects, as well as a growing number of other environmental research projects, are required to create and submit a management plan, and follow FAIR principles to enable data sharing [38, 50]. However, quality assurance and control[9] of the data in the management plan, and to what degree FAIR principles are followed, is not included in policy (and, arguably, achieving any increase in overall quality is nearly impossible, with one reason being that

[9] See section 2.2 for a further explanation and discussion of quality assurance and control.

FAIR data isn't necessarily an objectively measurable quality). Additionally, the curation practices that realize quality data are often done under the radar [18]. It is thus up to researchers and the community they are part of to create incentives for providing quality data for data sharing and reuse.

Traditionally, reward systems in research are centered on narrow metrics such as publications and research metrics[7], and do not take into consideration making research data and metadata reusable or available [37–39]. Research culture is described as one that '...values creative and independent research above secondary use of data' [37, p. 636], and despite an ever-growing reliance on data experts, they are greatly undervalued in academia [39]. The current reward systems have been addressed as a challenge for data sharing, and are by the EU considered the foremost obstacle in achieving FAIR data in research [38]. Researchers are seen as 'risking their careers' [38] by prioritizing curation practices for achieving FAIR data, and their efforts are seldom recognized. Additionally, data work that is necessary for data reuse is often not benefiting individual scientists [39, 56], making incentives less explicit. Generally, environmental researchers recognize the importance of data sharing, however, when prioritizing a never-ending queue of tasks, incentives in the form of resources and rewards are missing for data sharing to be at the top of the agenda [13, 14].

Furthermore, scientists find it more beneficial and *safe* to share data as part of a social exchange [37]. This sharing is often more of an ongoing exchange, referred by Wallis and colleagues as the 'gift culture of scholarship' [58]. When sharing data with colleagues, the incentive is that one can expect something in return, either directly in the form of data, or indirectly by solidifying social relationships. A cross-disciplinary study from 2014 [59] showed that regulative[10] pressures (such as by funding agencies) had no positive influence on data sharing behaviours (the authors noted that this might be a case of researchers not seeing data sharing obligations as 'pressures', especially since the sample group did not experience any explicit enforcement), while normative[11] and cultural-cognitive[12] pressures had a positive impact, supporting the idea that the socio-cultural environment of researchers is a deciding factor in data sharing efforts.

Additionally, one can expect a certain level of *trust* between colleagues in contrast

[10] Defined as 'the rules that an authoritative organization or actor sets for desirable behaviors of other organizations or its organizational members' [59].

[11] Defined as the 'social obligation caused by collective expectations in a community' [59].

[12] One where 'individuals observe others' activities and simply imitate their behaviors' [59].

to sharing data openly, where no trust relationship is formed. Consequently, the perceived risk is greater than the benefit. This is supported in further literature that suggests that trust between researchers, as well as their respective work organizations, is vital to reduce perceived risk of collaboration and shift the focus towards existing incentives [60].

Fear of (mis)interpretation

A cross-disciplinary international survey in 2011 [61] showed that the majority of participating researchers believed data could be misinterpreted due to complexity, and that poor quality of data may also be a reason for this misinterpretation[13]. Baker and Millerand [25] support the notion that one of the reasons for resistance to data sharing is the justifiable worry that data can be misinterpreted if shared. This stems from the '...scientifically salient concerns about the lack of maturity of data classification efforts...' [25, p. 25], as well as the fact that complex data is hard to interpret when you are not familiar with the environment of creation or lack context for this data [37, 40]. This will be discussed more in depth in the next section, section 2.3.3. What adds to this uncertainty is that researchers working on data do not have a clear answer to what 'good' data is themselves [18], and thus, the implications of this uncertainty in the context of data sharing are hard to establish [26]. One possible implication is that data is at a higher risk of misinterpretation.

Conversely, researchers are also worried about others *interpreting their data in new ways* [57]. This fear comes down to the social relationships researchers cultivate in their work [37]; if one shares one's own data with a trusted colleague, new interpretations of this data can be part of a social exchange, and notably, this oftentimes happens *before* one already has published findings based on this data. Sharing one's own data with the *'world'* however, creates the risk of reputation loss if this data leads to interesting discoveries that the original creator has overlooked [57].

Technical challenges

A hindrance to data sharing that comes up in several literary works is technical challenges [25, 26, 37]. Mosconi et al. [26] point out a significant challenge of data sharing: a unique research project will have unique needs, use unique tools, and produce metadata specifically attuned to the environment it is created in.

[13] 75% and 71% of participants believed that data misinterpretation stemmed from complexity of data and poor quality of data respectively. A total of 1329 scientists participated in the survey.

Designing a common database that works with such diverging needs is therefore incredibly difficult. Additionally, there is a need to create tools supporting 'data negotiation'[14][25, 26], giving data suppliers the control and transparency that is present in selective data sharing between colleagues, but absent in open data sharing today [18, 57].

If one manages to develop an information infrastructure that minimizes the technical challenges described above, it is still important to note that technology itself will not solve some of the important problems already discussed. Ethical challenges, data quality uncertainties and many other issues will still be of concern[25] - and most importantly, as technology, infrastructures, and communities evolve, one will need to find new solutions.

2.3.3 Reusing data

In regards to using data provided by external research projects, two issues will be discussed below: local context and technical challenges. Note that reuse can both be the act of using external data for what Wallis and colleagues [58] categorize as background purposes (such as context or calibration) and for foreground purposes (meaning it is associated with the main research question).

Local context
Context can be seen as relevant knowledge about data or metadata. Metadata is often considered crucial for providing context [25, 37], however Birnholtz and Bietz argue that it is often not enough *(the study conducted was interdisciplinary, but notably did not include environmental research)* [57]. When assessing the reusability of data, knowledge of local context is crucial for environmental researchers [26, 37]. For instance, '...researchers with direct access to all the original material and data from a study may struggle to understand it' [26, p. 757], which suggests that context is needed to gain an adequate understanding of data. However, gathering this context has shown to be a challenge, and is both deprioritized and resource demanding [26].

If one investigates research sites, one will arguably find one or more communities of practice [54]. Wenger [62, p. 4] describes this phenomenon: 'People belong to

[14] Data negotiation in this context means '...communication between data producers, data consumers and, potentially, data re-users...' [26, p. 781] to enable researchers that create data to choose what is shared and with whom.

communities of practice at the same time as they belong to other organizational structures. In their business units, they shape the organization. In their teams, they take care of projects. In their networks, they form relationships. And in their communities of practice, they develop the knowledge that lets them do these other tasks'. As the communities of practice found at research sites often do not include both the members supplying data and the members reusing this data, one will see an apparent problem: knowledge, and especially *tacit knowledge*[15], is gained through becoming part of the community [64] in which this data is sourced. Furthermore, learning, and thereby knowledge, cannot simply be separated from practice [26, 64]. Thus, data sharing and reuse faces a complex issue of on one side documenting context in a way that can be understood from the outside of the community of creation, which has proved to be challenging [54], and on the other side understanding this documentation without being a community member[25].

An adjacent view on this issue is that context can be gained through one's own field and/or laboratory knowledge [37], however this implies that the data one receives has been created and curated in a similar environment to one's own. This is supported by Birnholtz and Bietz's [57] notion that one needs to understand '... 1) the nature of the data itself, 2) the scientific purpose of its collection, and 3) its social function in the community that created it' [57, p. 346] in order to gain access to contextual information. The third criterion is especially hard to achieve without having insight into the relevant community of creation, or an existing social relationship with the researchers providing the data for reuse [57]. Furthermore, Karasti, Baker and Halkola [13] observed that despite personal insight and connections, ecologists need to work hard to comprehend the data they seek to reuse, oftentimes discarding it altogether to seek other sources as uncertainties about data are common. This suggests that the three criteria set by Birnholtz and Bietz all need to be present for data reuse to be successful.

Technical challenges

As has been discussed in section 2.3.2, divergent needs of different research sites means that designing common databases for data sharing is challenging. If one uses different information systems, potential users of data face other challenges such as incompatible hardware, software, and data structures [37]. However, even with shared information infrastructures, '...data can rapidly become unreadable because of software and hardware updates...' [26, p. 757]. Additionally, data structure and representation depends on the research approach used [26], meaning

[15] Knowledge gained through personal experience rather than theory [63].

that it is hard to get context as a result of technical restrictions.

2.4 Coordination and cooperative work

The process of preparing data and making it available for reuse usually requires multiple researchers. As the work of each influencing party at the very least influences data quality, if not defines it, one can expect that there exists a dependency between researchers which defines the success of data sharing. As Schmidt [65, p. 36] put it: 'An omniscient and omnipotent agent to match the multifarious environment of modern work does not exist'. Thus, cooperation is both inevitable and crucial to understand. To gain knowledge on cooperation in scientific work, we will take a closer look at cooperation and its coordination in this section, as well as how *trust* plays a role in cooperative work. Note that literature on these topics in the context of environmental research is scarce, and that I struggled with finding articles that address them. Thus, I largely draw on studies from other areas which I believe are relevant, and where I believe the observations are transferable to the setting of environmental research and data sharing.

2.4.1 Cooperation and its coordination

The term *cooperative work (CW)* of CSCW has historically brought along confusion [22, 66], and is understood only in a '...deficient, vague, [and] patchy' way [66, p. 284]. Thus, the definitions are often varied. However, this review will use the term as described by Schmidt and colleagues [22, 65, 67], which is further elaborated below.

Schmidt and Simone [67] describe: 'Cooperative work is constituted by the interdependence of multiple actors who, in their individual activities, in changing the state of their individual field of work, also change the state of the field of work of others and who thus interact through changing the state of a common field of work'. In the process of establishing the characteristics of cooperative work, Schmidt and Bannon [22] make an important clarification: 'The notion of mutual dependence in work does not refer to the interdependence that arises by simply having to share the same resource'. This distinguishes cooperation from coordination, which does not imply direct mutual dependence. Furthermore, coordination implies a *desire* to not have others' work affect one's own [22], as it is strictly not necessary. Cooperating parties on the other hand, are directly dependent on each others' work, and are thus dependent on cooperative involvement to reach a common goal [22, 65]. All things considered, our current understanding of cooperative work and its coordination is insufficient [66, p. 284], so more research is needed on these aspects in real life settings.

When looking at the coordination aspect of data sharing facilitation, it is imperative to take a look at what this coordination implies. As discussed in section 2.2, data needs continuous care throughout its life-cycle in order to ensure data quality. As this care is more often than not provided by multiple researchers, there needs to be a level of cooperation [15, 67]. This is not to say that all involved parties at every step of the data curation process will need to directly cooperate; in fact, '...since it involves multiple actors, cooperative work is inherently distributed, not only in the usual sense that activities are distributed in time and space, but also - and more importantly - in the sense that actors are *semiautonomous* [emphasis added] in terms of the different circumstances they are faced with in their work as well as in terms of their strategies, heuristics, perspectives, goals, motives, etc' [67, p. 4]. Thus, we can view data curation as a pipeline through which data is processed by different actors, which is variably dependent on other steps in the curation process, and in which it is expected to find multiple cooperative 'formations', sometimes overlapping, in a wider coordinated network. This coordination and management of distributed cooperative work is what often is referred to as *articulation work* [67].

If we continue this pipeline analogy for data curation, we can generally expect that not every individual dealing with data from its conception to its sharing will intentionally always work on producing data fit for reuse. However, as Tenenberg, Roth and Socha [68] expand on, the activities these individuals carry out have inherent shared intentionality in that it has a shared goal that is worked towards in a network of coordinated cooperative interactions. These cooperative interactions can be described as what Schmidt [65] calls *cooperative work arrangements*, in that they are '...shifting patterns of actually enacted relationships...' [65, p. 10]. Furthermore, Schmidt proposes that these cooperative work arrangements are performed in and across *work organization*, which in the case of research sites could be defined as the configuration of all the individuals that are assigned to a specific task.

Furthermore, coordinative work is dependent on expertise, where this expertise can be divided into two categories consisting of *process expertise* and *domain ex-*

pertise[69]. The former is determined by having '...situated knowledge of the social processes operating in a specific social context' which '...complement and amplify the work of domain experts' [69, p. 62] through communicative exchanges, while the latter concerns itself with the knowledge one gains from being an expert in a specific field. A study by Barley, Treem, and Leonardi [69] found that in high-stress environments, coordinative work is predominantly based on process expertise. However, as coordination of data workers in environmental research arguably is not a high-stress environment, studies have yet to show what expertise is needed to perform such tasks.

The role of knowledge exchange in cooperation

One important aspect of cooperative work arrangements is that if one is part of this arrangement, *sharing information* in the form of knowledge exchange or data becomes an integral part of the cooperation, and this is critical for the '...construction of knowledge and understanding, which is ultimately a social, not individual, task' [70, p. 341]. Furthermore, much of this knowledge exchange occurs in an informal manner, making informal communication the cornerstone of cooperation in tasks that 'on paper' are individualistic in nature [71–73].

With the rise of collaboration and communication technologies (such as Microsoft Teams), there exist studies that show co-location might not be the golden solution to knowledge exchange and the like [74, 75]. However, most literature [71, 76, 77] suggests that '...co-located work colleagues have opportunities for rich interactions simply because they can talk, listen, and watch each other' [71, p. 320], suggesting that co-location and casual, yet important communication, go hand in hand. One can however argue that it is highly dependent on context and how this knowledge exchange has materialized in the workers' environment, and as these studies are done in different fields and environments, literature does not provide answers to how co-location influences knowledge exchange between data workers in an environmental researching setting[16].

Challenges of interdisciplinary cooperation

Computer and data specialists are increasingly becoming key actors in research, and are vital in the realization of data sharing efforts[39]. It is therefore interesting to look at the cooperation between domain experts and technical experts that have little to no prior domain knowledge. Unfortunately, the cultural differences

[16] Or at the very least, I have failed to find such literature

resulting from a lack of co-evolution and differing backgrounds (with different reward systems and incentives) has in many cases made this cooperation challenging [39, 78]. Additionally, the lack of technology and infrastructure support has been brought up as a reason for these challenges [78]. This is not to say that interdisciplinary collaboration isn't working at all; in fact it has shown to be promising in solving complex issues in research on multiple occasions [78]. More importantly, it is unavoidable in modern research, and thus requires more recognition and attention [39].

In the process of examining cooperation between domain experts and technical experts, Mao et al. [78] use two central terms: *content common ground* and *process common ground*. The former '...depends on an abundant shared understanding of the subject and focus of work (know that)' [78, p. 6], while the latter '...depends on a shared understanding as well as a continual updating of the rules, procedures, timing and manner by which the interaction will be conducted (know how)' [78, p. 6]. The study showed that content common ground was rapidly evolving, and thus needed constant adaptation from both domain experts and technical experts. It is however unclear how well these findings would transfer to the case of facilitating data sharing, as the rapid evolution was due to participants' goal of advancing scientific discovery, and curation practices are arguably comparatively stable and slowly-evolving in comparison. It is however interesting to note that cooperation required that participants accumulated process common ground, meaning that they constantly and progressively learned from their cooperative work arrangements.

2.4.2 The role of trust in cooperative work

Trust between data workers in research, and specifically between colleagues, is an understudied area. Thus, this section will draw from other fields to establish an overview of interpersonal trust.

When examining cooperative work, one can arguably expect a level of trust as a foundation for data work and data sharing. As Shapin puts it, '...trust is a condition for having the body of knowledge currently called science' [79, p. 402], which arguably, can be used both in relation to trust between colleagues, trust in and between research communities, and the trust put into the cumulative work of science.

Butler and Cantrell [80] propose five determinants to measure trust, which include integrity (honesty and truthfulness), competence (technical and interpersonal knowledge and skills), consistency (reliability, predictability, and good judgement in handling situations), loyalty (benevolent motives, willingness to protect and save face for a colleague), and openness (mental accessibility and willingness to share ideas and information freely). The study in which the concepts were proposed was on trust relationships between subordinates and superiors, but can arguably be used to measure and describe trust between peers as well, as done by Schindler and Thomas [81]. Both Butler and Cantrell, as well as Schindler and Thomas, found that integrity and competence was ranked as the most important dimensions of trust, while loyalty and openness was ranked at the bottom[80, 81], independently of participants being subordinates, superiors, or peers. Schindler and Thomas argue that finding no difference in trust based on relative status of individuals may be due to the nature of the sample participants' lack of rigid hierarchy in the work place, as other studies do show a difference in trust between different roles, depending on where these roles are in the organizational hierarchy.

Arguably, the focus on quality assurance and control in environmental research, as well as a big prevalence of literature on systematic error detection and reduction (such as: [82–84] and many more), implies that environmental data is quite prone to systematic errors, meaning that trust in the data itself is also a rational concern. Mikalsen and Monteiro [43] found that data workers in environments where data is characterized by systematic uncertainty, *do not at all trust the data at face value*, however their work provides them with knowledge that cultivates beliefs about which data is 'safer', and which data should be approached with more caution. Notably, this was data directly coming from instruments and/or sensors. If we examine trust in cases where one has not been involved with curating data from its creation, literature suggests that the trust conditions differ, especially if one does not have a background in data science. A study in corporate settings showed that the trust individuals place on data often is a direct continuity of the trust they place on the people providing this data, resulting in trust sometimes being placed '...not in the analysis, but in the identity of the analyst' [85, p. 22].

Furthermore, researchers in recent years often have adapted a pragmatic approach to accuracy and elimination of errors, as the data quantities produced by sensors and other mass-collection devices mean that ensuring complete correctness is unattainable [86]. Consequently, trust can only depend on the 'good enough' practicality of results, rather than a standard of perfect data [85]. As environmental research often concerns itself with complex modelling of the real world [83, 87], attaining flawless data becomes even further away from reality. As Box proclaimed, 'all models are wrong, but some are useful' [88, p. 203], suggesting that models and data being 'right' is not the quality one should be most concerned about. This further suggest that data trust needs to be achieved through other means than through the measure of how accurate the data in question is. This is not to suggest that the process of data curation should discard all data accuracy concerns, but rather that there needs to be a balance that creates an '...intimate relationship between trust, skepticism, and action' [85, p. 21].

Finally, Passi and Jackson's findings [85] show that trust in data is built on communication, in the sense that diverse experts '... translate between different forms of knowledge' [85, p. 21] through narration. In a way, this form of communication takes on the quality of *communities of practice* (see subsection 2.3.3 for definition). In these communities of practice, learning and solving issues is done with the vital cumulative knowledge built through narration, in stories shared by community members [62, 64]. This narrativization of enormous amounts of often complex information builds a bridge between reality and data, creating a sense of intuitive understanding. By filling in gaps and giving meaning to data, narrativization enables researchers to trust it.

Chapter 3

Case description

This research project draws on a literature review [1] and an explanatory case study conducted in the fall of 2021 and spring of 2022 respectively. The unit of analysis for this project is to investigate data preparation and sharing practices in Norwegian environmental research infrastructure, and how these practices facilitate quality data.

Long-term Ecosystem Research in Europe, commonly referred to as eLTER, is an umbrella network established in 2003[89]. Since then, two of the four main objectives of eLTER have been (1) to develop criteria for eLTER sites (these being research sites that are part of the eLTER network) and (2) to improve cooperation between different stakeholders [90]. eLTER comprises environmental research sites in several European countries, with some of these sites being situated in Norway. A number of these are Norwegian research institutes with associated stations, and the work of retrieving and preparing data is usually shared between the stations and the institutes. eLTER sites are highly heterogeneous in that the research objectives, policies, pressures and resources vary substantially from site to site.

In this thesis, participants were recruited from three Norwegian institutes that specialized in different fields of environmental long-term monitoring:

The first group of participants did long-term monitoring of air pollutants where data was collected through several monitoring sites and stations across Norway, usually as a combination of sensor data and samples that were sent to the institute for analysis. The stations vary in size, and while some are staffed with multiple researchers, others have automated data transmission to the institute databases.

One such monitoring station was on the institute premises. A picture of this station can be found in Figure 3.1. The institute is also a chemical coordinating center, meaning it gets air pollutant samples from different countries to evaluate air pollution on a global level, as well as being in charge of other supra-national coordinative actives.

The second group of participants worked with agricultural runoff and its effects on water pollution. Here, most of the data was collected through automated sensors at unmanned stations, where field technicians would collect samples for laboratory analysis once a week while ensuring that the stations did not suffer from technical or practical errors.

The third group of participants worked at a bigger institute with several locations and ecological research objectives, with participating researchers mainly working in camera trap projects for wildlife observations. This included a quite different process of collection, as cameras all around Norway would take pictures when sensing movement and heat, and these would go through several rounds of semi-automatic and manual sorting before analysis in the form of species identification would start. Additionally, the cameras needed external care by local 'operators', who would collect memory cards and assist in problem solving when technical problems were identified.

In 2015, eLTER started the design phase of what will become eLTER RI, a research infrastructure with the goal to '...comprise National Research Infrastructures (NRIs), and European level Central Services (CS), such as data access, training and harmonized methods and parameters' [12]. The idea is that all participating research projects will be part of a highly integrated infrastructure with a shared eLTER service portal providing access to data and sites. The different projects will also follow agreed policies, something that is not the case of eLTER sites today.

In 2018, eLTER RI was officially added to the ESFRI roadmap[53]. ESFRI has a central role in the creation and implementation of policies that govern European RI's[91]. These policies include those concerning data sharing, and thus significantly influence researchers' work with data. With eLTER comprising numerous countries and research sites as of 2021, all with varying degrees of digital tools, resources and expertise, the status quo is a wide variety of issues pertaining to policies on data management, curation, sharing, and the use of this shared data.

Researchers working on environmental research projects are part of a complex multi-layered network, usually involving NRIs (such as Forskningsrådet in Norway) on a national level, as well as eLTER and ESFRI on a European level. The data work conducted by these researchers lays the foundation for environmental policy development and decision making, in addition to giving valuable insight to the public about environmental trajectories [92]. This is often done through long-term monitoring of environmental factors, such as air and water pollutants. Because of their influence, it becomes imperative that such long-term monitoring studies can ensure quality data and results. Furthermore, the institutes behind these environmental studies are often non-profit organizations, and thus rely on a constant stream of funding that can support the decade-long work.

As eLTER in Norway comprises several large research sites that are required to share data, it was natural to use eLTER member sites as an 'access point' for participant recruitment. Thus, some researchers interviewed in this project were employed at eLTER member institutes. However, as there are other environmental research bodies in Norway that are required to share data, some of the interviewees were not affiliated with eLTER.

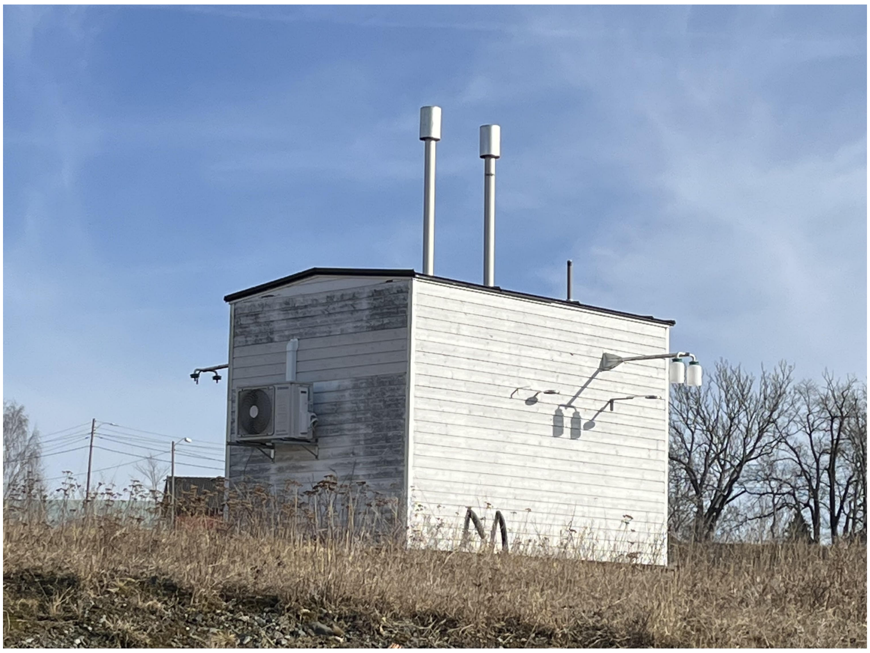


Figure 3.1: An air monitoring station at one of the visited sites. The station sends live data to the institute database. Additionally, samples for analysis are regularly collected at the station. *Picture is taken with permission from the participant showing us around during an observation visit.*

Chapter 4

Research Method

In this chapter, the research methods used in the study will be presented. First, the research strategy for this study will be presented, followed by a description of the methods of data collection, and the motivations behind them. Next, the process of participant recruitment and the preparations for this recruitment will be presented. Finally, the grounds on which the method for data analysis was chosen, as well as the steps in this data analysis and its subsequent concept development will be discussed. A resulting analytical framework can be found at the end of this chapter, in Table 4.2.

4.1 Research strategy and paradigm

The chosen research strategy for this project is the explanatory case study, with the deciding factors being my research questions (see section 1.2). The research questions aim to contribute with empirical insight into the key processes and relationships in environmental data work, and their influence on the preparation and sharing of data and its quality. Additionally, the goal of the research questions is to gain insight into the challenges of policy implementation, data sharing, and the relationship between them. It was thus natural to conduct a case study with interviews being the predominant data source, and use observations and document study as complementary sources to insight [93]. The focus on the *how and why* instead of arriving at a clear-cut answer, in addition to the importance of uncovering participants' tacit knowledge, further suggests that choosing a qualitative case study approach was the natural choice [93].

Furthermore, this research has been conducted with the belief that both my own, and the participants' views are subjective interpretations of reality. Thus, an interpretive approach was chosen for this case study [20, 94]. This was also motivated by the fact that the goal was to study the social context of researchers, and *make sense* of the data collected [95]. This meant using my initial theoretical understanding when planning and starting the data collection, which then continuously evolved throughout the interviews, data analysis, and even writing of this thesis [20, 94]. Thus, the sensemaking of the data collected was a continuous and iterative process. The chosen method of data collection, described and discussed in section 4.4, supported this iterative nature of the process.

As my research arguably covers a lot of ground for being a master thesis, the approach to interpreting my data was to actively employ the principle of the hermeneutic circle, by iteratively reflecting on the observed processes in context of each other, and in this way forming an overarching view of how the study objectives were interconnected [94].

4.2 Methods of data collection

The main methods of data collection were interviews and in-situ observations, with supplementary document analysis. An overview of the data sources can be found in Table 4.1. The importance of using multiple methods of data collection in qualitative case studies has been emphasized by Baxter and Jack [93], as this lets the researcher explore and reveal multiple sides of one phenomenon. A key benefit of using multiple methods of collection was that interviews afforded a great variety of perceptions and opinions, observations supplemented these perceptions and opinions with the knowledge of the environment of participants, and document analysis gave a high-level picture of policy interests and objectives.

The interviews were conducted either physically or virtually, depending on the nature of each participant's work and preference. Participants that could provide opportunities for observation were interviewed physically, while most of the others were interviewed virtually, especially since the ongoing Covid-19 pandemic meant that some participants preferred scheduling online interviews.

The interviews were semi-structured, to allow for a natural discovery of challenges and opinions about participants' work. This meant that many of the questions in the interview guide (see Appendix A) were not touched upon, but in return, new questions and answers were brought up. Furthermore, a casual conversational style, as well as unrushed introductions of the interview purpose and structure were made to make interviewees feel comfortable. This meant that less of the interview time was ’productive’, however, as Meyers and Newman[96] point out, these steps are important for interpretive interviews to minimize the risk of forced answers that do not reflect reality.

As seen in Table 4.1, a total of nine interviews were conducted, three of which were group interviews, consisting of two colleagues each. Additionally, four of the interviews were conducted on-site, three of which were paired with participant observations. This included presentations of data and the sharing practices/solutions (such as data sharing portals) of this data, tours of the laboratories/institutes, and external observations of (small) research stations. Additionally, one participant recommended visiting the National History Museum of Oslo, as a current exhibition showed the work and processes of data workers from the informant’s institute. The entirety of the observations provided context to the interviews, and in some cases, uncovered new knowledge as they prompted new questions along the way.

The document analysis in this project was primarily done in the fall semester of 2021. However, further policy documents were examined as the project progressed and new perspectives and hurdles emerged in the spring of 2022. The entirety of the document analysis provided insight into the status quo of data sharing policy, the background on eLTER and participating research sites, as well as relevant policy documents from the EU, ESFRI, eLTER, NFR, and NSD, in order to get a better understanding of the political constraints participants worked within and corroborate the data collected through the main methods of analysis [97].

Table 4.1: An overview of the data generation displays the data collection method (data source), the domain in which data has been collected, and the participants that have been interviewed.

Data source	Description of domain and participants
Semi-structured interview (duration: one hour per interview)	***Environmental monitoring*** • 3 environmental researchers • 3 work package managers • 2 project coordinators • 1 information and database manager • 1 software developer (tech support) ***Work package managers of eLTER RI (central services of eLTER RI*)** • 1 interviewee working with system development and requirement collection for eLTER RI • 1 interviewee involved with policy creation and information infrastructure development of eLTER RI
Observation	***Environmental monitoring*** • ½ day at 3 research institutes (1 and ½ day in total) • External observation + description of small research station • ½ day at the Natural History Museum of Oslo, exhibition: 'Climate research in the Arctic and Antarctica'
Document study	***EU*** • Strategy documents on Horizon Europe (funding criteria for research, open science policy, FAIR data in research, etc.) ***ESFRI*** • Roadmap and open science policy documents ***eLTER*** • Objectives, eLTER RI plans ***Research Council of Norway (Forskningsrådet)*** • Policy documents on open science, national RI roadmaps

4.3 Participant recruitment

Recruiting participants for the study was a four-step process, with the two first steps handling the ethical concerns related to participation, and the last two steps handling the actual participant recruitment. The steps are described below.

(1) Drafting an information letter, a consent form, and an interview guide. The information letter detailed what the project was about, what participation in interviews entailed, and the rights participants had with respect to their data, including the right to withdraw participation consent at any time during the study. The consent form detailed what data would be stored and for how long, and participants would need to sign this before interviews. Finally, interview guides for both the researchers working with data and eLTER work package managers were created. All files can be found in Appendix A.

(2) Submitting an application for the processing of participants' personal data to the Norwegian Centre for Research Data (NSD). The application included consent forms, interview guides, the collection method of oral and written data from interviews, and the steps to anonymization and protection of this data. The application and its acceptance letter can be found in Appendix B for further details.

(3) Deciding on relevant participants to contact. The participants in this study were chosen on the basis of relevance to the research questions, and were divided into two categories: environmental researchers, specifically those required to share their data with some other party, and informants who were directly tied to the eLTER council. The former category was composed of a wide range of data workers: field technicians, laboratory technicians, work package managers, and analysts. Additionally, project coordinators, software developers ('tech support') and database administrators belonged to this group. When creating a contact list for interviews concerning environmental researchers, individuals working at Norwegian research institutes were considered as relevant candidates. Furthermore, these institutes needed to be involved in data sharing with other parties, and thus it was natural to first look into researchers from eLTER member organizations, as eLTER is a strong advocate for the sharing of data. In the end, the individuals chosen as possible participants were a mix of previous candidates from my co-supervisor's contacts in the spring of 2021, as well as new candidates found through websites of eLTER member organizations and affiliated research institutes.

The second group of relevant candidates was identified on the basis of direct involvement with eLTER and the development of eLTER RI, as the goal was to study the influences on data sharing from the perspective of individuals placed between researchers and policy, to examine the processes and ensuing challenges they face. All members of the eLTER research council were contacted.

(4) Contacting participants. All relevant candidates were contacted by e-mail. In the invitation, the information letter, consent form and interview guide were attached. In a few instances, we would be redirected to other candidates that the contacted individual deemed more relevant.

4.4 Method for data analysis

When choosing a method to use for data analysis and conceptual model development, emphasis was put on a process that could tie specific statements and recurrent themes from the interviews together in an organized manner, ensuring a complete data overview when developing concepts from this data. Thus, a process that employs coding of transcripts was a natural choice. Additionally, an inductive[1] 'open' coding approach was decided upon as there was not much theory to draw from to create good predefined codes.The SDI (Stepwise Deductive Induction) approach described by Tjora [100] fit this criteria. Tjora proposes to stay as empirically close as possible to the raw data when creating codes, which makes it easier to remember what specific text a code refers to, as its uniqueness makes it a natural identifier when sorting through codes. Furthermore, the approach employs stepwise deductive feedback loops that lets the researcher review and reevaluate each step in the process, readjusting some of the data if necessary.

The process consists of six steps that go from raw empirical data to theory. The six steps are (1) Generate empirical data, (2) Process raw data (transcribe), (3) Code data (from transcripts), (4) Code grouping, (5) Develop conceptual categories

[1] Inductive coding is a process based on inductive reasoning where codes are not predefined, and rather emerge from examining raw data repeatedly [98, pp. 91–93], which means that one '...allows the theory to emerge from the data' [99], with the goal being to close the gap between speculated theory and reality. Deductive coding on the other hand, starts out with predefined codes, and is often used in well-studied areas with existing predefined concepts and theories.

and (6) Develop theory. These steps are taken from A. Tjora's book 'Qualitative Research Methods in Practice' (orig. title Kvalitative forskningsmetoder i praksis) [100], and were followed with a few adjustments. Step (5) Develop conceptual categories was renamed from 'Develop concepts', as the goal was to create categories corresponding to sets of overarching concerns and perceptions. Steps (3) Code data, (4) Code grouping and (5) Develop conceptual categories were done using the computer-assisted qualitative data analysis software NVivo. Finally, as the end goal of this project was to develop a conceptual model, step (6) has not been implemented. The steps and the adjustments made to them are explained below.

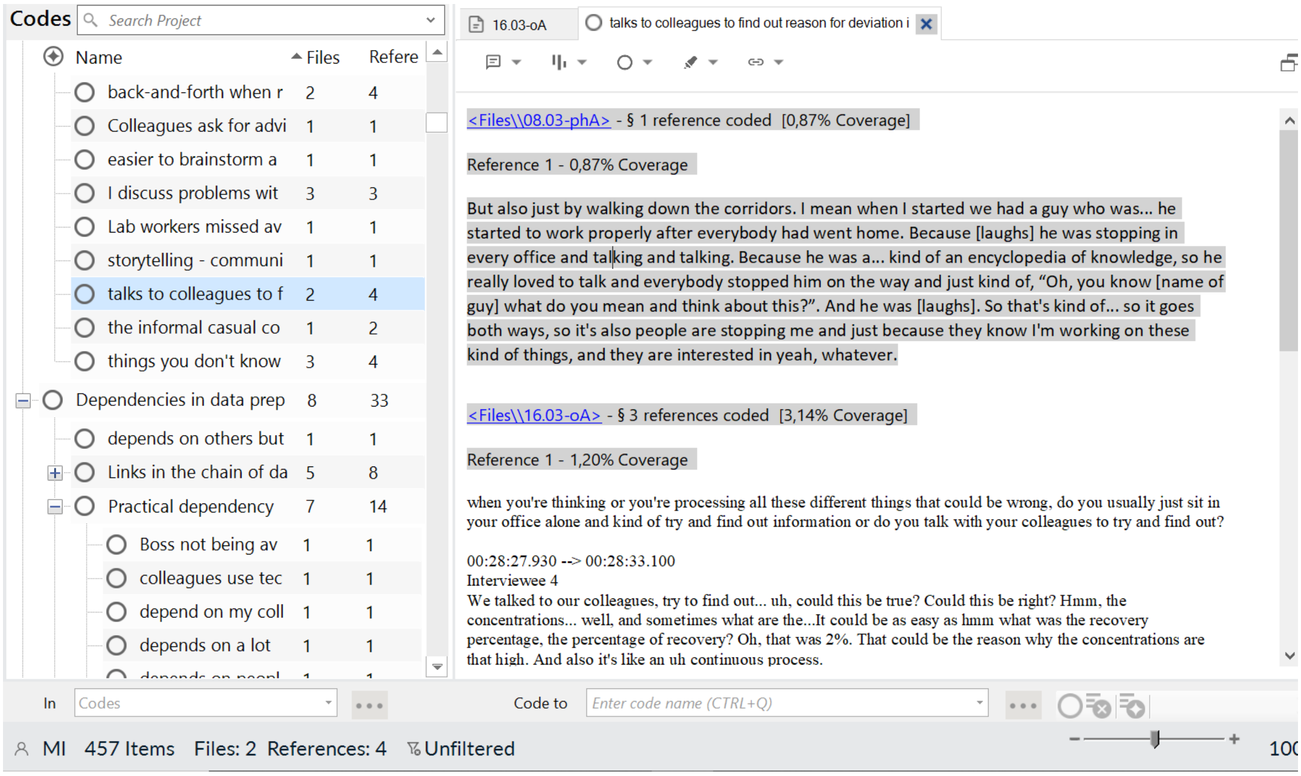
(1) Generate empirical data and (2) Process raw data.
Steps (1) and (2) can be looked at as a preparatory phase, and are thus grouped together for simplicity. As described in section 4.1, data generation was done predominantly in the form of interviews and observations, which is the basis of the empirical data used for this analysis. For transcriptions, a mixed approach of Microsoft Teams' live transcription service (which was later manually corrected and edited) and transcription by hand was used. Furthermore, field observation notes were refined and rewritten to be more complete.

(3) Code data (from transcripts and written field observation notes). The focus of this step was to identify any practices or concerns the participants had, and coding these in a way that would make them easily identifiable when later creating code groups. When coding the data, Tjora's approach of keeping the codes as empirically close to the source as possible was used in the majority of the coding work. This meant that more unique codes were created, and that these codes were rarely re-used. However, in some cases it was seen beneficial to generalize codes from the very start, and the codes created would diverge from the format recommended by Tjora. This was done on the basis that the uniqueness of the codes in question when following Tjora's approach was not seen as beneficial for the study. One such example is the code "Exchanges data with different institutes", which was preferred over the code "We and [External organization] often exchange data", as one could put all mentions of cross-institutional collaboration under one code from the start. In many ways, this generalization was a precursor to the next step of code grouping. The coding process can be seen in Figure 4.1, giving an overview of the NVivo interface. After finishing the step, a total of 496 codes were created.

(4) Code grouping. Since the majority of the codes created in step (3) were unique rather than generalized and categorical (and thus very few text segments were in any way connected at this point), it was especially important to create code groups that would reveal connections in the plethora of collected data and sort it in a thematic manner. This was done by grouping codes into recurring practices and/or concerns. Sometimes, a code group would be named after some code that represented the contents of the code group in a way that was easy to identify and remember. Some codes were left out of this step as they were not seen as relevant to the research as the coding and code grouping progressed. The code grouping differed from the SDI step described by Tjora in that some of the code groups created were strictly collections of process and participant background and description. This was done to effectively also collect contextual information while creating code groups for method development. Once the code groups had been collected, they were looked over, readjusted and reevaluated, as per the principle of ’feedback loops’ in the SDI approach.

(5) Develop conceptual categories. In Table 4.2, these are written in *italic* in order to separate them from conceptual categories. The biggest difference between code groups and conceptual categories in the SDI model is that conceptual categories are sufficiently abstract in relation to time, place, and people involved [100]. Once the conceptual categories had been collected, they were looked over, readjusted and reevaluated, as per the principle of ’feedback loops’ in the SDI approach. Once the conceptual categories were finalized, they were placed into four main themes, as seen in Table 4.2. This was done to create a better overview of the findings, and was not a step outlined in the original SDI.

Figure 4.1: A screenshot of the code grouping process in NVivo. The conceptual categories, code groups, and codes are placed on the left-hand side, in a nested structure. On the right-hand side, one can expand codes into their respective text excerpts. This is also where interview files appear when coding.



Name	Files	Refere
○ back-and-forth when r	2	4
○ Colleagues ask for advi	1	1
○ easier to brainstorm a	1	1
○ I discuss problems wit	3	3
○ Lab workers missed av	1	1
○ storytelling - communi	1	1
○ talks to colleagues to f	2	4
○ the informal casual co	1	2
○ things you don't know	3	4
○ Dependencies in data prep	8	33
○ depends on others but	1	1
○ Links in the chain of da	5	8
○ Practical dependency	7	14
○ Boss not being av	1	1
○ colleagues use tec	1	1
○ depend on my coll	1	1
○ depends on a lot	1	1
○ depends on peopl	1	1

<Files\\08.03-phA> - § 1 reference coded [0,87% Coverage]

Reference 1 - 0,87% Coverage

But also just by walking down the corridors. I mean when I started we had a guy who was... he started to work properly after everybody had went home. Because [laughs] he was stopping in every office and talking and talking. Because he was a... kind of an encyclopedia of knowledge, so he really loved to talk and everybody stopped him on the way and just kind of, "Oh, you know [name of guy] what do you mean and think about this?". And he was [laughs]. So that's kind of... so it goes both ways, so it's also people are stopping me and just because they know I'm working on these kind of things, and they are interested in yeah, whatever.

<Files\\16.03-oA> - § 3 references coded [3,14% Coverage]

Reference 1 - 1,20% Coverage

when you're thinking or you're processing all these different things that could be wrong, do you usually just sit in your office alone and kind of try and find out information or do you talk with your colleagues to try and find out?

00:28:27.930 --> 00:28:33.100
Interviewee 4
We talked to our colleagues, try to find out... uh, could this be true? Could this be right? Hmm, the concentrations... well, and sometimes what are the...It could be as easy as hmm what was the recovery percentage, the percentage of recovery? Oh, that was 2%. That could be the reason why the concentrations are that high. And also it's like an uh continuous process.

Table 4.2: Analytical framework. The table shows the overarching themes, conceptual categories, and code groups created in the analysis. Additionally, example excerpts from the interview transcripts are included.

Theme	**Conceptual category**	**Code group**	**Example of excerpt**
Coordinating data work	Coordinator types	Project coordinator responsibilities	’My role in [network of projects] is to coordinate most of it, and coordinate the people, coordinate the data flow.’
		Work package manager responsibilities	’I have to support them, and I have to know things, and I have to be the one that makes their job easy.’
	Coordinators being a necessity	Coordinators have an overall view of the project(s)	’I just make sure that everyone knows what they’re gonna do, and then ... if there are people working under them, they will sort of distribute tasks or make sure that things in their work packages are done.’
		Coordinators are communicators	’The researchers are not always able to give them [software developers] the most precise description of what [they] actually need ... And that’s the reason why I was hired, because I’m able to talk about both languages, [as] it’s quite different the way they talk.’
		Coordinators act as personal organizers	’I have like follow up meetings ... just to make sure that things are progressing and everything is fine.’

Preparing the data	Dependencies in data preparation work	Links in the chain of data flow	’I just create a table, and then send it to [colleague] who merges it together with the other data from the other sectors.’
		Trust dependencies	’I’m also dependent on my colleagues, that they do as good work as possible in the lab. That they are careful when they treat the samples [and] that they care about [and] try to avoid contamination and everything.’
		Practical (and verifiable) dependencies	’Or, sometimes we get a concentration from the analysis laboratory where we can see that it is completely outside of [the normal range]. [The data] can’t be right. And then we have to send it back for reanalysis’
	Participants’ experiences with quality work	Participants’ perceptions of quality work	’Quality control and assurance is the most important thing ... I think it’s really really important in order to be trustworthy.’
		Wide variance in standardization of processes	’And then you have to do quality control of the standards too. So we buy them with the certificate and we ... have to trace everything; the weight, and weight loss, and mixing things.’
	Knowledge exchange for realizing quality data	Intentional knowledge exchange	’We talk to our colleagues, try to find out, "Could this be true? Could this be right?" ... It’s like a continuous process.’

Preparing the data (contd.)	Knowledge exchange for realizing quality data	Spontaneous knowledge exchange	’Because we usually don’t have so much issues... well I have [laughs]. But then it’s not difficult to know who to contact, but it’s still the... kind of things you don’t know that you should know.’
Sharing the data	To share or not to share	Supplies open data through portal(s)	’So having open data has been important for us. But the EU expects open data.’
		Data sharing is done on a case-by-case basis	’We send data [to other people], but... the person who has been responsible [for env. monitoring program] hasn’t been quite willing to share. So we did it for a long time, but then... yeah, she wants it to be proper.’
		The pitfalls of the FAIR police	’If they [forms] turn into some sort of tick box exercises [for] which standard you’ve used to create your metadata... I don’t think that’s very good really. But having said that, there’s a lot of projects who are actually doing that.’
	Implementing data sharing policy	Challenges and considerations in implementation of policy	’So we can establish our policy, but if we don’t take account of existing policies, then our implementation plan is going to fail.’

Chapter 5

Findings

This chapter follows the structure of the analytical framework presented in Table 4.2 in section 4.4. First, a backdrop to the preparation of data will be given by introducing the findings on coordinative practices and identified coordinators in environmental monitoring projects. Afterwards, the findings on dependencies data workers experience when preparing data will provide practical insights into the flow of data, and what data workers depend on for this data to flow smoothly. Next, findings on participants' perceptions and communicative practices in the form of knowledge exchange will be presented. Then, a look at identified data sharing practices and challenges, both from researchers', policy advisors', and eLTER work package managers' perspectives will be presented. At the end of this chapter, you will find an overarching summary of the key findings.

Note that the titles by which participants are sometimes referred to do not always correspond to their actual job titles. This is done to emphasise their role in the context of data preparation and sharing. Note that sometimes, these roles do not correspond to what is usually associated with the titles. Primarily, these roles include data analysts, work package managers, data workers, and project coordinators. The role *data analysts* is used to specify that the individual is primarily working with analysis of data, in that they analyse it for flaws and incongruities. The role *work package manager* includes what is often referred to as team leader, as well as work package manager. The role *data worker* is used as a collective term and refers to researchers that primarily work with data. This includes field technicians, laboratory technicians and data analysts. Additionally, work package managers can be considered part of this group, as all interviewed work package managers did data work in addition to their management responsibilities. The role *project coordinator* involves both what is commonly referred to as project leader/ manager, and project coordinator. I do realize that the latter often has a different set of responsibilities, but I found that the title of ’project coordinator’ better reflects the responsibilities of the role presented and discussed here.

Finally, note that the quotes presented in this chapter have been refined for readability, as well as translated from Norwegian in certain instances.

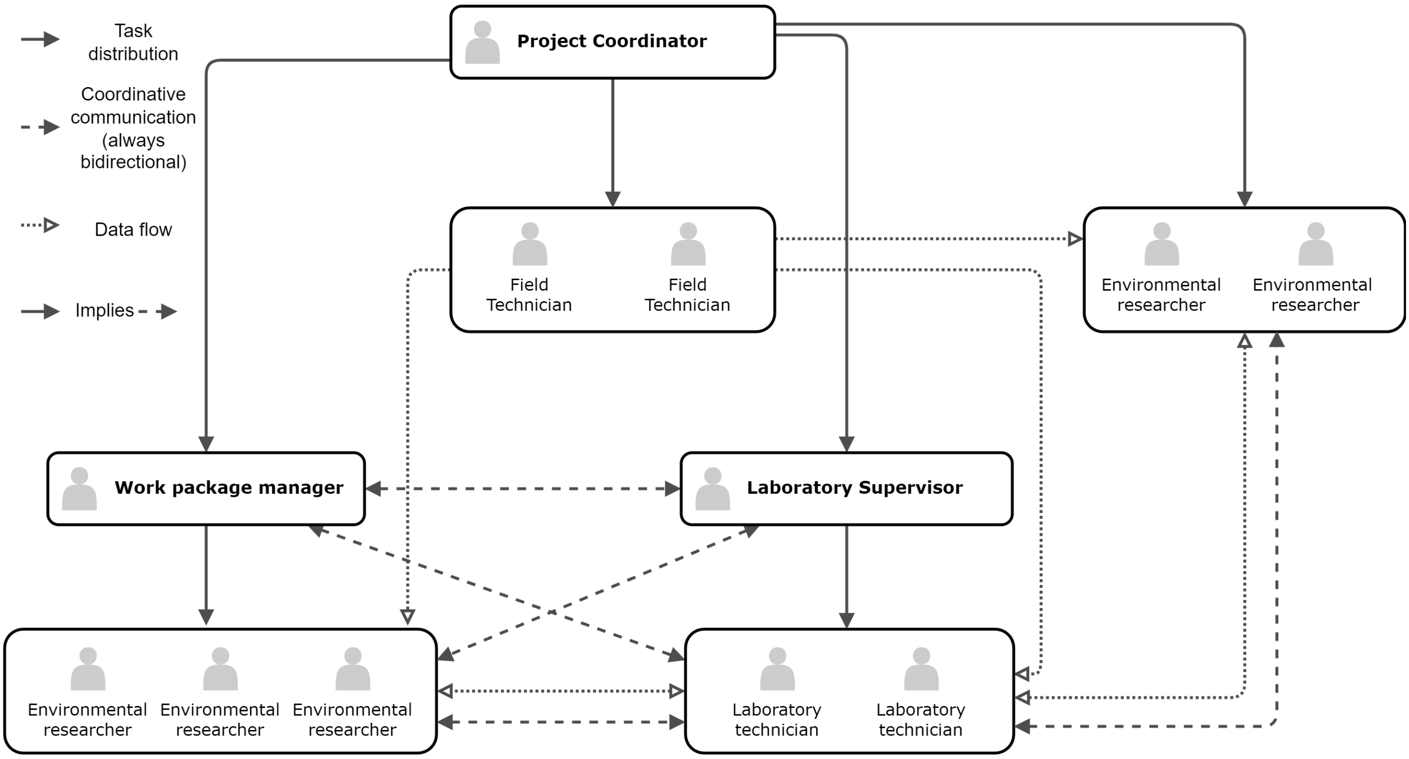
5.1 Coordinating data work

This section presents the coordinative roles and practises identified in the data analysis. Furthermore, it presents the three main identified functions of coordinators, these being (1) to have an overall view of the project(s), (2) to be communicators, either for their team or between re-searchers, and (3) to act as a personal organizer, helping people stay on task and prioritize. When interviewing participants, it became clear that project leaders, work package managers, and project coordinators all had very varied coordinative responsibilities depending on the institute they worked at. This implies that researchers’ tasks and responsibilities not only depend on their titles, but largely also on culture and general practice. Additionally, almost all participants worked on multiple projects at once, and differences were observed in how coordination was handled even on a project basis. Therefore, roles have been assigned to the participants based on what their reported tasks and responsibilities are rather than their job titles. The resulting findings are a generalized overview formed through the data collected.

5.1.1 Coordination overview

When coordinating data work in research projects, interviews with participants showed that there are two main *coordinators* involved: project coordinators and work package managers. In cases where laboratory work is involved as well, laboratory supervisors act as a third coordinator, performing similar coordinative tasks as the work package managers. An overall overview of coordination flow can be seen in Figure 5.1. In it, the role *environmental researcher* is used as a collective term for data analysts and report writers. Additionally, notice that the *field technicians* only have a direct communicative relationship with, and only get tasks from the *project coordinator*. This is not an absolute truth, but a simplification of what was mostly reported by participants.

Figure 5.1: Coordination flow of data work. The roles in this diagram are connected by three types of relationships, described in the upper left corner. Note that the *task distribution* relationship implies that a *coordinative communication* relationship exists as well. Furthermore, note that the coordinative communication relationship is bidirectional, as communication always flows both ways.



The coordination performed by the project coordinators and the work package managers can be divided into high-level coordination and low-level coordination. Project coordinators often do what is here referred to as high-level coordination: finding suitable people for projects, setting up timelines, budgets, drafting projects, and arranging meetings with work package managers to communicate the needs of the project. This work often involves coordination and communication with many people at once. This can be seen in Figure 5.1, where project coordinators coordinate and distribute tasks to field technicians, work package managers, (alternatively) laboratory supervisors, as well as environmental researchers in smaller projects. Interestingly, it was the project coordinators that reported spending a lot of time doing the lowest level of coordination as well; to simply act as a personal reminder, checking in and reminding colleagues of time constraints when necessary. This contrasts with the otherwise high-level nature of the project coordinator responsibilities. Additionally, when information systems are involved, project coordinators need to be able to communicate efficiently with software developers and tech support, as well as the researchers. This will be further discussed below, in subsection 5.1.2.

When projects are small, the project coordinators will take on the role of both project coordinator and work package manager, removing the intermediary work package manager from the chain of coordination. This alternative coordinative flow can be seen on the right-hand side in Figure 5.1, where the project coordinator directly distributes tasks to environmental researchers.

The work package managers usually take on the low-level coordinative tasks. This usually involves figuring out the logistics and distributing tasks received from the project coordinator, and solving issues that team members present with, both personal and work-related. In contrast to the project coordinators, most of the work package manager coordination happens on a one-on-one basis, only occasionally involving an additional party. Work package managers reported a highly autonomous work environment. This is reflected in the way that team members only seek out their work package managers if the issue at hand doesn't seem to have an obvious 'recipient', such as an expert in the field that they encounter difficulties in. The work package managers interviewed were also all part of the team in that they did analysis and/or ensuring data quality, making coordination a smaller part of their overall duties. Additionally, the coordinators reported that the coordinative work they were doing was not something they were formally educated for: both work package managers and coordinators came from either a strictly scientific background in the field they were coordinating, or a background in science and information systems. Thus, as one project coordinator described it, they were expected to 'jump into it' without much guidance or knowledge of supporting tools and systems for their managerial responsibilities.

5.1.2 Coordinators being a necessity

When asked about who participants most relied on in their daily work, multiple participants answered that this was either their work package manager, or their project coordinator. It is arguably a given that projects depend on funding applications being submitted, timelines being set, and budgets being drafted. Thus, the findings in this section will not focus on the technicalities of coordinative practices. Instead, a closer look will be taken at the vital role of project coordinators and work package managers in a social and organizational context. Three key functions of coordinators were identified. These are presented below.

(1) Coordinators have an overall overview of the project(s). As project coordinators deal with the high level-coordination described in subsection 5.1.1, they naturally end up best equipped to have overall insight into the different projects. The same goes for the work package managers, only in the context of the project (or part of project) that they are managing. Thus, the role of the coordinator becomes to use the overview of projects and tasks to keep everything consistent and adhere to the restraints of the given project(s).

(2) Coordinators are communicators, either for their team or between researchers. Project coordinators need to create communication between the necessary work package managers when needed. Additionally, when information systems are involved, a background in both the scientific domain of the coordinated data workers, as well as the technical domain of the software developers and/or tech support is needed. It is vital that project coordinators can communicate efficiently with software developers and tech support, as well as the data workers and other parties that might be involved. This often means going back and forth to relay information. This proficiency in 'both languages' is what one project coordinator described as 'a necessary middle way' as researchers do not always know how to communicate their needs to software developers, and software developers do not always know the necessary background for these needs. She goes on to explain:

> 'I'm like, the middle of the whole spider web. So my role is essential to make sure that this program works like it should, because I keep the communication running between all the different levels. So if I was removed, the communication would stop.'

Work package managers on the other hand, need to be able to facilitate open communication between their team members, and also communicate with external parties if needed. Thus, both project coordinators and work package managers are responsible for communication flow on a high and low level, respectively.

(3) Coordinators act as personal organizers, helping people stay on task and prioritize. Project coordinators spend a significant amount of time reminding others of their work, and correcting their prioritization of tasks. One project coordinator felt that her task was to act as a human ’pop-up reminder’, describing the reason for this need to actively monitor other people’s progress:

> ’I think people are in general... like, people in research are very busy people. People have loads to do, and several tasks at the same time. So I think it’s important for efficiency that you have someone who’s like, "Hey, OK, now we need to do this" ... Instead of you like having to have the total overview of all of the projects that you’re in all the time.’

In contrast to project coordinators, work package managers reported a highly autonomous work environment between team members. However, helping out with prioritizing tasks and doing other organizational[1] activities was mentioned as one responsibility. As one work package manager reported:

> ’If they come to me and say “I have this and I have that, what do I do?”, I’m the one that is responsible to say that “you do this”, and I take the responsibility for that. They don’t have to do that, they just work’

This suggests that work package managers not only provide necessary assistance in matters directly affecting individual work, but also take responsibility for it. Thus, the findings indicate that the work package managers both assist in personal organization, and act as a support system to fall back on when necessary.

5.2 Preparing the data

Arguably the most apparent part of how environmental researchers facilitate data sharing, is the preparation of data. In this section, dependencies participants experience when preparing data will first be presented. Two types of dependencies will be identified and explained: trust dependencies and practical dependencies. Thereafter, the findings on standardization of processes and the consequent effects on data work will be presented. Finally, the role of knowledge exchange and cooperation for data preparation will be presented. Note that the term *preparation* is used instead of *curation*, as curation often implies the care and management of data after it has been prepared, which are acts not discussed in these findings.

[1] In this context, organizational is used to describe that coordinators are organizers, *not* that coordinators are part of an organization and that the work they do is especially concerned with this fact.

5.2.1 Dependencies in data preparation work

One can divide dependencies in data preparation work into two categories: *internal* and *external* dependencies. The former is the dependencies researchers have inside their own organization and projects. The latter is the dependencies outside researchers' organizations, such as on data from external institutes or projects, and will not be discussed in this chapter. Note that in this section, there will be references to external data that is external in the sense that it has been collected by individuals not employed at the researchers' institute. However, this data is seen as creating *internal* dependencies, as the main reason for collection is to provide context to *internal* projects.

When talking to participants about dependencies in their work, two things became clear; the dependencies data workers experience vary substantially on a case-by-case basis, and data workers are often not aware of their dependency on colleagues or external factors. The latter is not an indication that some data workers suffer from poor self-reflection, but rather a reason to believe that working relationships are often overall good, and that good coordinative work has been done by project coordinators and work package managers, making the dependencies problem free and thus easy to overlook. When probed, this was acknowledged by one environmental researcher who agreed with the statement that she didn't really think about dependencies because of the rarity of problems occurring. She went on to explain:

> 'One doesn't really get further before the other [colleague] does what they have to do. But that mostly works out as long as we just agree on it.'

The dependency on one's colleagues delivering their work before being able to do one's own can be observed through the whole chain of field technicians, lab technicians, data analysts, and report writers. This will be discussed below. In the subsections after, findings on the two main types of dependencies identified between data workers will be presented. These are *trust dependencies* and *practical dependencies*.

Links in the chain of data flow

The participants reported between two and four 'links in the chain' of data flow before data was finalized and ready to be shared. The analogy of a chain where the '...chain is not working if not all the links are there' was used by a participant during an interview, communicating a strong dependency on all people, or 'links', that are involved in preparing data. These links include (1) Field technicians, (2) Lab technicians (often referred to as lab analysts), (3) Data analysts, including researchers ensuring data quality, and (4) Report writers. The number of links depends on project size and workplace. A visual representation of the data flow can be seen in Figure 5.2. The only links that were present in all scenarios discussed with participants were (1) and (3), as some data did not need to be treated in a laboratory (e.g. sensor data), and some researchers did both analysis and writing reports, taking on the role of both link (3) and (4). Even in the case of sensor data going straight to link (3), field technicians do need to make sure that sensors are performing well and do not show signs of technical failure, making (1) a necessity in all cases discussed with participants. As seen in Figure 5.2, data always flows from left to right, e.g. from (1) and onwards. Observe that there can be possible data flow backwards in the chain, as sometimes data has to be reevaluated and sent anew. This is always done with the *previous* link in the chain. The possible data flow also signals a dependency, as the *reason for reevaluation is a dependency on this data to be correct*.

Notice that report writing is included in the chain, despite data not going through any additional changes after the analysis. This is because data results are often reviewed and published in a report form, and the raw data will only be made available after, often by the project owners. Additionally, project coordinators are not part of the data flow chain, as the coordinating role does not directly interact with data. Project coordinators do however fulfill an essential role of overseeing and managing the whole process by being ’in the middle of the whole spider web’, as one project coordinator put it.

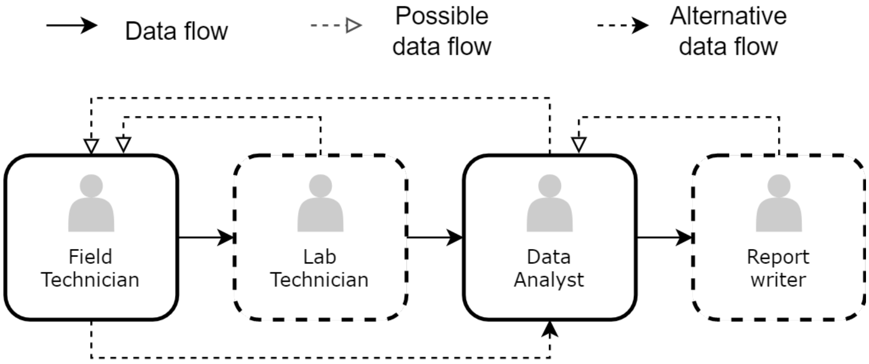


Figure 5.2: Roles that act as ’links in the chain’ of data flow. The roles that have dotted edges represent removable roles, e.g. they do not exist in all projects (roles (2) lab technicians and (4) report writers).

Trust dependencies

Trust dependencies imply that the dependency is one where the recipient[2] does not necessarily have any ways of verifying if the dependency is broken, and/or the process of doing so is cumbersome. Thus, the services[3] provided through these dependencies are based on trust in the provider[4] to deliver. The trust dependencies identified in interviews with participants are presented below.

Dependency on transparency. Data workers are trust dependent on other data workers being open and honest when they discover uncertainties, weaknesses or limitations in their data. One participant expressed that in order to be trustworthy, one should 'at least deliver results with some attachments. Like OK, the recovery is low, we can see traces of these and these and these compounds. This openness... it's really important for me'.

Dependency on field technicians to identify errors. As field technicians are the first link in the chain of data preparation *(see 'Links in the chain' in section 5.2.1)*, the subsequent data workers that receive data from the field technicians rely on their ability to identify signs of technical failure in sensors, other errors that can occur in sensor data collection, as well as errors in data collected through other means. Many participants had field experience, and were to a certain extent capable of identifying errors in raw data, however, they expected and trusted field technicians to do this job.

Dependency on lab technicians to do proper quality work. Similar to the trust dependency data workers have on field technicians, researchers that work with ensuring data quality and writing reports depend on lab technicians to do what one participant described as 'proper quality work'.

Dependency on external data submission accuracy. Multiple researchers reported that complementary, yet crucial data was sometimes collected by external individuals, such as farmers filling out information about their plowing activities. This data is a necessity for water[5] data analysis, as it directly affects sensor data.

[2] Here, I use the term *recipient* to describe the individual entrusting another with some task or responsibility, and relying on the resulting outcome.

[3] The outcome the recipient expects from the trust dependency.

[4] I chose to use the term *provider* to emphasise the role of the individual a *recipient* is dependent on, and their responsibility in providing some service.

[5] In this example, the participants were specifically talking about agricultural runoff, which is

If, for instance, a farmer reports the wrong dates for when his fields have been plowed, identifying this as the source of the subsequent water data fluctuations will become challenging. Thus, data workers are trust dependent on the *accuracy* of the reported data. The trust dependency on external data accuracy from farmers was the one case that participants reported feeling distrustful of. Both participants working on this project reported that they often did not believe that the submissions were accurate. This made their data analysis work challenging, as they did not know if anomalies in their sensor data were due to incorrect dates in the farmers' submissions, or if these anomalies were rooted in other issues.

Dependency on trust between colleagues to discuss and evaluate results. As will be presented in subsection 5.2.3, an integral part of achieving quality data lies in the knowledge exchanged between researchers. When data workers collectively work together to discuss and evaluate results, they depend on the trust relationships that enable them to work together to solve problems. If, for instance, a data analyst doesn't trust in their colleague's expertise, the knowledge exchange cannot take place.

Dependency on other people's expertise and competence. Researchers working in a modularized fashion, in the sense that highly specialized data workers constitute one part in the chain of multiple data preparation steps *(see 'Links in the chain' in section 5.2.1)*, often do not have the expertise to effectively evaluate the validity of their colleagues' work, or the work of researchers providing external data. Even in cases where researchers *do* have this expertise, it is hard to identify errors unless the data has evident abnormalities. Consequently, researchers are dependent on the quality of data work done by colleagues and external data providers, in order to achieve quality in their own data. Note that this dependency differs from the 'dependency on trust between colleagues to discuss and evaluate results', as the latter concerns itself with the back-and-forth nature of solving problems together. The dependency researchers experience in regards to other people's expertise and competence is also a *verifiable dependency*. This will be discussed below, in the next section.

Practical (and verifiable) dependencies

Practical dependencies often materialize in more direct ways, where the recipient depends on explicit acts of the provider that are easy to verify. Note that practical

analyzed in order to track the time trends of water contaminants as a result of farming.

dependencies still imply that there is a trust relationship between the provider and recipient. However, the verifiable nature of these dependencies means that one does not have to rely on *trust in that the resulting services exist*, as is the case with trust dependencies. Five *practical* dependencies are presented below.

Dependency on colleague(s) to supply data. As discussed in section 5.2.1, all 'links in the chain' of data flow need to be present in order for data preparation to happen. Consequently, data workers depend on the previous 'link in the chain' to supply data in order to do their own job.

Dependency on project coordinator and/or work package manager. As discussed in subsection 5.1.2, data workers depend on their project coordinator and/or work package manager for communication with different researchers, prioritizing tasks, task reminders, and guidance in both work and personal matters. These all require that the recipient has confidence in the provider, as this confidence is a precondition of the recipient handing over responsibility instead of fixing their problems on their own. This dependency is verifiable in the sense that the outcome from this dependency can be assessed (if, for instance, a project coordinator forgets to remind a data worker of his deadlines). Note that guidance given in both work and personal matters has an element of *quality* that creates a trust dependency (specifically, a *dependency on other people's expertise*), however, in this case we only look at the act of providing this guidance, which is a practical dependency.

Dependency on colleagues to discuss and evaluate results. Data workers depend on having colleagues to discuss and evaluate their data as elaborated in subsection 5.2.3. Note that this differs from the trust dependency 'Dependency *on trust between colleagues* to discuss and evaluate results', as this practical dependency concerns itself with *having a conversation partner to discuss ideas with* rather than *trusting that their conversation partner has correct ideas*.

Dependency on other people's expertise and competence. In this instance meaning a dependency on other experts' ability to do what others cannot. This is different from the *trust dependency* on other people's expertise and competence, as the emphasis is put on the act of *doing the work* rather than *this work being correct or of quality*. This dependency can for instance be observed between data workers and tech support (including in-house developers), where the former needs help with putting atypical data into databases, correcting atypical data, converting data into atypical formats, needing new functionality in the software they are using, and similar issues. Interestingly, when interviewing participants from one institute, it became clear that some data workers depended on their tech-savvy colleagues to do tasks that require nothing more than running data through a pre-programmed script. This suggests that sometimes the perception of other people's expertise and specialized competence has just as much, if not more say in peoples' dependencies on this matter. Furthermore, this suggests a need for education that addresses the use of tools and software for data work.

Dependency on external data submission. E.g. farmers filling out information about their plowing activities, which affects sensor data. This dependency is concerned with the actual act of receiving data. Participants reported that they experienced considerable delays in getting external data from farmers, and that field technicians sometimes had to physically show up at farmers' doors, sometimes repeatedly, in order to get the data they needed for data analysis. This created an unfortunate bottleneck, as much of the data the participants worked with was influenced by farmers' plowing activities, and the lack of this external contextual data meant that variations in data were hard to explain.

When things don't run smooth

A final note on dependencies is that while they often are problem-free and thus go unnoticed, they do in fact become very evident when problems do arise. This was the case of one environmental worker, who reported that 'everything we do is actually dependent on [colleague]', and that it was 'very scary actually'. She went on to explain that her colleague was the biggest bottleneck in her work, and she expressed that she often felt like they were not motivated and would therefore take much longer to do the tasks she depended on. She did not feel like the colleague listened to, or cared when she would ask and remind her of the tasks she would wait on. This led to quite a lot of tension and issues on her side. On the other hand, when later interviewing the colleague, she reported that she really did not have any pressing matters usually, and that she spent a lot of time at work solving issues she had found and wanted solving. She also seemed to see the importance of solving the issues that colleagues came to her with, but did not seem to find their matters urgent. Furthermore, she reported no tension in the work environment, and seemed to have no issues with the colleague who had previously painted a picture of a somehow tense workplace atmosphere.

5.2.2 Participants' experiences with quality work

The majority of the participants in the researcher sample reported spending most of their time ensuring data quality[6]. This included people that had split positions where only half of their time was supposed to be spent doing research. Below, the findings are divided into two parts: participants' perception of quality work[7], and standards for realizing quality data.

Participant's perceptions of quality work

Participants' perception of what 'good' quality data implies was relatively consistent, with the common denominator being *data accuracy and completeness*. Several participants reported spending a significant amount of time making sure that both their own work, *as well as the work of their colleagues* was accurate. However, there was one diverging opinion on the meaning of 'good quality data', being that of quality metadata in relation to reuse and understanding by external parties. Here, some environmental researchers felt that other people not understanding your data implies poor metadata, while others expressed that seeing data and metadata differently is unavoidable, suggesting that how people understand your data doesn't necessarily reflect its quality:

> 'One person can look at the metadata and see things. And another person can look at the same metadata and not see it. And a third person could want to have different metadata to look for different things.'

The diverging beliefs on metadata and its importance were also evident when interviewing the participants about their own work practices in relation to creating metadata. While researchers from one institute reported strict standards and forms that needed to be filled out, others had much more freedom. The researchers reporting strict standards also had stronger opinions on metadata, and interestingly, they were the ones that showed diverging opinions about the relation between misunderstandings of this data in the context of sharing and the quality of a data set. Notably, participants from this institute had very little experience with data reuse. Researchers with few guidelines on the other hand, seemed to view metadata as a much more flexible concept, taking a case-by-case approach when judging its definition, use and importance.

[6] Participants reported spending around 60%-80% on quality ensuring activities such as cleaning data (such as fixing incorrect or incomplete data values) and adjusting model parameters to make sense of this data.

[7] Quality work is here used to describe the work researchers do to ensure quality data.

Apart from the diverging opinions about metadata and the importance of external parties understanding your data, participants shared similar beliefs about the significance of quality data. All participants reported that they thought quality control was an integral part of data work. Additionally, the following opinions were expressed:

Accuracy of data is important. As mentioned above, all participants agreed that the *accuracy* of data was a big determinant of its quality. However, the term *accuracy* wasn't consistently used across the participant sample. Data workers at one institute reported that even though accuracy is a desirable trait in data, it is not realistic to achieve complete correctness of individual data values. As especially sensor data is often incomplete and one needs to fill out a great number of 'blanks' in this data, it wasn't seen as realistic either. Thus, the participants took on a more pragmatic approach to accuracy, where it was used to signify the overall accuracy in what data *indicates*, in terms of trends and cause-effect relationships. One data analyst explained the process of achieving such accuracy:

> 'Then I look [at the data], "Yes, that looks right. I remember that it rained on Tuesday. It looks like there's been more rain on Tuesday. And there I see an absence [of data]. Why is there no data? Let's see if I can find the data somewhere else, since there's a few hours missing". Or, "Was there a power outage? Was it maybe a lightning strike?" ... So stuff like that, so that we have a continuity in the data, so that we don't have *too many errors* [emphasis added].'

This sits in contrast to the participants working closely with laboratory analyses, to whom the term *accuracy* signified precise individual data values.

It's important to feel like what you deliver is of quality. As one environmental researcher put it, 'quality control and quality assurance is the most important thing, because if you can't deliver good results, why bother to write about it?'. One participant reported that as a senior researcher she would hold her colleagues accountable and give constructive feedback when needed, both in relation to their execution of tasks and the resulting data, as she believed this would foster the desire to care about the quality of data in her colleagues.

Quality work doesn't happen by itself. All participants involved with laboratory and data analysis expressed that a lot of time and effort was put into ensuring that their work is of quality, consequently bringing along compromises. Participants reported spending most of their time solving tasks related to anomalies in the data. Both identifying anomalies and finding suitable solutions for these anomalies means that one needs a high level of expertise, especially since this work is highly context dependent. One participant explained, 'there is not a single type of quality check'. Another participant expressed that quality work is 'quite the hassle', and that complicated metadata standards had often created bottlenecks in the data flow. A third participant confessed that the importance of overall accuracy in her work as well as discovering deviations and errors as soon as possible was the thing that kept her going as she exclaimed that quality assurance and control is not only a lot of work, but also 'so boring'. Additionally, some participants reported that the amount of work and the timelines they needed to adhere to meant that there was a constant juggle between balancing quality with deadlines and efficiency.

Wide variance in standardization of processes

Participants reported a wide variance in standardized methods, requirements, and controls to ensure data quality. While participants from one institute reported that most, if not all work for ensuring quality data was done on their own initiative with the motivation being that one has to be able to 'stand by it in the long run', the rest of the participants reported that while they saw the importance of quality results, they also needed to employ various methods and standards enforced by their institute's quality assurance and control measures. This included using analytical/internal standards[8] when analysing standards in the laboratory, filling out deviation forms, running data through automatic checks before sharing it in internal and/or external data sharing portals, and similar measures. Additionally, participants from one institute reported external assessments and audits, and that the results of these were included in metadata when sharing data externally. They stressed that this was an important part of transparency, and that they saw this as an important aspect of ensuring data quality and credibility of their institute.

Note that although filling out forms and running data through quality checks was a significant part of participants' tasks, this was not reported to be the most time consuming activities. A lot of the work participants reported spending time on was not anything outlined in standards and methods explicitly, but rather the byproduct of these, such as the act of solving questions about why data doesn't adhere to these standards. This meant working with the knowledge and existing data they have to fill out gaps in or correct lacking or uncertain data. Consequently,

[8] Controlled samples that one analyses with one's own samples.

a lot of the work was making informed judgements about whether or not data corresponded to the real world, and using the resources they had to decide what to do with these judgments. Communication was an important resource in this problem solving, and will be discussed in the next section, subsection 5.2.3. Interestingly, the participant group that reported little to no quality assurance and control in the workplace, reported spending approximately the same amount of time[9] doing quality checks and solving questions related to data abnormalities.

5.2.3 Knowledge exchange for realizing quality data

The majority of participants reported that communicating with colleagues and exchanging knowledge while doing so, was an integral part of doing their work, and doing it with quality. I choose to use the term ’knowledge exchange’ to emphasise that this form of communication is mutually beneficial, and facilitates an ongoing, continuously evolving relationship. Presented below are the findings on two main identified categories of knowledge exchange; *intentional* and *spontaneous* knowledge exchange.

Intentional knowledge exchange

Participants whose work primarily consisted of ensuring data quality reported that while working on data, an important part of progression involved a high degree of back-and-forth[10] with either other colleagues doing the same work, laboratory technicians, or sometimes even field technicians, depending on the size of the project. This included mostly discussing specific data sets, data samples, data values, or overall trajectories. Most of the time this communication happened through physically stopping by or sending a message online. Sometimes, if deemed necessary, a meeting would be scheduled either online or in person, where data would be discussed in depth. This form of predominantly informal knowledge exchange enabled researchers to execute their tasks in an environment that was still perceived as highly independent, while simultaneously and effectively using

[9] In percent, as reported by the researchers

[10] The term back-and-forth is here used to convey a similar process to that assigned to this term in Parmiggiani, Østerlie and Almklov’s paper [36] on the work of finding and preparing data. Except here, I chose to use back-and-forth as an expression of a ’...movement where the new data are interpreted in light of the old data and the old data might be reinterpreted in light of the new data’ [36], and then further expand it to include not only new and old *data*, but new and old *knowledge* that is related to this data, as well as all the communication between colleagues facilitating this data and knowledge exchange.

colleagues' unique experience and expertise.

Importantly, this knowledge exchange enabled colleagues to build a shared body of knowledge over time, which was evident in the symbiosis of some of the observed participants, some of whom had worked alongside for decades. At one point, when interviewing two tenured colleagues together, one said to the other jokingly, '...you can close your ears now, you've heard this a thousand times before'. This exchange demonstrated that data workers become aware of their colleagues' own body of knowledge in parallel with building their repertoire of exchanged stories and experiences. Furthermore, this knowledge exchange enabled participants to learn from past researchers indirectly, thus, as one participant put it '...building on experience from people before'.

Additionally, participants reported that they would schedule pre-planned meetings to discuss data and results, both at the finishing stages of projects, as well as throughout whole projects. One environmental researcher described an online chat where researchers would contribute ideas and knowledge over time, and how this was used:

> 'And then we were having a Teams chat with about 10-15 people where we looked at preliminary data, and we discussed it on this chat. And it was really exciting because people came in with- some are modellers, and some had trajectory, and some had some satellite data which they showed ... And in the beginning we didn't know if it was dust or if it was forest fire, and we wanted to quantify how much it was from each. And yeah, took half a year and then we... [laughs] had a paper.'

Spontaneous knowledge exchange

For many, the Covid-19 pandemic resulted in relocation to home office, bringing along abrupt change. This abruptness presented a unique opportunity to look closer at the spontaneous informal communication that happens as a result of co-location, such as in hallways, break rooms, and similar places, as participants were acutely aware of the effects of the sudden change to remote communication. The resulting findings outline how researchers experienced the differences in communication before and after national and regional workplace measures in Norway took place.

Interestingly, most of the participants reported that little changed for them work-wise when working from home instead of in office. They reported that they did not experience any decline in tasks requested by colleagues, and that the job they did was not affected by their physical location, as communication with colleagues could happen over Microsoft Teams or email. However, when probed further, many participants expressed that by missing the informal conversations and interactions with coworkers, they also missed out on a lot of important discussions and brainstorming activities. One data analyst described how the absence of her daily ’coffee machine breaks’ affected her chances of knowledge exchange with colleagues:

> ’I think sometimes you don’t really understand that there is an issue before you talk to people. I mean, just among the... by [the] coffee machine you can talk about, you know what your... whatever problem, and then, "Oh yeah, maybe that is-", and then there can kind of be discussions which you haven’t thought about before.’

As this wasn’t something they specifically sought out when working in office, the ’coffee machine talks’ disappeared once people started working more from home. One participant described this as the *’...kind of things you don’t know that you should know’*. This notion of ’things one doesn’t know yet should know’ separates the spontaneous from the intentional knowledge exchange, as people cannot consciously, or purposefully, seek out this knowledge. Thus, the absence of co-location during remote work effectively removes researchers’ chances of gaining new and *unexpected* knowledge. Finally, this spontaneity also allowed for information to be exchanged across teams and projects, as both informal conversations and informal spaces allowed for people to join in on conversations with colleagues they would otherwise not talk to.

5.3 Sharing the data

In this section, findings on current data sharing activities and identified incentives for these activities will be presented, followed by findings on the challenges of implementing data sharing policy in highly heterogeneous research infrastructures.

5.3.1 To share or not to share

This section will first address findings on the current sharing practices of data workers. Furthermore, findings on the perceived drawbacks of standardization will be presented.

When interviewing the participants of this study, the predominant consensus was that the extent of the current data sharing practices and efforts are internally driven. Participants from one institute reported that they felt no pressures to share or open their data. When asked about the nationally funded environmental monitoring project one data worker worked on, she replied that they ’don’t seem interested’ and that she wished they were more involved, explaining that the only way the project was held accountable was by a yearly one-page summary of their findings. They also expressed that they had been trying to push for the development of a portal for sharing data, expressing the desire to have an easy way to share their data.

Some participants from other institutes reported some degree of standardized sharing, usually through internally or externally operated portals, as well as project requirements stipulating the sharing of (predominantly) processed and finalized data as part of project deliveries. These participants also reported stricter metadata standards and quality assurance and control (ref. section 5.2.2), and had a noticeably more aware perception of the implications of sharing data (such as the misunderstandings that can occur when others reuse it, as discussed in section 5.2.2), and the role of metadata.

However, especially in non-standardized, case-by-case sharing relations (such as through email exchanges between institutes), it was up to the data workers, and/or their superiors (such as section leaders or data sharing program leaders) if data was being shared, and to which degree it was shared. An example of such a case-by-case data sharing relation is described below:

> ’We have shared data with many... I have a lot of scientific publications where

> I have been in a meeting, and then I say, "Yes well, we also have data from cold climates", or clay, or livestock or something, and then I contribute to a publication'.

This type of data sharing, often involving somewhat coincidental conversations and groups of people, seemed to be prevalent in the participant descriptions. This meant that data workers often did not intend for their data to be shared from the very beginning, and would only add contextual metadata *when they were asked about sharing their data*, in order to minimize misunderstandings. Furthermore, another non-standardized data sharing relation was described by an environmental researcher, in this case getting data from another institute:

> 'I just send an e-mail to a person at [institute name], and ask if they could send data from... from this specific station.'

Both the participant above, and other participants, reported that this was common practice. Some reported that these non-standardized instances of sending or receiving data had become a regularity, in many ways making a 'standard' out of sending an email, getting a file, and doing so repeatedly. Most of the time the participants already had established connections at external organizations, however, it was not uncommon to ask colleagues for connections if they needed new types of data.

Multiple participants reported that the extent of this type of data sharing was often internally controlled, either by project leaders or upper management within the organization, not necessarily basing these decisions off of any external policies or regulations. When asked about what external parties could get data from their institute, a participant explained:

> 'That's not up to me to say, but I guess... the one who runs these pictures. I guess... he's a very [laughter], he's a guy that really likes to say yes. So I guess he would like to say yes. But in regarding this, we own the pictures and they need to... credit us and stuff.'

In another institute, an environmental researcher described her own desire to share data and how important she viewed it, but reported an obstacle in management. The new leader of the environmental program through which they used to share previously wasn't willing to share data on a case-by-case basis, as she wanted data sharing to happen through standardized sharing agreements only.

These findings suggest that there can be resistance to sharing data not only from researchers themselves, but also the management, and that the antithesis of ’to share or not to share’ can be conflicting within organizations. When asked about the struggles of resistance to sharing data, one information manager proclaimed ’data is power’, and that this was, in his opinion, why many were not willing to share their data. Consequently, when data sharing is driven by internal motivation and without policies or regulations to incentivize this sharing, the stability of data sharing relations and standards is diminished as one skeptical link in the system of researchers and management can prevent a whole project (or, in the case of resistance from management, a whole institute) from sharing data externally.

The pitfalls of the FAIR police

While some participants reported very slack guidelines for the format and contents of datasets, all participants from one research institute reported a plethora of metadata standards that one needed to adhere to in order to achieve FAIR results. Although these standards were seen as important and necessary, they were often overwhelming for some of the data workers. Participants reported that the complicated standards data had to adhere to sometimes prevented lab workers from sending data in after analysis, as they had unresolved questions in regards to filling out forms, consequently leading to data congestion in the laboratory.

The eLTER work package managers expressed worry[11] about funding bodies moving in the direction of evaluating FAIRness of data through controls and checks. One participant elaborated their concern:

> ’There is a lot of work going on [with] the implementation of FAIR principles. There’s a lot of work on standards ... And I think we’re also moving toward checklists and evaluation plans for how you check the FAIRness of particular data resources. Which I refer to as the FAIR police. So, you know, if these people come to your door and say, well, these don’t meet FAIR... So what? Are we now going to move to a situation where in the same way as, you know, if your data aren’t open, you won’t be funded? Is the European Commission also going to move to a thing where if you don’t meet the checklist on the FAIR police’s document, then you won’t get funding?’

The skepticism towards a ’FAIR police’ was further elaborated by pointing out that checking off all the boxes on a checklist doesn’t necessarily equate to good quality

[11] With emphasis on this being a personal view.

research.

When interviewing a participant who partially worked as a coordinator of a data sharing network in ecology, the biggest challenge she identified in data preparation and data sharing was *know-how*. This included the lack of knowledge surrounding standards, where one should put data, how one can retrieve data, and what one could do with other researchers' data.

From one eLTER work package manager's perspective, the absence of understanding standards was seen as a product of research culture. The participant proposed a better alternative:

> 'If you put it into the working procedures of open science, such as the idea of research objects. So, the record of research is increasingly digital. Therefore, the research record should be open as [an] additional asset all the way through the models and the data and the procedures and everything like that. And that is seen [in] how you publish your work. I think then it isn't.... and we've got experience of talking to scientists about this. Then it's very familiar to them. They understand it, and they understand how it works. It's then just a matter of what are the bits that implement it. Culturally, they get that, they understand it. But coming in with a whole lot of metadata standards, *they just don't understand that at all*[emphasis added].'

The worry of the participant was that over-standardization risks disincentivizing researchers from joining networks and infrastructures. The excerpt stresses the importance of meeting researchers where they're at, proposing education and improvement of existing procedures before standardization.

5.3.2 Implementing data sharing policy

One of the eLTER work package managers had an active role as a research policy advisor in addition to managing the development of eLTER RI. This presented a unique chance to examine implementation challenges and considerations of environmental research policy. Thus, findings on policy implementation will be presented from both a policy advisor's perspective as well as a RI developer's perspective. Note that eLTER RI work package managers will be referred to as eLTER work package managers for the remainder of this section.

When discussing policy creation and implementation with the eLTER work package managers, several challenges and considerations were addressed. When asked about the creation of policies, one eLTER package manger replied:

> 'That's the easy bit. The difficult bit is turning that [policy] into an actual procedure and an implementation plan in terms of how that happens. So we would say we support FAIR principles, we believe in open data, all that good stuff. Then we have to actually write the procedures and implement the tools that would actually do that, and I guess also... what happens if there's a breach of those policies, so GDPR for example. We have to have procedures as to what happens if we think we're not implementing that policy correctly or if any of the parties involved aren't implementing that policy correctly.'

The excerpt demonstrates that there is a long chain of processes that need to be skillfully crafted in order for policy to be successfully implemented in research practices. Creating procedures, implementation plans, and response plans in case of policy or implementation breaches, is what transforms policy intent into reality.

Another aspect one has to take into consideration when implementing policies, is the *existing policies* that govern researchers, as well as the institutes and networks they are part of. For instance, if a research site is part of both an institution and a network, who should they publish their data through? Who should get credit for this data? One eLTER work package manager elaborates this challenge in the context of a DOI[12] wrestling problem:

> 'So if a DOI is put against a data set, who is the institution that will actually do the minting of the DOI? So, if we created a data set in [redacted research institute name] and we minted the DOI against that data set, how would eLTER then handle the fact that they want credit as well because it's part of their network?'

The challenge of the DOI wrestling problem is partially dependent on history and research culture. However, a substantial part of it comes down to battles between policies. If one tries to establish a policy without consideration of the complexity of existing policies and their interaction, the implementation plan will fail.

The current state of individually managed data sharing practices was also brought up as challenge from a policy implementation perspective. When interviewing eLTER work package managers, one interviewee described the process of creating an eLTER RI as one '...moving from the current situation where it's a network of

[12] 'A DOI, or Digital Object Identifier, is a string of numbers, letters and symbols used to permanently identify an article or document and link to it on the web' [101]

existing sites, which is basically this coalition of the of the willing'. This was an important goal, as the current degree of decentralization and autonomy makes it especially challenging to effectively and collectively implement policies. The participant explained that in the case of eLTER, the status quo is that individual eLTER research sites have the freedom to choose how to go about open data, and which policy they are going for. Thus, the challenge is that one needs a suitable governance infrastructure to enforce the implementation and monitoring of the plethora of different policies sites have chosen to follow.

The rich collection of differing research sites further complicates implementation of new policies as all participating eLTER sites have different pressures and criteria they need to adhere to. There is no unified funding for eLTER sites, let alone environmental monitoring projects, and funding is often comprised of institutional, national, and EU support. Consequently, funding and other resources vary depending on the project. Furthermore, other criteria and pressures regarding standards, data management and publications vary substantially from project to project, site to site, and institute to institute.

Additionally, the eLTER work package managers experience the pressure and subsequent friction of being positioned between research sites and policy, and balancing this in a way that appeals to both researchers and policy advisors:

> 'Typically, policy advisors don't want access to data, they want access to insight. They want to know what's going to happen. They don't want to know what the time trend looks like. And that's what we need to do to add value to the data that's coming through the network.'

As funding decisions are ultimately the product of the opinions of policy advisors, whether directly or indirectly, research infrastructures need to take not only researchers' needs into consideration, but also the needs and wants of policy advisors. The pressure to deliver *access to insight* and not only *data* to policy advisors, means that infrastructure developers need to think of how to provide additional functionality on top of the data they incentivize to share.

In summary, the main challenges of balancing users of RIs and the relating policy in the development of eLTER RI were identified as (1) Creating procedures, implementation plans, and response plans in case of policy or implementation breaches), (2) Taking existing policies into account, (3) Having suitable governance infrastructure to enforce the implementation and monitoring of policies, (4) Taking policy advisors' needs into consideration in addition to researchers'. (5) The overarching challenge of doing steps (1) through (4) in a highly heterogeneous network of varying resources, funding, and motivations, with research sites that in the case of eLTER, are currently highly autonomous.

5.4 Summary of findings

The findings present the key processes and relationships in environmental data work, and how these influence the preparation and sharing of data with extra attention to quality. Additionally, the sharing of data was examined from a policy perspective, specifically how research infrastructure developers approach and deal with data sharing policy. An overview of the findings and how they are situated in relation to one another, can be seen in Figure 5.3.

Broadly, the findings provide empirical support for the following: (1) There are two types of coordinators, both of which are essential in supporting data preparation work, (2) The process of preparing data includes data flowing through several 'links in a chain', which sometimes implies dependency between links. Dependencies, and *trust* in these dependencies, is a necessary foundation for data preparation, (3) Perceptions of quality and knowledge exchange drive the processes involved in preparing data, and doing so with *quality*. The degree of standardization in data preparation was not found to be a significant driver of quality by some participants. Furthermore, the findings suggest that (4) Standardized *open* sharing of data makes data workers more aware of the role of standards for external understanding of data. Non-standardized, case-by-case sharing activities are both highly dependent on the data workers themselves, and subject to interference from within organizations. Finally, (5) the findings present the struggles of RI developers in enforcing, monitoring, and balancing policy in highly diverse research sites and institutes.

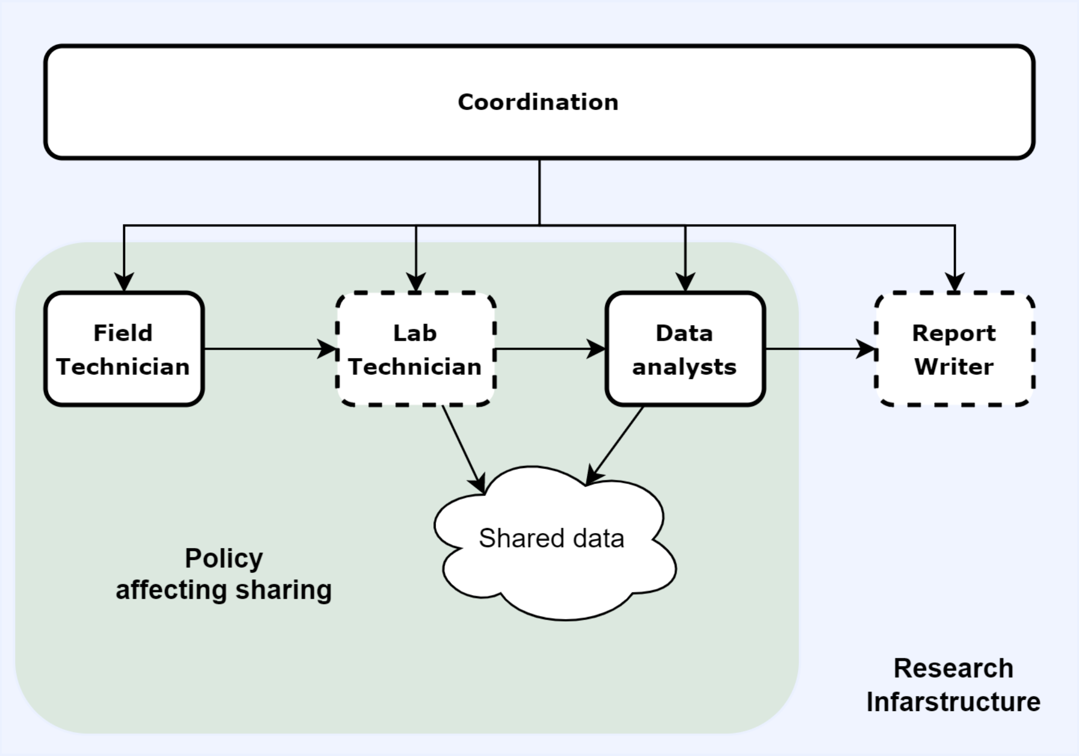


Figure 5.3: A total overview of the presented findings, showing the 'links in the chain', how these links distribute data sharing activities, and to which degree the identified data workers are affected by policy. In the context of this study, the coordinators were not found to be influenced by policy. Note that the diagram is a simplification, as the detailed coordinative processes, as well as the data flow are shown in Figure 5.1 and Figure 5.2 respectively. The arrows are a simplistic representation of data and information flow.

Chapter 6

Discussion

In this chapter, a socio-technical approach will be employed to discuss the findings, how they seek to answer the research questions identified in section 1.2, and how this ties in with the existing research presented in chapter 2. The discussion uses research infrastructures as a backdrop for the findings and what these suggest in an environmental research context. This chapter thematically follows the overarching chronology presented in the analytical framework in Table 4.2 in section 4.4, with the categories and most code groups appearing in order. There is however one important change: the drivers of quality in data preparation will be discussed first to answer the overarching research question RQ1, and give an overall picture of how drivers of quality may influence the sub-questions of RQ1. Furthermore, this chapter is divided into one section and subsection per research question and sub-question respectively. Lastly, as the discussion is especially broad thematically, an additional section presenting the ’big picture’ of the findings and what they imply will be found at the end of this document.

6.1 Data preparation: achieving quality data

Quality of environmental research data is arguably an important requirement. Quality assurance and control in the form of training, standardized methods, use of supporting software, and other practical measures has been recognized as an essential foundation to achieve quality data [16, 17, 47, 48]. In this thesis, a work-centric approach with special focus on social and cultural influences will instead be taken [21], as it is often overlooked in the context of facilitating quality work. Accordingly, the following research question is addressed: *What are the drivers of quality in data preparation?*

It is important to note that one can find many drivers of quality in the complex processes of data preparation. The trust dependencies presented in section 5.2.1 give a clear indication that there are many influences at play when striving for quality data. Thus, this section will address the quality drivers that were found to be both evidently important, and seen as interesting to look into. Furthermore, coordinating data work, as well as knowledge exchange, were found to facilitate quality work. However, as these both are addressed by research sub-questions, they will not be a part of the discussion in this section. Note that as the term 'quality' is highly context-dependent, this discussion will use the two main qualities that the participants associated with the term. These were data accuracy and data completeness.

The role of perception in quality data preparation

All participants in the researcher sample group agreed that ensuring data quality is as resource-demanding as it is important. These findings support previous research that elaborates on the tremendous efforts researchers put into preparing, sharing, and maintaining data [13–15, 18]. Furthermore, views on metadata varied greatly between the participants. Tenopir and colleagues' [61] interdisciplinary survey suggests that the majority of data workers believe both complexity and poor quality of data might be to blame if data is misinterpreted. In this thesis, the admittedly much smaller sample of study participants seemed to have stronger views, as some believed data will be misinterpreted no matter what you do, while others believed that misinterpretation was directly caused by poor data. This was especially interesting as the participants with the strongest variance in beliefs were from the same institute, suggesting that the work environment and local practices in regards to metadata are not enough to foster the same views on the issue. Additionally, these participants had little experience with data reuse. As multiple studies argue that 'good' metadata is often not enough for reusers of data [26, 37, 40, 57], one can argue that believing misunderstandings in data are directly caused by poor metadata indicates a lack of reuse experience.

Previous studies show that data workers take a pragmatic approach to data accuracy [15, 85]. The result of the study substantiates that this was indeed true with researchers that worked with big amounts of data at a time, who thought that the accuracy of a data set was determined more so by the overarching indications of this data set and consequent models, rather than the individual values that constitute it. Arguably, this is a necessary approach, as the rise in sensor technology and other methods of mass collection of data means that one has to look at trends over numbers[86].

In contrast to high-level data analysts that need to see the bigger picture in big datasets, lab technicians were found to value the accuracy of numbers. This belief was further encouraged through a myriad of standards that they had to adhere to. Arguably, this level of precision is also necessary, as small human errors can lead to significant, yet undetectable errors in data [82, 102]. The focus on data accuracy was motivated by a strong opinion of the importance of accuracy to achieve quality data. Interestingly, senior researchers saw the source of this motivation, and would hold their colleagues accountable to communicate the importance of quality results. This suggests that researchers not only feel like the data they deliver should be of quality, but that they are aware of how to make their colleagues feel similarly, and adjust their behaviour to achieve a collective feeling of caring about quality results.

The role of standards in quality data preparation

The presented findings suggest that despite all participants working at government-supported long-term monitoring projects, quality assurance and control measures and the prevalence of these greatly varied significantly. Participants from one institute reported very little regulation in regards to standards, even expressing that they wished for *more involvement* from their primary funding body, as this would motivate them. This is contrary to previous studies that indicate that heavy involvement from donors in the form of audits and the like have a negative effect on motivation and workers' behaviour in NGOs, despite these donors often being the only ones holding the NGOs accountable [103]. This is not to say that the findings indicate a need for heavily audited or standardized projects, but to suggest that there needs to be a balanced relationship between NGOs and their donors that signalises interest without being overly controlling or overbearing. Participants that reported audits and other control measures of laboratory performance were not of the opinion that these were negatively impacting motivation or performance. This further supports the idea that external involvement to a certain degree doesn't necessarily have negative consequences.

Additionally, despite a significant variance in standards and work procedures, both lab data analysts *and lab technicians* reported spending less time 'filling out the check boxes' outlined in standards, and most time trying to answer questions relating to *why* data doesn't adhere to standards or expected outcomes. This problem-solving is brought up by Mikalsen and Monteiro [43], who argue that researchers working with data that is especially prone to systematic uncertainty spend significant amounts of time doing *triangulation*, using multiple sources of information to create a more complete picture of the data [42]. The methods for this information collection included checking other data, using previous experience, and *knowledge-exchange*. This indicates that achieving data quality requires a high degree of both technical and social skills, and that these additionally need to be applied in a creative manner, as the problems data workers encounter are often new and unique.

The role of knowledge exchange in quality data preparation

Previous studies have shown that informal communication plays an active role in performing work activities [71–73]. What separates these previous works from the current findings, is that often this communication was observed in settings where workers used informal communication as a tool to perform work tasks simultaneously as this communication takes place. The communication thus led to an explicitly interdependent work environment. This was also the case of the data workers in the study, as they often employed a 'back-and-forth' communication strategy when solving problems such as assuring completeness or accuracy of data. However, they would additionally use informal communication as a vital tool to expand their knowledge and apply this onto their individual tasks later.

Furthermore, this communication was often not only casual, but also *spontaneous*, in that it happened in hallways, break rooms, and similar spaces. Just as in Brown and Duguid's [64] research on communities of practice where the retelling of Orr's [104] study describes how Xerox machine technicians would 'learn the craft' through exchanging stories by the coffee machine, one participant reported the same phenomenon. This reaffirms the importance of such informal spaces, and the conversations that happen in them. What was interesting in the case of the study participants, was that the spontaneity of their interactions with colleagues enabled them to acquire knowledge about what one participant described as the *'...kind of things you don't know that you should know'*. Consequently, data workers cannot actively seek out this knowledge, and are dependent on casual, unplanned conversations to discover new and important information.

Interestingly, this spontaneous knowledge exchange happens across teams and projects, a trait that characterizes communities of practice [62]. This was often found to be different from the intentional knowledge exchange, which would mostly stay within the boundaries of teams, projects, or even offices. The intentionality of this knowledge exchange also made it easier for participants to identify the significance of it, which was not always the case with spontaneous knowledge exchange which was often overlooked by participants during interviews. This trait of *overlooking its significance,* as well as the presence of communal learning as a result of spontaneous knowledge exchange, and the importance of 'coffee machine talks' and conversations in similar spaces, further indicates an existence of one or several communities of practice in environmental research institutes [62].

The informal nature and spontaneity of the conversations facilitating knowledge exchange seems to make this system prone to interference, such as in the case of changing from co-located to remote work. The role of co-location in informal communication has been studied with conflicting results on whether or not remote work effectively removes the possibility of informal yet crucial knowledge exchange [71, 76, 77], or can foster it through the use of technology [74, 75]. The general consensus supports the interdependent nature of co-location and casual communication. The participants were especially aware of this interdependency, as the relevance of the Covid-19 pandemic at the time of writing this thesis meant that remote working arrangements had put an abrupt end to this form of communication. Despite all participants reporting that they used Microsoft Teams for virtual meetings and chatting with colleagues, they reported that this was not a substitute for casual conversations in the office. The absence of a virtual alternative for casual communication (and especially the spontaneous kind) in the workplace is especially worrying, at a time when tools for remote work facilitation are in rapid development, and peoples' perception of remote work is changing.

6.1.1 Coordination of data preparation

Coordination and its importance has been discussed extensively in CSCW theory [65, 67, 105] as well as literary works from other fields [68, 69]. Thus, its importance seems evident in cases of cooperative work. When looking at the chain of data workers needed to prepare data, the question of how these data workers are coordinated arguably becomes an important one to answer, in order to gain a better picture of how the data preparation and its quality is achieved. Therefore, it seems fitting to propose the research question *What role do coordinators have in data preparation, and what expertise is necessary to realize this role in practice?*. The discussion in this section attempts to answer this question, with the help of this study's findings and previous literature on coordinative work.

Two main coordinator types were identified in the findings; project coordinators and work package managers. Their main responsibilities were socio-organizational in that they needed to effectively communicate with all involved parties, and often with organizational purposes, such as communicating the issues of one data worker with another, or organizing a data workers' tasks so that they would be able to complete them within certain time constraints. Their roles were often consisting of the starting points, endpoints, and all the intermediary connections that made data work possible, while remaining on the outside. This included distributing tasks at the start of projects, following these up, solving the intermediary problems pertaining to organization and communication that occurred during the project(s), and making sure all work was wrapped up at specified deadlines. Most of the participants were still involved with the data when they did not do coordinative work, however, this was due to having split positions, and the tasks involved in their coordinative responsibilities were still distinctly separated from their data work.

Furthermore, the analysis found that the role of the coordinators was to coordinate the cooperative work arrangements [65] that existed in the data preparation process. The data workers were concerned with their own data and their immediate dependencies, while coordinators needed to make sure that the data work was performed from the retrieval of data and all the way to its finalization. Thus, their tasks were to do the bulk of the articulation work [67] needed for the data to successfully flow from 'link' to 'link'.

Finally, the findings indicate that one needs both process expertise and domain expertise when coordinating data work. In contrast to coordinators in high-stress environments [69] where process expertise was the main necessity in order to know the 'who, what, and when', coordinating data work required domain expertise too. Coordinators needed process expertise to know who to contact when tasks needed expertise or just manpower, what to coordinate between researchers, and when certain things had to be done. Coordinators also needed to be able to 'speak multiple languages' to carry communication between the researchers, provide guidance when researchers had questions, and have enough domain knowledge to have a correct overview of all projects and tasks. Thus, domain expertise was just as vital in doing overall coordinative work. Interestingly, none of the coordinators interviewed reported having a background in management or other education that teaches coordinative skills, suggesting that process expertise, as well as tools and software knowledge for coordination, were accumulated through experience. This might indicate that educational institutions, research organizations, and governing bodies do not take into consideration the necessity of coordinators and the importance of their coordinative competence.

All in all, the role of coordinators was to be aware of, and coordinate the 'when, what, and who' of data work. Furthermore, the findings suggest that coordinators need both process expertise and domain expertise to do these tasks.

6.1.2 Dependencies in data preparation

Previous literature from other disciplines has studied the trust relationships of colleagues, as well as the trust data workers put in their data [43, 80, 81, 85]. Trust is undeniably an important factor in interpersonal relationships, in data work, and in science in general [79–81, 85]. Arguably, trust often translates to dependency. By putting one's trust in another entity, whether it be colleagues, tools or data, one often depends on some outcome, big or small. Thus, it was interesting to investigate possible dependencies in data work, and how, on the basis of trust, these dependencies define data preparation. For this purpose, the following research question is addressed: *What dependencies do data workers face when preparing data?*

Four 'links in the chain' of data flow were identified in the data preparation process. Edwards and colleagues [40] use the term 'data friction' to describe the friction that occurs when data goes from one 'link' to the next, costing energy and attention in order to move along. From a *dependency perspective*, this data friction can be seen as a result of the dependency between one 'link' and the previous 'link' that supplies the data. The findings suggest that this friction is often solved through the back-and-forth knowledge exchange (see more in section 6.1) researchers have both directly with the supplier in the 'link' dependency relationship, as well as others that they have a dependency relationship with, such as a trusted colleague with suitable expertise.

Furthermore, the 'chain of links' can be viewed as a cooperative arrangement [65], where data workers prepare data with *shared intentionality* [68], which sometimes is to make data prepared for sharing, and other times to simply find meaning and significance in this data. The data workers were found to be highly independent in the majority of their work, however, when needed, they would come together to cooperate on finding solutions, or depend on each other to provide the necessary data to continue their work. Thus, the dependencies from the 'link' formation as well as the plethora of other dependencies would appear, disappear, and reappear between different links at different moments in data workers' individual tasks.

Data workers experienced two types of dependencies; *trust dependencies(TD)* and *practical dependencies(PD)*. Interestingly, these dependencies were always observed together, where a trust dependency always implied that there was a practical dependency to build on. One can thus look at these dependencies as two halves coming together to form a composite, total dependency (such as depending on colleagues to supply data(PD), *as well as this data being of quality(TD)*). Arguably, to realize this total dependency in practice, one needs to trust *both the trust dependency component and the practical dependency component of the total dependency*. Thus, dependencies were found to be intrinsically *trust driven*. Furthermore, the findings indicate that trust dependencies were what influenced the *quality* of data preparation, while practical dependencies *enabled the processes* of this data preparation.

If one compares the findings on this trust driven total dependency to Butler and Cantrell's findings on what colleagues perceive to be the most important trust determinants [80], one sees that the two determinants, integrity (honesty and truthfulness) and competence (technical and interpersonal knowledge and skills), coincide with the foundation for the trust dependencies identified when interviewing data workers. Data workers trusted that their colleagues would be honest and truthful when needed, and that their colleagues had sufficient knowledge and skills to provide them with quality services. The opposite applies as well: In the few cases participants reported mistrusting the integrity and competence of the providers in their trust dependency relationships, they expressed that not having the means to verify these services was detrimental to their work (ref. 'Dependency on external data submission accuracy', section 5.2.1). If one looks at the *practical dependencies* found in section 5.2.1, it becomes clear that the trust qualities data workers are most concerned with change from integrity and competence to consistency (reliability and predictability). Data workers are dependent on consistently having individuals providing services, such as data or discussions about an issue they have discovered in their data.

Mayer and colleagues[106] point out that in interpersonal trust in work settings,

’...being consistent is insufficient to integrity, as the trustee may consistently act in a self-serving manner’ [106, p. 720]. This certainly rings true in cases of *trust dependencies*, and thus, it is reasonable that integrity is an important part of these dependencies. However, the concept of *practical* trust separates the need for integrity, as acting in a self-serving manner while being consistent would still fulfill the role of providing the service a recipient expects through this dependency. Note that as *trust dependencies* and *practical dependencies* form a composite trust, the role of integrity is still vital in the overall trust relationship between data workers.

Data workers were found to not be aware of many of their dependencies when preparing data. However, as the interviews progressed, the conversations revealed that dependencies were often the glue that held data preparation work together through practical dependencies (such as needing data from a colleague that is before oneself in the ’chain’ of data flow). Additionally, the dependencies data workers relied on enabled them to do quality work through being certain in the quality of the data they depended on (e.g. when being further along in the ’chain’ of data flow and using data that has been retrieved or processed by another colleague) and by seeking out knowledge exchanges with colleagues that had other experiences and expertise. The lack of awareness in regards to *just how much* they depend on each other seemed to be a positive attribute, as data workers became much more aware of their dependencies on colleagues when these dependencies created problems.

6.2 Sharing data: practices, challenges, and concerns

Previous studies have shown that sharing data has been deprioritized in research communities as reward systems are often narrow, and exclude the work that goes into sharing data and making it reusable [7, 37, 39, 56]. For this study, investigating (dis)incentives of sharing data was thus found to be an important task in order to achieve a complete view of what influences researchers not only in their sharing activities, but also in the totality of their practice. Thus, the discussion in this section seeks to address the research question *What (dis)incentivizes researchers to share data?*

Policy, and the lack thereof, was shown to influence data sharing activities. It was evident that researchers that were required to share data *openly* through portals were much more concerned with standards and their importance (ref. section 5.2.2), both for data and its metadata. This might just be a by-product of stricter quality assurance and control at institutes that have open data sharing mechanisms in place, however, it does suggest that the overall relationship to data changes once data workers need to prepare data for open sharing, without knowing who the recipient might be.

Interestingly, the findings suggest that the pressures of funding and policy on data sharing aren't always as prominent as previous literature on the topic shows [18, 26]. For instance, participants reported that even though they did share data, this was not because of funding incentives. Usually, the reasons were internally motivated teams and colleagues that found data sharing important, creating an organizational culture that promoted sharing efforts. This is supported by a survey done by Kim and Stanton [59], showing that normative and cultural-cognitive pressures are highly influential factors in determining data sharing behaviours. Another possible explanation for the evident willingness to share can be that researchers often shared data with institutes that were both geographically close and dealt with similar research, thus requiring less 'hassle' when preparing contextual metadata for the reusers, as data created in a similar environment to one's own is often easier to interpret [37, 57].

The findings further propose that since data sharing activities of this kind are unregulated, data workers do not work with a shared intentionality of making this data reusable from the beginning of the data sharing 'chain'. Since effects on data were not studied in this project, it is difficult to determine the consequences. However, the quality of data and metadata may be affected by this behaviour, especially since some participants reported that some metadata was created for data only after external parties requested this particular data.

Interestingly, the analysis found that data workers themselves are not necessarily the people that oppose data sharing activities; participants from one institute reported that upper management had restricted non-standardized sharing of data in recent years. This suggests that it is not only the data workers themselves that are hesitant in sharing data, which is what previous research often addresses [25, 61], but also individuals that do not have direct attachments to these data. Harris and Lyon [60] suggest that collaborations between different research organizations are highly dependent on trust, and that even incentives (such as funding) are not adequate for successful collaborations. Thus, when viewing data sharing relations as collaborative exchanges, their success is not only determined by trust between researchers sharing and receiving data, but also the interorganizational trust of the involved institutes.

Additionally, even though the participant sample of this research project did not consist of management that showed hesitance in data sharing, one can argue that an overall lack of reward systems might be the reason for such hesitance. Arguably, the researchers themselves gain new connections and build their network by establishing non-standardized data sharing exchanges in an environment of preexisting 'gift culture' [58], while management has to take responsibility for potential negative consequences. Furthermore, prior experience with data sharing *and reuse* has shown to decrease the perceived risk of sharing data [107]. Thus, if management lacks such experience, and in addition is not driven by normative or cultural-cognitive pressures if they are not actively involved in data work activities themselves, the perceived risk might outweigh the benefits of data sharing. Thus, it seems that EU's statement about researchers 'risking their careers' [38] with the current lack of risk regulation and sharing incentives might be just as applicable to the management of research institutes.

Concerns regarding rigid policy of standardization were expressed by both eLTER work package managers and researchers. Some data workers experience downsides of ensuring quality through complicated metadata standards, as these sometimes become the reason for *not* sharing data. This struggle was acknowledged by one eLTER work package manager who expressed that the focus on rigid standardization of data was unfortunate as the researchers 'just don't understand [metadata standards and reasons for them] at all'. Another participant reported that 'know-how' was the most prominent challenge when it comes to data sharing, supporting the notion that there is an existing lack of education and understanding in regard to metadata. This is unfortunate, as metadata is crucial for assessing data quality, as well as making data reusable and minimize risks of misunderstandings [25, 37–39].

One participant expressed skepticism in regards to an increase in focus on FAIR metrics, and a future of FAIR becoming a 'checklist to tick off'. The worry of a 'FAIR police' enforcing strict measures and controls of datasets, was that one risks no improvement in data quality while simultaneously disincentivizing researchers to contribute with FAIR data. The worry is substantiated, as EU's action plan on FAIR data proposes the development of metrics for assessment of both FAIRness support in infrastructures and FAIRness of data sets [38, pp. 68, 74], effectively increasing the rigidness of FAIR standards. Altogether, the action plan consists of several countermeasures consisting of education, development of further specification of metrics by communities themselves (thus allowing researchers to participate actively), and resource distribution that will allow for researchers to benefit from the standardization. However, the order in which this is and will be implemented, and to what extent the different measures succeed, is not clear. Thus, one arguably risks a skewed distribution of implementation successes and failures that potentially can result in overwhelming standards with little to no knowledge or resources to handle them. Finally, the efforts to unify inherently heterogeneous work practices through standardization risks unintentional new exclusions and gaps in the existing infrastructure [9], leaving questions to whether or not the tremendous amounts of resources planned for unification might be spent just for similar problems to emerge.

All in all, the findings indicate that regulative open data sharing makes researchers more aware of standards for data, and that missing regulative data sharing policy leads to an unstable case-by-case sharing environment that is dependent on both researchers' and their management's motivation to share. Furthermore, over-standardization of data risks being a disincentive for sharing data if researchers are not adequately aware of how to employ the enforced standards.

6.2.1 The influence of policy fragmentation on data sharing.

Jackson and colleagues [24] suggest that practice and policy are deeply intertwined, and that these have to be studied in relation to one another in order to see the ways in which they interact. Data sharing in the context of environmental research is especially relevant in this regard, as related policy is concerned with data being shared to solve environmental challenges [13, 53, 54]. Furthermore, this policy has been shown to be fragmented in research infrastructure settings, complicating an already complex situation[8, 9]. It therefore seems that employing a top-down policy perspective is important to understand the possible influences policy and its fragmentation has on data sharing. The research question that seeks to answer this is thus *How does policy fragmentation influence data sharing?*

Stahlecker and Kroll [8] discussed the challenges that resulted from a fragmented research policy in the EU in 2013. The findings show that this challenge pertains almost a decade after, as eLTER work package managers reported difficulty in implementing, enforcing, and monitoring policy in research sites that have vastly different pressures (e.g. requirements in data management and results), and resources (e.g. funding and tools), as supranational and national policies vary substantially. Thus, enforcing data sharing, and regulating this data sharing in a way that is feasible for all involved eLTER sites is a challenge that can not be solved through one single all-encompassing solution.

The effects of sustained challenges as a result of insufficient funding and other resources have been documented from a bottom-up perspective as well [14], where the suggestion is that a 'top-heavy' infrastructural template neglects the vital role of data curation, and thus the work that needs to be done to make data reusable, in the overall infrastructure ecosystem. If one looks at this issue from a broader perspective, one can view the resulting system as a negative feedback loop, where policymakers struggle with implementing and enforcing policy on their side, while researchers struggle with conforming to this policy as the result of disproportionate resource allocation.

6.3 Seeing the big picture

As the intent of this thesis is to examine the process of data preparation from multiple angles, as well as examining possible influences on this process, this section will attempt to give insight into the 'big picture' that connects the discussed findings, and highlights the relationships between them.

Firstly, it is important to stress that the findings were studied and presented in the context of RIs. In practice, this means that environments of less complexity, and with fewer outside influences and pressures, might behave differently. Secondly, the heterogeneous nature of RIs means that these findings will vary from other, similar cases. This variance was observed between institutes and even projects.

The discussed findings suggest that data workers prepare data in cooperative work arrangements, and with shared intentionality. The participants' experiences with sharing data also showed that often, the focus of this shared intentionality was not to make data shareable or reusable, but to make sense of it in order to explain trends and connections in the real world. Thus, the meticulous work put into making it shareable was often the last step of the process, and it only applied if and when researchers decided they were going to share this data.

The findings on challenges in policy fragmentation suggest that the challenges of implementing policy in highly heterogeneous environments such as eLTER RI may affect data sharing regulations. These, in turn, affect data preparation, as the focus of data workers' shared intentionality shifts towards at the very least making data understandable, if not reusable, for external parties. The findings on data workers preparing data for regulated *open* sharing, support the idea of a shifted shared intentionality, as the participants were much more aware of standards and metadata in such situations.

Furthermore, coordination was found to be a supporting, yet vital part of the processes constituting data preparation. This *articulation work* concerned itself with all the in-between, communicative and organizational tasks that let data workers focus their domain expertise on preparing data. Thus, the preparation process and all researchers involved were dependent on coordinators to coordinate. Additionally, the dependencies outlined in the findings show that data workers often relied on others in ways that were hard to verify, making a vital part of their dependencies *trust based*. All of these directly influenced the quality of their work. Overall, the findings demonstrate the interwoven relationship of data workers, coordinators, and policy in a research infrastructure setting.

Chapter 7

Conclusion

This thesis provides insight into the complex socio-technical processes of preparing data, with attention to the coordinative and cooperative measures supporting this, as well as the policy, and lack thereof, that governs the sharing of this data. It is important to stress that the study covers a broad selection of themes and study objectives. Effectively, this means that the work presented in this thesis acts as an overall insight into the factors that influence data from its conception to its finalization, and is by no means exhaustive. It does however present new empirical knowledge about the roles of coordinators, data workers, and the significance of their socio-cultural relations in the context of research infrastructures. Additionally, the study presents empirical insights into the relationship between data sharing, policies, and how these in turn might affect data preparation processes.

Project coordinators and work package managers were recognized as the two main coordinators of data preparation work. Despite coordinators being recognized as key actors in data preparation work through solving communicative and organizational tasks, none of the participants had any formal training for their coordinative responsibilities. This suggests that educational institutions, research organizations, and governing bodies can improve processes and workflow of data preparation through acknowledging coordinators and education for coordinative roles as an important part of their overall agenda.

Quality work was found to be driven by data workers' perceptions, their problem-solving abilities (which were both technical and social), and their knowledge exchange, all indicating that data quality is highly dependent on not only competence, but also the social and cultural aspects of data work. This is further supported by the findings on the different dependencies experienced by researchers, which, on closer inspection were found to have a practical component that enabled the tasks of data preparation, and one component that influenced the quality of this data.

Furthermore, the findings show that data sharing practices and regulations varied between the three institutes at which participants were employed. The findings do not indicate that environmental researchers struggle with balancing resources and policy demands, as previous literature suggests. Participants were found to be driven by social and cultural motivations when sharing data. Additionally, many participants expressed that they engaged in case-by-case non-standardized data sharing, where for some these were restricted by upper management, effectively destabilizing data sharing practices. Lastly, the findings brought forward issues of policy implementation, and how policy fragmentation continues to make this implementation challenging.

7.1 Limitations

As this case study was conducted over one semester with a relatively small participant sample (12 in total), this work has inherent limitations to its scope and generalizability. Thus, further research in different settings will be needed to test the generalizability of the findings presented.

A significant amount of the findings concerning implementation and enforcement of policy are based on the interviews with two eLTER RI work package managers. Thus, the data sample is not as diverse as with the other participant sample group. However, I do believe that the findings hold some validity as they are managers of the core eLTER RI development, and consequently, their views have significant influence.

7.2 Suggestions for future work

The findings have uncovered several avenues to be worth exploring beyond this thesis. Below, two of these are presented as suggestions for future work.

One fact that adds to the complexity of data curation and the resources required, is that environmental research includes a great amount of longitudinal studies. Thus, environmental, human, and infrastructural conditions are often subject to change throughout the data's life cycle [15]. Consequently, data needs to be adapted to continuously changing influences and ontology. This need for upkeep of data to ensure long-term value despite changes in technology, policy, and user requirements (often referred to as digital preservation [26, 108]), means that sharing and reusability ultimately requires a considerable amount of resources even after data initially has been through a preparation process [13, 15]. Digital preservation for data reuse can be seen as a continuation of the data preparation process explored in this thesis. Investigating the possible challenges, considerations, and influences of digital preservation could be of great use in gaining empirical insight into the overall infrastructure ecosystem and its interconnections.

Another path to pursue is to look into the role of 'tech support' for data preparation and sharing. Multiple participants reported that they often needed technical help with their work. This was only tangentially touched on as part of the dependencies data workers experience. The question that remains unanswered is whether this is an educational issue where the evolving digitalization of research demands a higher level of mastery in regards to digital tools and general software, or if data, for some reason, *needs* to go through researchers with high information system expertise in order to be successfully prepared and/or shared.

Bibliography

[1] M. Ivanova, *Coordination of scientific work for open data sharing in environmental research*, Norwegian University of Science and Technology, 2021.

[2] R. D. Kelemen, 'Globalizing european union environmental policy,' *Journal of European Public Policy*, vol. 17, no. 3, pp. 335–349, 2010.

[3] J. Carmin and S. D. VanDeveer, *EU enlargement and the environment: institutional change and environmental policy in Central and Eastern Europe*. Psychology Press, 2005, vol. 9.

[4] European Commission. 'Research for environmental policymaking: how to prioritise, communicate and measure impact.' Accessed: 10.06.2022. (2016), [Online]. Available: `https://ec.europa.eu/environment/integration/research/newsalert/pdf/research_for_environmental_policymaking_54si_en.pdf`.

[5] D. Ribes and C. P. Lee, 'Sociotechnical studies of cyberinfrastructure and e-research: Current themes and future trajectories,' *Computer Supported Cooperative Work (CSCW)*, vol. 19, no. 3-4, pp. 231–244, 2010.

[6] E. Monteiro, N. Pollock, O. Hanseth and R. Williams, 'From artefacts to infrastructures,' *Computer supported cooperative work (CSCW)*, vol. 22, no. 4-6, pp. 575–607, 2013.

[7] D. Ribes and T. A. Finholt, 'The long now of infrastructure: Articulating tensions in development,' 2009.

[8] T. Stahlecker and H. Kroll, 'Policies to build research infrastructures in europe: Following traditions or building new momentum?' Arbeitspapiere Unternehmen und Region, Tech. Rep., 2013.

[9] E. Parmiggiani, H. Karasti, K. Baker and A. Botero. 'Politics in environmental research infrastructure formation: When top-down policy-making meets bottom-up fragmentation.' Accessed: 11.12.2021. (2018), [Online]. Available: `https://blog.castac.org/2018/06/research-infrastructure/#_ftn3`.

[10] K. S. Baker and H. Karasti, 'Data care and its politics: Designing for local collective data management as a neglected thing,' in *Proceedings of the 15th Participatory Design Conference: Full Papers-Volume 1*, 2018, pp. 1–12.

[11] S. Rennie, K. Goergen, C. Wohner, S. Apweiler, J. Peterseil and J. Watkins, 'A climate service for ecologists: Sharing pre-processed euro-cordex regional climate scenario data using the elter information system,' *Earth System Science Data*, vol. 13, no. 2, pp. 631–644, 2021.

[12] eLTER. 'About eLTER RI.' Accessed: 04.06.2022. (2020), [Online]. Available: `https://elter-ri.eu/about-elter-ri`.

[13] H. Karasti, K. S. Baker and E. Halkola, 'Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the long term ecological research (lter) network,' *Computer Supported Cooperative Work (CSCW)*, vol. 15, no. 4, pp. 321–358, 2006.

[14] H. Karasti, A. Botero, K. S. Baker and E. Parmiggiani, 'Little data, big data, no data? data management in the era of research infrastructures,' University of Oulu, Finland, Tech. Rep., 2018.

[15] L. Gitelman, V. Jackson, D. Rosenberg, T. D. Williams, K. R. Brine, M. Poovey, M. Stanley, E. G. Garvey, M. Krajewski, R. Raley, D. Ribes, S. J. Jackson and G. C. Bowker, 'Data bite man: The work of sustaining a long-term study,' in *"Raw Data" Is an Oxymoron*. MIT Press, 2013, pp. 147–166.

[16] C. W. Whitney, B. K. Lind and P. W. Wahl, 'Quality assurance and quality control in longitudinal studies,' *Epidemiologic reviews*, vol. 20, no. 1, pp. 71–80, 1998.

[17] P. Quevauviller and E. Maier, 'Quality assurance and quality control for environmental monitoring,' in *Quality assurance in environmental monitoring-sampling and sample pretreatment*, VCH Weinheim New York, 1995, pp. 1–25.

[18] E. Parmiggiani and M. Grisot, 'Data curation as governance practice,' *Scandinavian Journal of Information Systems*, vol. 32, no. 1, 2020.

[19] H. Karasti, A. Botero, J. Saad-Sulonen and K. S. Baker, 'Configuring devices for phenomena in-the-making,' *Science & Technology Studies*, 2021.

[20] G. Walsham, ‘Interpretive case studies in is research: Nature and method,’ *European Journal of information systems*, vol. 4, no. 2, pp. 74–81, 1995.

[21] K. L. Mills, ‘Computer-supported cooperative work (CSCW),’ in *Encyclopedia of Library and Information Sciences, Third Edition*, CRC Press, 2009, pp. 1234–1249.

[22] K. Schmidt and L. Bannon, ‘Taking CSCW seriously,’ *Computer Supported Cooperative Work (CSCW)*, vol. 1, no. 1, pp. 7–40, 1992.

[23] M. S. Ackerman, ‘The intellectual challenge of CSCW: The gap between social requirements and technical feasibility,’ *Human–Computer Interaction*, vol. 15, no. 2-3, pp. 179–203, 2000.

[24] S. J. Jackson, T. Gillespie and S. Payette, ‘The policy knot: Re-integrating policy, practice and design in CSCW studies of social computing,’ in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 2014, pp. 588–602.

[25] K. S. Baker and F. Millerand, ‘Infrastructuring ecology: Challenges in achieving data sharing,’ in *Collaboration in the new life sciences*, Routledge, 2010, pp. 133–160.

[26] G. Mosconi, Q. Li, D. Randall, H. Karasti, P. Tolmie, J. Barutzky, M. Korn and V. Pipek, ‘Three gaps in opening science,’ *Computer Supported Cooperative Work (CSCW)*, vol. 28, no. 3, pp. 749–789, 2019.

[27] H. Karasti, K. S. Baker and F. Millerand, ‘Infrastructure time: Long-term matters in collaborative development,’ *Computer Supported Cooperative Work (CSCW)*, vol. 19, no. 3-4, pp. 377–415, 2010.

[28] S. L. Star, ‘The ethnography of infrastructure,’ *American behavioral scientist*, vol. 43, no. 3, pp. 377–391, 1999.

[29] D. Ribes and J. B. Polk, ‘Flexibility relative to what? change to research infrastructure,’ *Journal of the Association for Information Systems*, vol. 15, no. 5, p. 1, 2014.

[30] H. Karasti, F. Millerand, C. M. Hine and G. C. Bowker, ‘Knowledge infrastructures: Part i,’ *Science & Technology Studies*, vol. 29, no. 1, pp. 2–12, 2016.

[31] P. N. Edwards, ‘Knowledge infrastructures: Intellectual frameworks and research challenges,’ 2013.

[32] ESFRI. 'ESFRI Vision and Mission.' Accessed: 10.11.2021. (2021), [Online]. Available: `https://www.esfri.eu/esfri-white-paper/esfri-vision-and-mission`.

[33] LTER. 'LTER Site Profiles.' Accessed: 13.12.2021. (2021), [Online]. Available: `https://lternet.edu/site/`.

[34] M. H. Cragin, P. B. Heidorn, C. L. Palmer and L. C. Smith, 'An educational program on data curation,' *2007 STS Conference Poster Session*, 2007.

[35] E. Rahm and H. H. Do, 'Data cleaning: Problems and current approaches,' *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000.

[36] E. Parmiggiani, T. Østerlie and P. G. Almklov, 'In the backrooms of data science,' *Journal of the Association for Information Systems*, vol. 23, no. 1, pp. 139–164, 2022.

[37] A. S. Zimmerman, 'New knowledge from old data: the role of standards in the sharing and reuse of ecological data,' *Science, technology, & human values*, vol. 33, no. 5, pp. 631–652, 2008.

[38] European Commission. 'Turning FAIR into reality.' Accessed: 06.06.2022. (2018), [Online]. Available: `https://data.europa.eu/doi/10.2777/1524`.

[39] European Commission. 'Realising the European open science cloud: First report and recommendations of the Commission high level expert group on the European open science cloud.' Accessed: 12.12.2021. (2016), [Online]. Available: `https://data.europa.eu/doi/10.2777/940154`.

[40] P. N. Edwards, M. S. Mayernik, A. L. Batcheller, G. C. Bowker and C. L. Borgman, 'Science friction: Data, metadata, and collaboration,' *Social studies of science*, vol. 41, no. 5, pp. 667–690, 2011.

[41] M.-W. Dictionary. 'Systematic error.' Accessed 16.05.2022. (), [Online]. Available: `https://www.merriam-webster.com/dictionary/systematic%5C%20error.`.

[42] R. Nancy Carter, D. Bryant-Lukosius and R. Alba DiCenso, 'The use of triangulation in qualitative research,' in *Oncology nursing forum*, Oncology Nursing Society, vol. 41, 2014, p. 545.

[43] M. Mikalsen and E. Monteiro, 'Acting with inherently uncertain data: Practices of data-centric knowing,' *Journal of the Association for Information Systems*, vol. 22, no. 6, pp. 1715–1735, 2021.

[44] L. M. Jahnke and A. Asher, 'The problem of data: Data management and curation practices among university researchers,' *The Problem of Data*, pp. 3–32, 2012.

[45] L. L. Pipino, Y. W. Lee and R. Y. Wang, 'Data quality assessment,' *Communications of the ACM*, vol. 45, no. 4, pp. 211–218, 2002.

[46] ISO 9000:2015(en), 'Quality management systems - Fundamentals and vocabulary,' International Organization for Standardization, Geneva, CH, Standard, 2015.

[47] R. L. Habig, P. Thomas, K. Lippel, D. Anderson and J. Lachin, 'Central laboratory quality control in the national cooperative gallstone study,' *Controlled Clinical Trials*, vol. 4, no. 1-2, pp. 101–123, 1983.

[48] P. Konieczka and J. Namieśnik, *Quality assurance and quality control in the analytical chemical laboratory: a practical approach*. CRC Press, 2018.

[49] I. Erickson, K. Eschenfelder, S. Goggins, L. Hemphill, S. Sawyer, K. Shankar and K. Shilton, 'The ethos and pragmatics of data sharing,' in *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing*, 2014, pp. 109–112.

[50] Forskningsrådet. 'Open access to research data.' Accessed: 4.6.2022. (2021), [Online]. Available: `https://www.forskningsradet.no/en/Adviser-research-policy/open-science/open-access-to-research-data/`.

[51] Forskningsrådet. 'The Research Council Policy for Open Science.' Accessed: 28.10.2021. (2021), [Online]. Available: `https://www.forskningsradet.no/en/Adviser-research-policy/open-science/policy-for-open-science/`.

[52] European Commission. 'Horizon Europe.' Accessed: 2.12.2021. (2018), [Online]. Available: `https://ec.europa.eu/info/research-and-innovation/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe_en`.

[53] ESFRI. 'Roadmap 2018: Strategy Report on Research Infrastructures.' Accessed: 29.10.2021. (2018), [Online]. Available: `http://roadmap2018.esfri.eu//`.

[54] A. S. Zimmerman, *Data sharing and secondary use of scientific data: Experiences of ecologists*. University of Michigan, 2003.

[55] GO FAIR. 'FAIR Principles.' Accessed: 29.10.2021. (2017), [Online]. Available: `https://www.go-fair.org/fair-principles/`.

[56] P. Arzberger, P. Schroeder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, D. Moorman, P. Uhlir and P. Wouters, 'Promoting access to public research data for scientific, economic, and social development,' *Data Science Journal*, vol. 3, pp. 135–152, 2004.

[57] J. P. Birnholtz and M. J. Bietz, 'Data at work: Supporting sharing in science and engineering,' in *Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work*, 2003, pp. 339–348.

[58] J. C. Wallis, E. Rolando and C. L. Borgman, 'If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology,' *PLOS One*, vol. 8, no. 7, 2013.

[59] Y. Kim and J. M. Stanton, 'Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis,' *Journal of the Association for Information Science and Technology*, vol. 67, no. 4, pp. 776–799, 2016.

[60] F. Harris and F. Lyon, 'Transdisciplinary environmental research: Building trust across professional cultures,' *Environmental Science & Policy*, vol. 31, pp. 109–119, 2013.

[61] C. Tenopir, S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff and M. Frame, 'Data sharing by scientists: Practices and perceptions,' *PLOS One*, vol. 6, no. 6, 2011.

[62] E. Wenger *et al.*, 'Communities of practice: Learning as a social system,' *Systems thinker*, vol. 9, no. 5, pp. 1–10, 1998.

[63] C. Dictionary. 'Tacit knowledge.' Accessed 16.05.2022. (), [Online]. Available: `https://dictionary.cambridge.org/dictionary/english/tacit-knowledge`.

[64] J. S. Brown and P. Duguid, 'Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation,' *Organization science*, vol. 2, no. 1, pp. 40–57, 1991.

[65] K. Schmidt, *Cooperative work and coordinative practices: Contributions to the conceptual foundations of Computer-Supported Cooperative Work (CSCW)*. Springer Science & Business Media, 2011.

[66] K. Schmidt, "Keep Up the Good Work!': The Concept of 'Work' in CSCW,' in *Proceedings of COOP 2010*, Springer, 2010, pp. 265–285.

[67] K. Schmidt and C. Simone, 'Coordination mechanisms: Towards a conceptual foundation of cscw systems design,' *Computer Supported Cooperative Work (CSCW)*, vol. 5, no. 2-3, pp. 155–200, 1996.

[68] J. Tenenberg, W.-M. Roth and D. Socha, 'From i-awareness to we-awareness in cscw,' *Computer Supported Cooperative Work (CSCW)*, vol. 25, no. 4, pp. 235–278, 2016.

[69] W. Barley, J. Treem and P. Leonardi, 'Experts at coordination: Examining the performance, production, and value of process expertise,' *Journal of Communication*, vol. 70, pp. 60–89, Feb. 2020. DOI: 10.1093/joc/jqz041.

[70] N. A. Van House, M. H. Butler and L. R. Schiff, 'Cooperative knowledge work and practices of trust: Sharing environmental planning data sets,' in *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, 1998, pp. 335–343.

[71] J. D. Herbsleb, A. Mockus, T. A. Finholt and R. E. Grinter, 'Distance, dependencies, and delay in a global collaboration,' in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, 2000, pp. 319–328.

[72] C. Heath and P. Luff, 'Collaborative activity and technological design: Task coordination in London Underground control rooms,' in *Proceedings of the Second European Conference on Computer-Supported Cooperative Work ECSCW'91*, Springer, 1991, pp. 65–80.

[73] B. A. Nardi and J. R. Miller, 'An ethnographic study of distributed problem solving in spreadsheet development,' in *Proceedings of the 1990 ACM conference on Computer-supported cooperative work*, 1990, pp. 197–208.

[74] M. Song, H. Berends, H. Van der Bij and M. Weggeman, 'The effect of IT and co-location on knowledge dissemination,' *Journal of product innovation management*, vol. 24, no. 1, pp. 52–68, 2007.

[75] M. E. Warkentin, L. Sayeed and R. Hightower, 'Virtual teams versus face-to-face teams: An exploratory study of a web-based conference system,' *Decision sciences*, vol. 28, no. 4, pp. 975–996, 1997.

[76] T. J. Allen and G. Henn, *The Organization and Architecture of Innovation: Managing the Flow of Technology*. Taylor & Francis Group, 2011, vol. 1.

[77] J. Xie, M. Song and A. Stringfellow, 'Antecedents and consequences of goal incongruity on new product development in five countries: A marketing view,' *Journal of Product Innovation Management*, vol. 20, no. 3, pp. 233–250, 2003.

[78] Y. Mao, D. Wang, M. Muller, K. R. Varshney, I. Baldini, C. Dugan and A. Mojsilović, 'How data scientists work together with domain experts in scientific collaborations: To find the right answer or to ask the right question?' *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, pp. 1–23, 2019.

[79] S. Shapin, 'Trust, honesty, and the authority of science,' in *Society's Choices: Social and Ethical Decision Making in Biomedicine*, The National Academies Press, 1995, pp. 388–408. DOI: https://doi.org/10.17226/4771.

[80] J. K. Butler Jr and R. S. Cantrell, 'A behavioral decision theory approach to modeling dyadic trust in superiors and subordinates,' *Psychological reports*, vol. 55, no. 1, pp. 19–28, 1984.

[81] P. L. Schindler and C. C. Thomas, 'The structure of interpersonal trust in the workplace,' *Psychological Reports*, vol. 73, no. 2, pp. 563–573, 1993.

[82] A. H. Schweiger, S. D. Irl, M. J. Steinbauer, J. Dengler and C. Beierkuhnlein, 'Optimizing sampling approaches along ecological gradients,' *Methods in Ecology and Evolution*, vol. 7, no. 4, pp. 463–471, 2016.

[83] M. Power, 'The predictive validation of ecological and environmental models,' *Ecological modelling*, vol. 68, no. 1-2, pp. 33–50, 1993.

[84] H. M. Regan, H. R. Akçakaya, S. Ferson, K. V. Root, S. Carroll and L. R. Ginzburg, 'Treatments of uncertainty and variability in ecological risk assessment of single-species populations,' *Human and Ecological Risk Assessment*, vol. 9, no. 4, pp. 889–906, 2003.

[85] S. Passi and S. J. Jackson, 'Trust in data science: Collaboration, translation, and accountability in corporate data science projects,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–28, 2018.

[86] C. Anderson, 'The end of theory: The data deluge makes the scientific method obsolete,' *Wired magazine*, vol. 16, no. 7, pp. 16–07, 2008.

[87] J. C. Refsgaard, J. P. van der Sluijs, A. L. Højberg and P. A. Vanrolleghem, 'Uncertainty in the environmental modelling process–a framework and guidance,' *Environmental modelling & software*, vol. 22, no. 11, pp. 1543–1556, 2007.

[88] G. E. Box, 'Robustness in the strategy of scientific model building,' in *Robustness in statistics*, Elsevier, 1979, pp. 201–236.

[89] eLTER. 'About LTER-Europe.' Accessed: 8.11.2021. (2016), [Online]. Available: https://www.lter-europe.net/lter-europe/about.

[90] eLTER. 'Objectives.' Accessed: 4.6.2022. (2014), [Online]. Available: https://www.lter-europe.net/lter-europe/about/objectives.

[91] European Commission. 'European Strategy Forum on Research Infrastructures (ESFRI).' Accessed: 8.11.2021. (2019), [Online]. Available: https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/european-research-infrastructures/esfri_en.

[92] Forskningsrådet. 'Norwegian roadmap for Research Infrastructure.' Accessed: 03.06.2022. (2020), [Online]. Available: https://www.forskningsradet.no/sok-om-finansiering/midler-fra-forskningsradet/infrastruktur/norsk-veikart-for-forskningsinfrastruktur/omradestrategier/klima-og-miljo/.

[93] P. Baxter and S. Jack, 'Qualitative case study methodology: Study design and implementation for novice researchers,' *The Qualitative Report*, 2008.

[94] H. K. Klein and M. D. Myers, 'A set of principles for conducting and evaluating interpretive field studies in information systems,' *MIS quarterly*, pp. 67–93, 1999.

[95] G. Walsham, 'Doing interpretive research,' *European journal of information systems*, vol. 15, no. 3, pp. 320–330, 2006.

[96] M. D. Myers and M. Newman, 'The qualitative interview in is research: Examining the craft,' *Information and organization*, vol. 17, no. 1, pp. 2–26, 2007.

[97] G. A. Bowen, 'Document analysis as a qualitative research method,' *Qualitative research journal*, 2009.

[98] Y. Chandra and L. Shang, *Qualitative research using R: A systematic approach*. Springer, 2019.

[99] A. Strauss and J. Corbin, 'Basics of qualitative research techniques,' 1998.

[100] A. Tjora, *Kvalitative forskningsmetoder i praksis*. Gyldendal akademisk Oslo, 2012, vol. 2.

[101] U. of Illinois Chicago Library. 'What is a DOI and how do I use it in a citation?' Accessed: 1.6.2022. (2018), [Online]. Available: https://researchguides.uic.edu/doi.

[102] L. H. Keith, *Environmental sampling and analysis: a practical guide*. Routledge, 2017.

[103] J. G. Townsend and A. R. Townsend, ‘Accountability, motivation and practice: Ngos north and south,’ *Social & Cultural Geography*, vol. 5, no. 2, pp. 271–284, 2004.

[104] J. E. Orr, ‘Narratives at work: Story telling as cooperative diagnostic activity,’ in *Proceedings of the 1986 ACM conference on Computer-supported cooperative work*, 1986, pp. 62–72.

[105] L. R. Christensen, *Coordinative practices in the building process: An ethnographic perspective*. Springer Science & Business Media, 2012.

[106] R. C. Mayer, J. H. Davis and F. D. Schoorman, ‘An integrative model of organizational trust,’ *Academy of management review*, vol. 20, no. 3, pp. 709–734, 1995.

[107] A. Yoon and Y. Kim, ‘The role of data-reuse experience in biological scientists’ data sharing: An empirical analysis,’ *The Electronic Library*, 2020.

[108] A. Currie and W. Kilbride, ‘FAIR Forever? Accountabilities and Responsibilities in the Preservation of Research Data,’ 2021.

Appendix A

Interview Guides

This appendix contains all interview guides used in this project, as well as the information letter and consent form all interviewees signed before the interviews. These were all combined into one file that was sent out by email to all participants. The documents appear in the following order:

1. Information letter, consent form, and interview guide for interviews with researchers concerned with data work.
2. A Norwegian version of the interview guide above.
3. Information letter, consent form, and interview guide for interview with eLTER work package leaders.

Interview Guide:

Investigating how environmental researchers work with scientific data: Coordination & (Re)Use

What is this about?

This is an inquiry about the practices, work processes, information systems, and infrastructures adopted by participants who work in environmental monitoring research stations to prepare, integrate, and make sense of heterogeneous datasets on a day-to-day basis. The aim is to contribute with empirical insights on how participants in research stations work on data including how they work with each other; and how these work practices are affected by the technologies they use and the workplace policies on a daily basis.

Who is responsible for the research project?

Norwegian University of Science and Technology (NTNU, Trondheim).

What does participation involve for you?

By taking part in this study, you are contributing towards the advancement of science by taking an active role to provide information that can help in extending the boundaries of knowledge.

How will your data be used?

The research methodology (qualitative research) incorporates information that will be collected through interviews. This information will be recorded on paper and/or by audio recording. The notes will be taken manually and stored electronically.

Participation is voluntary

Participation in this interview is voluntary. There will be no negative consequences for you if you choose not to participate or later decide to withdraw. You can withdraw your consent at any time by contacting us at [redacted] or [redacted], and your data will immediately be deleted.

Your privacy – how we will store and use your personal data

- We will only use your personal data for the purpose(s) specified in this information letter. We will process your personal data confidentially and under EU data protection legislation (the General Data Protection Regulation and Personal Data Act).
- Only the project leader, project supervisors and members of the project affiliated with the institution responsible for the project will have access to the personal data.
- Your name and contact details will be replaced with a code. The list of names, contacts, and respective codes will be stored separately from the rest of the collected data. The data will be stored on a research server.
- The participants will not be recognizable in publications/reports submitted to NTNU.

Your rights

So long as you can be identified in the collected data, you have the right to:

- access the personal data that is being processed about you
- request that your personal data is deleted
- request that incorrect personal data about you is corrected/rectified
- receive a copy of your personal data (data portability), and
- send a complaint to the Data Protection Officer or The Norwegian Data Protection Authority regarding the processing of your personal data

What gives us the right to process your personal data?

We will process your personal data based on your consent.

Based on an agreement with The Norwegian University of Science & Technology (NTNU), NSD – The Norwegian Centre for Research Data AS has assessed that the processing of personal data in this project is in accordance with data protection legislation.

Where can I find out more?

If you have questions about the project, or want to exercise your rights, please contact:

- NTNU via Maria Ivanova: [redacted] or Nana Kwame Amagyei: [redacted]

- NSD – The Norwegian Centre for Research Data AS, by email: (personverntjenester@nsd.no) or by telephone: +47 55 58 21 17

Yours sincerely,

Project Leader:	Elena Parmiggiani (Associate Professor), [redacted]
Students:	Nana Kwame Amagyei (PhD Candidate), [redacted]
	Maria Ivanova (Masters Student),[redacted]

Consent form

I have received and understood information about the project "Investigating how environmental researchers work with scientific data: Coordination & (Re)Use". I have been given the opportunity to ask questions.

I give consent for my personal data to be collected through interview and observation, and for this data to be stored and processed until 10.01.2024, the end date of this project.

(Participant name, participant signature, date)

Questions for Respondents

Note: the following questions are only meant as a guide, and might be altered and/or elaborated during the interview process.

About you

1. What is your primary role?
2. Take me through your average work day (both pre and post pandemic)
   a. Any specific tools you use?
3. In your experience, is your job different from your job description?
   a. Challenges?
   b. Do you use more time/resources on certain things (especially those not present in job description)?

Policies

1. Are you required to make data available to other research stations? Who makes these requirements? Data management plan
2. What guidelines (e.g., agreed practises, standards) are used for quality assurance and data reusability?
3. How is data saved, located and accessed?

Coordination

1. Do you follow a defined process when dealing with data? Describe this (if affected by pandemic, describe before and after).
2. What time constraints do you face?
   a. How do you make sure that you get your work done on time?
   b. Does this depend on others(/external factors)? How?
3. Do you use any tools that provide an overview over the process of collecting, filtering, and prepping data, or is this solved through verbal/written communication with colleagues?
   a. What do you think of this/these tool(s)?
4. What can be done differently to improve overall coordination (of work/data)?

Cooperation

Now I was thinking we could talk a little bit about cooperation between you and your colleagues, and maybe even groups of people. Now, when I say I want to take a closer look at your cooperation I mean that I want to look at everything that makes data go from creation, and all the way to a pretty and structured dataset ready to be shared. And so I want to look at not only what, *but also* who makes it happen *and* how.

1. What do you think of when I say cooperation?
2. Do you think the nature of your work is one where you wish to work as independently from others as possible, or one where you need to cooperate to get tasks done? Why?
3. Do new technologies (new software, sensors, servers, databases) for collection and prepping change often?
    a. How does this affect the work with your colleagues?
2. What is important for your colleagues to understand about the work you do?
3. Do you use any collaboration or communication tools with colleagues? (as simple as Teams or project management tools such as Trello/Asana)
    a. Do these work well? Why/why not?
4. Who (what roles) do you most depend on to get your work done?
    a. What needs to be in place?
    b. How does the quality of their work determine the quality of yours?
5. Who (what role) depends on you to get their work done?
6. Name one aspect of cooperation between you and your colleagues that works well.
7. Name one aspect of cooperation between you and your colleagues that could be improved (could be technical or governance related).
8. What can be done differently to improve cooperation between you and your colleagues?
9.

Open Data

1. Do you feel that you are being adequately incentivized (e.g., to share data, to make the data you share valuable and reusable for others, etc.)?
2. What do you think needs to be in place cooperatively for open data sharing to be successful (should they or I define success? Would be interesting to see what they think is actually successful data sharing since it probably influences their work)?
    a. Do you think this has been achieved?

b. Do you think this is hard to achieve?

c. Are there any sacrifices?

3. What extra work do you put into data prepping for it to be reusable?
   a. How does the workload depend on the work your colleagues do?
   b. How does the difficulty of the task (to make data reusable) depend on the work your colleagues do?

NORSK VERSJON

Spørsmålsskjema for intervjuobjekter

Om deg

1. Hva er det du primært jobber med?
    a. Hva er en viktig del av jobben din?
2. Vil du ta meg gjennom en standard arbeidsdag? (både før og etter Corona)
    a. Bruker du noe spesielle verktøy du vil fortelle om?
3. Vil du si at det arbeidet du gjør passer jobbstillingen din - altså det som sto på papiret du signerte?
    a. Utfordringer?
    b. Bruker du mere tid/ressurser på visse ting enn andre/antatt (spesielt tin som kanskje ikke står i arbeidsbeskrivelsen din)?

Policies

1. Er det noen krav i forhold til datadeling (f.eks. med andre forsningsstasjoner) som du må forholde deg til? Og hvem er det som definerer og stiller disse kravene i så fall?
2. Vil du fortelle meg om du må forholde deg til noen retningslinjer (standarder, avtalte måter å gjøre ting på, osv) av den typen som påvirker kvalitetssikring og datagjenbruk?
    a. hvordan påvirker disse arbeidet ditt?
    b. Vil du si disse krever mye samarbeid?
3. Hvordan blir data lagret, funnet frem, og hentet ut?

Koordinasjon(?)

1. Følger du en definert prosess/plan når du håndterer data? Forklar denne (hvis Covid har påvirket dette, kan du forklare før/etter).
2. Hvilke tidsbegrensninger må du ta hensyn til når du jobber?
    a. Hvordan sikrer du at du får gjort det du skal innen relevante frister?

   b. Er dette noe som avhenger av andre folk(/faktorer)? Hvordan?

3. Bruker du noe form for verktøy som gir deg oversikt/ innblikk i prosesser slik som samling, filtrering, tilberedelse etc av data? Eller er dette noe som koordineres skriftlig/muntlig med kollegaer?
   a. Hva synes du om disse verktøyene?
4. Hva tror du kan bli gjort annerledes for å forbedre koordinasjon(koordinering av arbeid/data)?

Samarbeid

Nå tenker jeg at vi kanskje kan snakke litt om samarbeid blant deg og dine kollegaer, og kanskje til og med mellom forskjellige grupper dere har på arbeidsplassen - i den forstand at jeg vil se litt på hva som skal til for å få data til å gå fra skapelse til at den er klar for å deles, og hvem *som skal til for at dette skal gå som det skal.*

1. Hva legger du i ordet samarbeid?
2. Vil du si at du på generell basis har arbeidsoppgaver der du vil jobbe i fred og ro og egentlig ganske uavhengig fra resten, eller har du kanskje mer arbeidsoppgaver som krever en del samarbeid for å få gjort ting? Hvordan da?
3. Er det ofte at dere må bytte til nye verktøy og generelt ny teknologi? (ny programvare, sensorer, servere, databaser)?
   a. Hvordan har dette eventuelt påvirket samarbeid og generelt arbeid med kollegaer?
4. Bruker du noen form for samarbeidsverktøy eller kommunikasjonsverktøy med kollegaene dine? (så simpelt som Teams til prosjektmanagement-verktøy som Asana eller Trello)
5. Vil du si disse fungerer bra? Hvorfor/hvorfor ikke?
6. Hvem(hvilke roller) avhenger du mest av for å klare å gjøre arbeidet ditt?
   a. Hva må være på plass?
   b. Hvordan påvirker kvaliteten på deres arbeid kvaliteten på arbeidet ditt?
7. Hvem(hvilke roller) avhenger av deg for å klare å gjøre sitt arbeid?
8. Nevn et eksempel på godt samarbeid mellom deg og kollegaene dine.
9. Nevn et eksempel på samarbeid mellom deg og kollegaene dine som har forbedringspotensiale (kan være direkte påvirket av eksterne faktorer som f.eks. teknologi eller policy).

10. Hva tror du kan bli gjort annerledes for å forbedre samarbeid mellom deg og dine kollegaer?
11. Hva er det som er viktig at kollegaene dine forstår om arbeidet du gjør?

Åpen Data

1. Føler du at du får nok ekstern motivasjon? (både med tanke på det å dele data, men også å gjøre dataen du deler både verdifull og gjenbrukbar)?
2. Hva tror du må være på plass samarbeidsmessig for at åpen datadeling skal lykkes? (should they or I define success? Would be interesting to see what they think is actually successful data sharing since it probably influences their work)?
    a. Er dette noe du ser på arbeidsplassen din (at pen datadeling har lykkes)?
    b. Tror du det er vanskelig å oppnå suksessfull åpen datadeling?
    c. Føler du at man må ofre visse ting for å lykkes med åpen datadeling?

3. Hvilket ekstra arbeid legger du i data for at den skal være gjenbrukbar?
    a. Hvordan blir denne arbeidsmengden påvirket av arbeidet kollegaene dine gjør?
    b. Hvordan blir vanskelighetsgraden av å gjøre data gjenbrukbar påvirket av arbeidet kollegaene dine gjør?

Interview Guide for Work Package Leaders:

Investigating how environmental researchers work with scientific data: Coordination & (Re)Use

What is this about?

This is an inquiry about the practices, work processes, and thoughts of participants who work on the organizational side of eLTER, with an emphasis on policy and its relation to eLTER. The aim is to contribute with empirical insights on how participants in research stations are affected by policy in their coordination and open data sharing practices, and, from the other side, what processes and practices are behind the forces acting on policy in environmental research.

Who is responsible for the research project?

Norwegian University of Science and Technology (NTNU, Trondheim).

What does participation involve for you?

By taking part in this study, you are contributing towards the advancement of science by taking an active role to provide information that can help in extending the boundaries of knowledge.

How will your data be used?

The research methodology (qualitative research) incorporates information that will be collected through interviews. This information will be recorded on paper and/or by audio recording. The notes will be taken manually and stored electronically.

Participation is voluntary

Participation in this interview is voluntary. There will be no negative consequences for you if you choose not to participate or later decide to withdraw.. You can withdraw your consent at any time by contacting us at [redacted] or [redacted], and your data will immediately be deleted.

Your privacy – how we will store and use your personal data

- We will only use your personal data for the purpose(s) specified in this information letter. We will process your personal data confidentially and under EU data protection legislation (the General Data Protection Regulation and Personal Data Act).
- Only the project leader, project supervisors and members of the project affiliated with the institution responsible for the project will have access to the personal data.
- Your name and contact details will be replaced with a code. The list of names, contacts, and respective codes will be stored separately from the rest of the collected data. The data will be stored on a research server.
- The participants will not be recognizable in publications/reports submitted to NTNU.

Your rights

So long as you can be identified in the collected data, you have the right to:

- access the personal data that is being processed about you
- request that your personal data is deleted
- request that incorrect personal data about you is corrected/rectified
- receive a copy of your personal data (data portability), and
- send a complaint to the Data Protection Officer or The Norwegian Data Protection Authority regarding the processing of your personal data

What gives us the right to process your personal data?

We will process your personal data based on your consent.

Based on an agreement with The Norwegian University of Science & Technology (NTNU), NSD – The Norwegian Centre for Research Data AS has assessed that the processing of personal data in this project is in accordance with data protection legislation.

Where can I find out more?

If you have questions about the project, or want to exercise your rights, please contact:

- NTNU via Maria Ivanova: [redacted] or Nana Kwame Amagyei: [redacted]
- NSD – The Norwegian Centre for Research Data AS, by email: (personverntjenester@nsd.no) or by telephone: +47 55 58 21 17

Yours sincerely,

Project Leader:	Elena Parmiggiani (Associate Professor), [redacted]
Students:	Nana Kwame Amagyei (PhD Candidate), [redacted]
	Maria Ivanova (Masters Student), [redacted]

Consent form

I have received and understood information about the project "Investigating how environmental researchers work with scientific data: Coordination & (Re)Use". I have been given the opportunity to ask questions.

I give consent for my personal data to be collected through interview and observation, and for this data to be stored and processed until 10.01.2024, the end date of this project.

(Participant name, participant signature, date)

Questions for Respondents

Note: the following questions are only meant as a guide, and might be altered and/or elaborated during the interview process.

About you

- What is your primary role in eLTER?
- What do you think is important in regards to eLTER and its evolution?
- Take me through your average work day (both pre and post pandemic)
  - What are your main tasks?
- What are your main responsibilities (decisions that you make)?
- Who are key people (roles) in your work?
- Who supports you in decision making?
- Name an unexpected challenge that comes with your work.

Policies

- How do you relate to policies? What do you think of when you think of policy?
- How do you interact with policy in your work?
  a. How do you think eLTERs relation/attitude to policies evolves?
- What do you believe are the key influences when it comes to policy creation (do people involved in forming policies seek out information/needs or do the needs ‘seek’ these people out)
  - How are those affected by policy taken into account?

Open data sharing

- What do you think about when you hear open data (sharing)?
- Describe the value of open data sharing
- Do you think open data sharing fosters reusability of data, or is it more of a requirement / bump in the road that people brush over?
  - How do you think policy affects coordination needed for data sharing?

    - What do you think is the role of data curation(term definition can be given) in achieving valuable open data?
    - How do you think this potentially could be implemented further in funding and policy?
- (In the context of open data sharing:) What do you believe is the upside of RI's and the subsequent unification and integration of research sites and systems?
    - (In the context of open data sharing:) What do you believe is the downside of this unification and integration?

Appendix B

NSD Application and Assessment

Below you will find the application, "Meldeskjema", to NSD (Norwegian Centre for Research Data) and the assessment, "Vurdering", received (approved).

NORSK SENTER FOR FORSKNINGSDATA

Meldeskjema

Referansenummer

887930

Hvilke personopplysninger skal du behandle?

- Navn (også ved signatur/samtykke)
- E-postadresse, IP-adresse eller annen nettidentifikator
- Lydopptak av personer
- Gps eller andre lokaliseringsdata (elektroniske spor)

Prosjektinformasjon

Prosjekttittel

Researching coordination of scientific work for open data sharing in environmental research

Prosjektbeskrivelse

This case study examines the coordination of day-to-day practices, work processes, information systems, and infrastructures adopted by researchers to collect and prepare data to achieve open data sharing. Additionally, I will be looking at how policy setters view this coordination by interviewing one or more project leaders.

Begrunn behovet for å behandle personopplysningene

Name and e-mail address: Contact info and consent form signature.
Geolocation data: I will need information on what researcher belongs to what research station. Thus, research stations will be tied to individual researchers.
Audio recording: To gain adequate data from interviews. Things said can thus be examined further, and the process of the interview can be analyzed.

Ekstern finansiering

Type prosjekt

Studentprosjekt, masterstudium

Kontaktinformasjon, student

Maria Ivanova, [redacted]

Behandlingsansvar

Behandlingsansvarlig institusjon

Norges teknisk-naturvitenskapelige universitet / Fakultet for informasjonsteknologi og elektroteknikk (IE) / Institutt for datateknologi og informatikk

Prosjektansvarlig (vitenskapelig ansatt/veileder eller stipendiat)

Elena Parmiggiani, [redacted]

Skal behandlingsansvaret deles med andre institusjoner (felles behandlingsansvarlige)?

Nei

Utvalg 1

Beskriv utvalget

Data workers of environmental research stations in Norway.

Rekruttering eller trekking av utvalget

This project will be conducted with my co-supervisor Nana Kwame Henebeng Amagyei, who already has a list of relevant contacts for his ongoing project. Thus, I will start by contacting them, and then contacting new contacts that will be recommended by them.

Alder

18 - 70

Inngår det voksne (18 år +) i utvalget som ikke kan samtykke selv?

Nei

Personopplysninger for utvalg 1

- Navn (også ved signatur/samtykke)
- E-postadresse, IP-adresse eller annen nettidentifikator
- Lydopptak av personer
- Gps eller andre lokaliseringsdata (elektroniske spor)

Hvordan samler du inn data fra utvalg 1?

Personlig intervju

Grunnlag for å behandle alminnelige kategorier av personopplysninger

Samtykke (art. 6 nr. 1 bokstav a)

Deltakende observasjon

Grunnlag for å behandle alminnelige kategorier av personopplysninger

Samtykke (art. 6 nr. 1 bokstav a)

Informasjon for utvalg 1

Informerer du utvalget om behandlingen av opplysningene?

Ja

Hvordan?

Skriftlig informasjon (papir eller elektronisk)

Utvalg 2

Beskriv utvalget

Project leader(s) of research networks

Rekruttering eller trekking av utvalget

Recruitement by e-mail.

Alder

18 - 70

Inngår det voksne (18 år +) i utvalget som ikke kan samtykke selv?

Nei

Personopplysninger for utvalg 2

- Navn (også ved signatur/samtykke)
- E-postadresse, IP-adresse eller annen nettidentifikator
- Lydopptak av personer

Hvordan samler du inn data fra utvalg 2?

Personlig intervju

Grunnlag for å behandle alminnelige kategorier av personopplysninger

Samtykke (art. 6 nr. 1 bokstav a)

Informasjon for utvalg 2

Informerer du utvalget om behandlingen av opplysningene?

Ja

Hvordan?

Skriftlig informasjon (papir eller elektronisk)

Tredjepersoner

Skal du behandle personopplysninger om tredjepersoner?

Nei

Dokumentasjon

Hvordan dokumenteres samtykkene?

- Manuelt (papir)
- Elektronisk (e-post, e-skjema, digital signatur)

Hvordan kan samtykket trekkes tilbake?

Data subjects will be free to withdraw anytime during the study by sending an email to me, my co-supervisor or the project leader of this study (my supervisor).

Hvordan kan de registrerte få innsyn, rettet eller slettet opplysninger om seg selv?

Data subjects can gain access to correct or delete their data by sending an email to the project leader.

Totalt antall registrerte i prosjektet

1-99

Tillatelser

Skal du innhente følgende godkjenninger eller tillatelser for prosjektet?

Behandling

Hvor behandles opplysningene?

- Maskinvare tilhørende behandlingsansvarlig institusjon
- Mobile enheter tilhørende behandlingsansvarlig institusjon

Hvem behandler/har tilgang til opplysningene?

- Student (studentprosjekt)
- Interne medarbeidere
- Prosjektansvarlig

Tilgjengeliggjøres opplysningene utenfor EU/EØS til en tredjestat eller internasjonal organisasjon?

Nei

Sikkerhet

Oppbevares personopplysningene atskilt fra øvrige data (koblingsnøkkel)?

Ja

Hvilke tekniske og fysiske tiltak sikrer personopplysningene?

- Flerfaktorautentisering
- Adgangsbegrensning
- Opplysningene anonymiseres fortløpende

Varighet

Prosjektperiode

17.01.2022 - 12.06.2022

Skal data med personopplysninger oppbevares utover prosjektperioden?

Ja, data med personopplysninger oppbevares til: 10.01.2024

Til hvilket formål skal opplysningene oppbevares?

Forskning

Vil de registrerte kunne identifiseres (direkte eller indirekte) i oppgave/avhandling/øvrige publikasjoner fra prosjektet?

Nei

Tilleggsopplysninger

Under 'Duration of processing' I have informed that data will be stored until 10.01.2024. This is only if the data is relevant for my co-supervisor Nana Kwame Henebeng Amagyei, and if he needs it for the whole duration of his Phd.

NSD NORSK SENTER FOR FORSKNINGSDATA

Vurdering

Referansenummer

887930

Prosjekttittel

Researching coordination of scientific work for open data sharing in environmental research

Behandlingsansvarlig institusjon

Norges teknisk-naturvitenskapelige universitet / Fakultet for informasjonsteknologi og elektroteknikk (IE) / Institutt for datateknologi og informatikk

Prosjektansvarlig (vitenskapelig ansatt/veileder eller stipendiat)

Elena Parmiggiani, [redacted]

Type prosjekt

Studentprosjekt, masterstudium

Kontaktinformasjon, student

Maria Ivanova, [redacted]

Prosjektperiode

17.01.2022 - 12.06.2022

Vurdering (1)

28.01.2022 - Vurdert

Our assessment is that the processing of personal data in this project will comply with data protection legislation, so long as it is carried out in accordance with what is documented in the Notification Form and attachments, dated 28.01.2022. Everything is in place for the processing to begin.

TYPE OF DATA AND DURATION
The project will be processing general categories of personal data until 12.06.2022.

General categories of personal data will for research purposes be stored after the project end date until 10.01.2024

LEGAL BASIS
The project will gain consent from data subjects to process their personal data. We find that consent will meet the necessary requirements under art. 4 (11) and 7, in that it will be a freely given, specific, informed and unambiguous statement or action, which will be documented and can be withdrawn.

The legal basis for processing general categories of personal data is therefore consent given by the data subject, cf. the General Data Protection Regulation art. 6.1 a).

PRINCIPLES RELATING TO PROCESSING PERSONAL DATA

Data protection services finds that the planned processing of personal data will be in accordance with the principles under the General Data Protection Regulation regarding:

- lawfulness, fairness and transparency (art. 5.1 a), in that data subjects will receive sufficient information about the processing and will give their consent
- purpose limitation (art. 5.1 b), in that personal data will be collected for specified, explicit and legitimate purposes, and will not be processed for new, incompatible purposes
- data minimisation (art. 5.1 c), in that only personal data which are adequate, relevant and necessary for the purpose of the project will be processed
- storage limitation (art. 5.1 e), in that personal data will not be stored for longer than is necessary to fulfil the project's purpose

THE RIGHTS OF DATA SUBJECTS

As long as the data subjects can be identified in the data material, they will have the following rights: access (art. 15), rectification (art. 16), erasure (art. 17), restriction of processing (art. 18), data portability (art. 20).

Data protection services finds that the information that will be given to data subjects about the processing of their personal data will meet the legal requirements for form and content, cf. art. 12.1 and art. 13.

We remind you that if a data subject contacts you about their rights, the data controller has a duty to reply within a month.

FOLLOW YOUR INSTITUTION'S GUIDELINES

Data protection services presupposes that the project will meet the requirements of accuracy (art. 5.1 d), integrity and confidentiality (art. 5.1 f) and security (art. 32) when processing personal data.

To ensure that these requirements are met you must follow your institution's internal guidelines and/or consult with your institution (i.e. the institution responsible for the project).

NOTIFY CHANGES

If you intend to make changes to the processing of personal data in this project it may be necessary to notify Data protection services . This is done by updating the Notification Form. On our website we explain which changes must be notified: https://www.nsd.no/en/data-protection-services/notification-form-for-personal-data/notify-changes-in-the-notification-form

Wait until you receive an answer from us before you carry out the changes.

FOLLOW-UP OF THE PROJECT

Data protection services will follow up the progress of the project at the planned end date in order to determine whether the processing of personal data has been concluded.

Good luck with the project!

NTNU

Norwegian University of
Science and Technology